A HYBRID CONTEXTUAL EMBEDDING BASED CLUSTERING AND CLASSIFICATION TECHNIQUE FOR UNSUPERVISED IMPLICIT ASPECT CATEGORIZATION IN INDONESIAN REVIEWS

Nur Hayatin[1], Suraya Alias[2*], Lai Po Hung[3]

[1]Department of Informatics, Faculty of Engineering, University of Muhammadiyah Malang, Malang, Indonesia

[1, 2, 3] Faculty of Computing and Informatics, Universiti Malaysia Sabah, Kota Kinabalu, Malaysia

Emails: noorhayatin@umm.ac.id[1], suealias@ums.edu.my[2*], laipohung@ums.edu.my[3]

ABSTRACT

Aspect categorization is a grouping of reviews based on aspect categories that follow the review domain. The problem arises when only sentiment features appear as a clue to predict implicit aspects. On the other hand, implicit aspects play an important role in generating a summary. Without implicit aspect, we probably lose some important words needed for analyzing user's reviews. Existing techniques face difficulties in utilizing the implicit aspects due to limited resources and computationally expensive problems. Hence, we propose an implicit aspect categorization model based on a hybrid contextual embedding-based clustering and classification technique. We developed the model using an unsupervised learning approach which is no need labelled data in training. A contextual embedding-based clustering technique generated train data from explicit sentences which will be used to classify implicit aspect categorization. Four steps of the proposed implicit aspects categorization model, i.e. preprocessing data, sentence feature selection, generating train data based on clustering, and finally categorizing implicit aspect using classification technique. We experiment with several classification techniques to get the best combination of the proposed technique (i.e. Logistic Regression, Support Vector Machine, Naïve Bayes, Decision Tree, and Random Forest). Based on the experiment, the combination of contextual embedding-based clustering and Random Forest algorithm produces higher accuracy than other classification techniques, with accuracy tent to 72.04% and F1 score in 0.6788.

Keywords: Aspect categorization; Contextual embedding; Clustering; Classification; Unsupervised learning; Indonesian reviews.

1.0 INTRODUCTION

An aspect is a feature or component of a product or service that can be expressed directly as a word or phrase in a sentence. Aspect extraction involves automatically identifying and categorizing specific features of a product or service mentioned in the review. Meanwhile, aspect categorization is a grouping of reviews based on the aspect categories that follow the review domain. Aspect extraction plays an important role in sentiment analysis and is one of the three essential stages in sentiment summarization [1].

Aspect is the key point of what is being talked about and commented on, while sentiment is a point of view that expresses feelings. For example, sentence: "*The dress is pretty*", the word "*dress*" represents the aspect, and the word "*pretty*" represents the sentiment term. Extracting aspects from user-generated content is the most fundamental task to get a fine-grained of the sentences [2]. In general, aspects can be explicitly expressed in sentences called explicit aspect, but customers' reviews can be written without explicitly mentioning the target of the opinion, called implicit aspect [3]. They also usually used specific terms to express their opinion. For example, sentence: "*too small..disappointed*", the sentence is talking about *size*, but this aspect is not explicitly mentioned in the sentence. Study in [4] argued that most studies do recommend implicit as a feasible future direction because it is considered as the latest aspect extraction area.

From previous studies, several unsupervised techniques for aspect categorization have been developed such as statistical [5], semantic [6], topic modeling [7], and dependency parsing [8]. However, it was found that most researchers face difficulties in utilizing the implicit aspects due to limited resources, complexity, and high computational cost. In fact, without the implicit aspect feature, we probably lose some important words needed for

analyzing user's reviews. It is also found in some cases there is only opinion as a clue to predict implicit aspects [9].

Indonesian reviews have potential to be analysis to support business and potential customers. However, there are still limited studies focusing on implicit aspect categorization due to the complexity and restricted resources. In [10], the study extracted pairs of aspect and opinion word in implicit opinion sentences on Indonesian hotel reviews using statistical approach based on Rules and co-occurrence matrix. They proposed Indonesian complex implicit opinion sentences based on complex compound sentences, a combination of equivalent compound and complex sentence.

Another study by [11] have identified an implicit aspect from Indonesian opinion sentences using combination of Rules based on Opinion Word Similarity (OWS). This research extracts implicit aspects by recognizing the relation between opinion words with aspect categories. Start by identifying a pair of opinion sentences and explicit aspect from train data, then extracting rule-based knowledge on user reviews based on co-occurrence and evaluating it using frequency and Association Rule Mining. Extracting knowledge is necessary to identify opinion sentences with explicit aspects and get pairs of aspects and words of opinion with rules generated from regular expressions. Finally, in the testing phase the knowledge will be used to identify an implicit aspect of opinion sentences from test data.

We hypothesis that implicit aspect category can be predicted effectively by learning sentences containing explicit aspects and considering contextual approach. Predicting the category of implicit aspect is complex because the opinion target is written following the user's style. We show that the absence of labeled data does not hinder supervised learning in implicit aspect categorization.

Therefore, the objective of our study is to develop implicit aspect categorization model using a hybrid contextual embedding-based clustering and classification technique for Indonesian reviews. Even though we implement classification in our proposed technique, the proposed is developed in unsupervised learning so that no need labelled data in training. Instead, we implement clustering to generate train data which will be used to classify implicit aspect categorization. Four steps for implicit aspect categorization using the proposed technique, i.e.: preprocessing data, sentence feature selection, generating train data based on clustering, and finally categorizing implicit aspect using classification models. We expect that the combination of clustering and the contextual embedding model will generate aspect categorization clusters that are more resilient to noise and outliers while also reducing the variance within each cluster. Finally, good clusters result as train data can increase the performance of the implementation in implicit aspect categorization for Indonesian reviews.

2.0 RELATED WORKS

Several unsupervised techniques for aspect categorization have been developed such as statistical [5], semantic [6], topic modeling [7], and dependency parsing [8]. The clustering approach is a method of unsupervised learning to group data based on similarity items or features. This approach is not as well-known as other unsupervised techniques. Meanwhile, several studies have shown that clustering gave promising results in terms of aspect categorization without the need for data labels.

In [12], the researcher developed a clustering algorithm to construct product feature categories. Based on the constructed feature categories, FBIOPs can be mined from the extracted implicit-opinionated clause dataset. the study strengths the current research on implicit opinion analysis for reviews in Chinese language. Another work by [13] constructed classifier to identify and predict target implicit aspects using non-negative matrix factorization (NMF). Experimental results demonstrated that the proposed approach outperforms compared to the baseline methods CR and ABSA15 dataset in English dataset.

In term of Indonesian reviews, research conducted by [14] focused on multi-label aspect classification, they used combination of CNN and Extreme Gradient Boosting (*XGBoost*). This research utilizes Indonesian hotel reviews while some aspect categories that specific for hotel domain had been determined manually. Input text will be proceeded by CNN to extract the features then the result will be an input for *XGBoost* classifier to get the final output aspect categorization. The model is quite sensitive to misspelling issues and does not handle imbalance dataset issues.

Study of [15] analyzed Indonesian Traveloka Hotel reviews, they extracted service words and opinion words using Rule-based method as well as extracted sentiment analysis. Meanwhile, sentimental terms are identified based on lexicon which are generated manually, this process also can identify negation words. Finally, the services are classified based on department or hotel functions using some machine learning methods. The result of sentiment

identification and negation problem can be solved. However, there are still incorrect results from classification tasks. Moreover, the study has shown that for verb opinion, not all verbs are opinions.

In [16], the study proposed unsupervised approach to extract explicit aspect as well as opinion word from Indonesian hotel reviews which used Rule-based generated from Regular Expressions (RE). A pair of aspects, opinion, and sentiment value are extracted based on RE then determine the aspect categorization while initial aspect seeds extracted before manually. If the sentence does not meet the regular expression, the opinion sentence will not be used. The summary of previous studies related to aspect extraction in Indonesian reviews can be seen in Table 1.

Table 1: Related studies in aspect extraction for Indonesian reviews.

Ref.	Approach	Technique	Feature	Dataset	Label	Aspect	Result
[14]	Supervised	CNN & XGBoost	-	Airy rooms (Indonesian)	Air conditioner, hot water, smell, general, cleanliness, linen, service, sunrise meal, general, television, Wi-Fi	Explicit	f1= 0.9316
[15]	Supervised	Rule-based	Important words (service, Opinion, negation)	Traveloka (Indonesian)	F&B, Security, Housekeeping, Sales, family vacation, hotel location, service, unknown.	Explicit	Precision= 0.99, recall= 1, f1= 0.99
[11]	Supervised	Association Rule, Rule based, Word Similarity	Support, confidence, initial seed aspect	Traveloka (Indonesian)	F&B, room, facility, surrounding, location, service, dan guest perspective.	Implicit	precision= 0.83, recall= 0.65, f1= 0.73
[10]	Supervised	co-occurrence matrix, Rule-based	objective features	Traveloka (Indonesian)	F&B, room, facility, service, hotel & surrounding.	Implicit	accuracy= 79.99%
[16]	unsupervised	Regular expressio, word similarity	Aspect candidate, opinion words.	Traveloka (Indonesian)	F&B, room, facility, surrounding, location, service, guest perspective.	Explicit	Precision= 0.82, recall= 0.70, f1= 0.75

Note: symbol "– " indicates the detail unreported.

Indonesian Implicit Aspect Extraction

In [10], they extracted pairs of aspect and opinion word in implicit opinion sentences on Indonesian hotel reviews using statistical approach based on Rules and co-occurrence matrix. This study proposed Indonesian complex implicit opinion sentences based on complex compound sentences, a combination of equivalent compound and complex sentence. Implicit Opinion Words are extracted using Co-occurrence Rule, it calculates the occurrence frequency of two words in one sentence which was built based on explicit opinion corpus extracted from train data which was parsed using the compound sentence rules. A compound sentence is a sentence consisting of two or more clauses. In this research, supporting aspects and data will be obtained based on word structure in sentences. However, this research needs a lot of training data to produce various rules so that it would be used to recognize various implicit opinion words. Other than that, if the clause is a complex sentence, the parsing cannot find the implicit aspect. Not all implicit aspects are found, for example Clause is a perfect sentence but has no aspect.

The work of [11] have identified an implicit aspect from Indonesian opinion sentences using combination of Rules based on Opinion Word Similarity (OWS). This research extracted implicit aspects by recognizing the relation between opinion words with aspect categories. Start by identifying a pair of opinion sentences and explicit aspect from train data, then extracting rule-based knowledge on user reviews based on co-occurrence and evaluating it using frequency and Association Rule Mining. In this step, co-occurrence matrix was developed, the co-occurrence

matrix is the frequency of occurrence of an opinion word appearing together with the aspect word in an opinion sentence. Extracting knowledge is necessary to identify opinion sentences with explicit aspects and get pairs of aspects and words of opinion with rules generated from regular expressions. Finally, in the testing phase the knowledge will be used to identify an implicit aspect of opinion sentences from test data. This research shows an escalation of result performance; however, aspect categories are extracted manually and will be expanded limited to 5 words, other than that the rule pruning is proceeded based on frequency.

We have differences from previous works. In this research, we construct unsupervised implicit aspects categorization models using a hybrid contextual embedding-based clustering and classification technique. The clustering result then was used to generate train data, we implemented two highly regarded clustering techniques i.e. K-means [17] and K-medoids [18]. We also implemented BERT as a contextual embedding model to represent the review sentences. The features used are aspect and sentiment terms which are generated from review sentences using keyBERT model and tag patterns. We expect that the combination of clustering and BERT contextual embedding will generate aspect categorization clusters that are more resilient to noise and outliers while also reducing the variance within each cluster. Finally, good clusters result as train data can increase the performance of the implicit aspect classification.

3.0 METHODOLOGY

To extract implicit aspects from Indonesian reviews based on unsupervised learning, we construct a model of hybrid contextual embedding-based clustering and classification technique. There are four main phases of the proposed model, i.e. preprocessing data, sentence features selection, generating train data based on clustering, and implicit aspect categorization. Fig 1 shows step by step the proposed model. The detail for each process is explained in the next paragraphs:

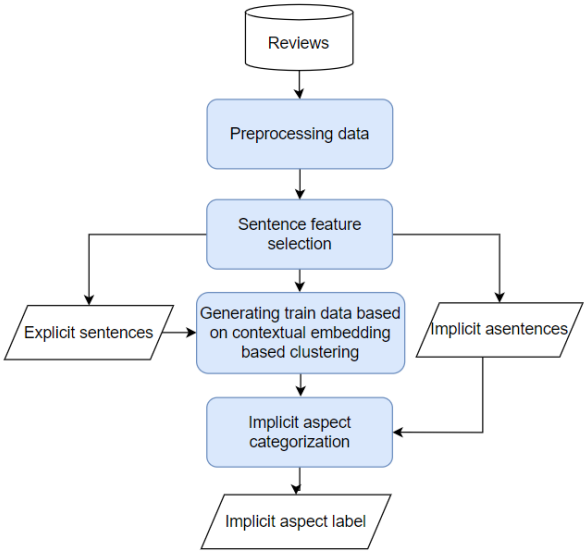

Fig. 1: The main processes of the proposed model based on hybrid contextual embedding-based clustering and classification technique.

3.1 Preprocessing Data

The first is preprocessing data, we need to prepare the data using preprocessing techniques to get cleaned row data from the original review. In [19], the study has analyzed the effect of various preprocessing techniques on the sentiment analysis model performance. We implemented preprocessing techniques include: 1) Removing Punctuation, symbols, and emoticons; 2) lowercase conversion; 3) Sentence normalization, we managed repetition letters of sentences and handled slang words by implementing Indonesian colloquial words for text normalization [20]; and 4) POS tagging, we implemented stanza library to tag POS of sentences [21]. The result for each step of preprocessing can be seen in Table 2.

Table 2: Sample of preprocessing results to produce cleaned reviews from the Indonesian original reviews.

Original	Aspect	Cleaned
Hotel ny *nyaman dan bersih* (In English: The hotel is comfortable and clean)	(implicit) kamar/room	hotel nya nyaman dan bersih
Lokasi *strategis dekat* ***dg*** *pusat kota* (In English: Strategic location close to the city center)	(explicit) Lokasi/location	lokasi strategis dekat dengan pusat kota
sky longue live music dengan panorama ***yg*** *mengesankan* (In English: Sky Longue Live Music with impressive panoramas)	(implicit) sky lounge	sky longue live music dengan panorama yang mengesankan
Cuman *shower air ga berfungsi dan channel* ***iflix+hooq*** *putus putus ga nyaman dipake nonton* (In English: It's just that the shower water doesn't work and the iflix hooq channel is broken, it's not comfortable to use to watch)	(implicit) shower	cuman shower air enggak berfungsi dan channel iflix hooq putus putus enggak nyaman dipakai menonton
Nah *ini review* ***nya :*** *Hotelnya bersih* (In English: Now this is a review of the hotel is clean)	(implicit) kamar/room	nah ini review nya hotelnya bersih
Yang *paling aku suka dari hotelnya adalah lokasinya* ***(bukan karena Sarkem ya ??)*** *karena dekat dari malioboro tapi gak crowded* (In English: What I like most about the hotel is the location is not because of sarkem yes because it is close to Malioboro but not crowded)	(explicit) Lokasi/location	yang paling aku suka dari hotelnya adalah lokasinya bukan karena sarkem ya karena dekat dari malioboro tapi enggak crowded
Fasilitas : - Kolam Renang -> *bagus* (In English: Nice swimming pool facilities)	(explicit) fasilitas/facilities	fasilitas kolam renang bagus

3.2 Sentence features selection

A review can contain one or more sentences with the same or different aspects and sentiment polarity, called multi-aspects and multi-polarities, respectively. Therefore, after we get cleaned reviews, we need to extract sentence features of reviews as input for clustering. First, we split reviews into sentences, then we extract all words represented aspects. We identified these types of words by selecting all words tagged as NOUN and proper nouns (PROPN). Furthermore, using a rule-based we classified sentences into two groups: explicit and implicit sentences. The group of explicit sentences contain sentences with explicit aspect categories, for example, "*kamarnya bersih dan nyaman*" (in English: The room is clean and comfortable), we classify this sentence as explicit sentence because it contains the explicit category "*kamar*" (room). Meanwhile, the sentences that do not contain aspect categories explicitly are classified as implicit sentences, for example, "*hotelnya strategis dan dekat dengan pasar tradisional*" (in English: the hotel is strategic and close to the local market), this sentence talks about the hotel's location but the "*lokasi"* (location) aspect does not appear explicitly in the sentence.

To represent this condition, we use rule-based while S is a sentence and $L = \{t_1, t_2, \ldots, t_n\}$ is the aspect categories, therefore $I(S, L)$ represents the indicator function that returns 1 if the sentence S contains any term from L and 0 otherwise. The formula can be written as (1).

$$I(S, L) = \begin{cases} 1, if\ \exists t \in L\ such\ that\ t \in S \\ 0, \qquad\qquad otherwise. \end{cases} \quad (1)$$

The formula (1) will check whether at least one aspect category t from list L exists in sentence S. The formula will produce output 1 if the condition is true, and output 0 if the condition is false. The sentences with output 1 are grouped as explicit sentences, while the group of implicit sentences for vice versa. Total sentences with explicit aspect are 1315 while total implicit sentences are 2365. In the next phase, the explicit sentences will be labelled automatically using clustering technique. Finally, the selected sentences are classified as explicit sentences which are used as input in clustering for generating train data.

3.3 Generating train data based on contextual embedding based clustering

In generating train data, we used a contextual embedding-based clustering algorithm which combines clustering concept and contextual approach. The goal of clustering is to group and to label explicit sentences based on the category aspects. We adopted two well-known clustering techniques (i.e. K-means and K-medoids). K-means algorithm is adopted for clustering sentence embeddings, and K-medoids is adopted to initialize centroid positions [17]. Meanwhile, for contextual approach, we implemented BERT embeddings to transform explicit sentences and aspect categories into the embedding format which is used as input points for clustering. The benefits of BERT

include automated contextual feature generation, fewer data requirements, faster development, and improved performance [22].

A combination of clustering, especially the k-means algorithm, and embedding model has been conducted in previous studies such as [23], [24], [25], [26]. In this research, we specifically implemented clustering concept and contextual embedding model for aspect categorization task. This clustering technique works by calculating the similarity between the items and organizing them into clusters based on these similarity scores. The process to produce clusters from explicit sentences using contextual embedding-based clustering contains four steps. The detailed processes of clustering technique used in this research can be seen in Fig. 2.

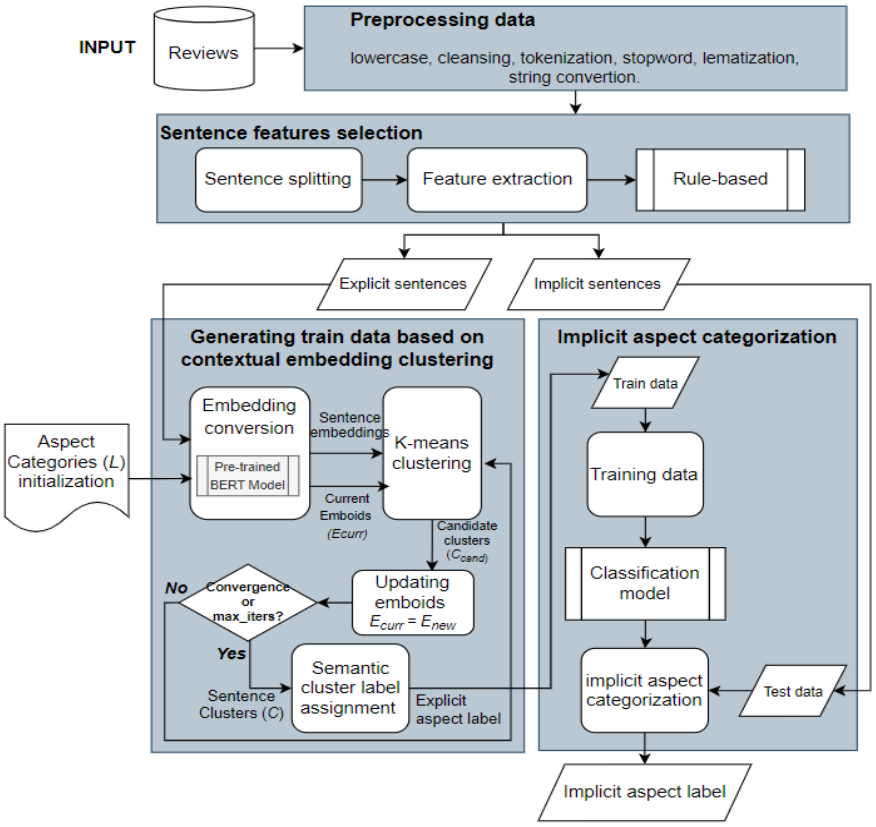

Fig. 2: The proposed hybrid contextual embedding-based clustering and classification technique for implicit aspect categorization model.

The first step is embedding conversion, a process for transformation of explicit sentences into embedding format. We implemented a sentence embedding mechanism to generate embedding by utilizing the Pre-trained Sentence-BERT Model, a variant of the original BERT (Bidirectional Encoder Representations from Transformers), as described in [27].

The second step is initialization of centroid positions (called *emboids*) based on aspect category embeddings. We transformed each aspect category into vectors using the pre-trained Sentence-BERT model following the same approach used for converting explicit sentence embeddings. This initialized *emboids* is called *current emboids (Ecurr)*.

The third step is clustering, the model clustered the explicit sentence embeddings based on the similarity with the *current emboids* using contextual embedding-based clustering technique. For clustering technique, K-means is adopted for clustering the sentence embeddings [17]. On the other hand, K-medoids is adopted to initialize centroid positions [18]. The similarity between sentence embeddings and the *emboids* is calculated based on Euclidean distance referring to equation (2) and (3).

$$C_i = argmin_k \, d(r_i, e_k) \tag{2}$$

$$d(r_i, e_k) = \sqrt{\sum_{j=1}^{m}(r_{ij}, -e_{kj})^2} \quad (3)$$

where r_i is the value of the *i*-th sentence's embedding, C_i is the cluster assignment for point r_i, and e_k is the *emboid* of cluster *k*, while *m* is the number of dimensions of the embeddings. In this research, we implemented SBERT which has 1x384 dimensional embeddings.

We updated the *current emboids* to obtain robust clusters. For updating *emboids,* we need to calculate the *new emboids* by taking *mean* of all sentence embeddings assigned to each candidate cluster. Repeating the assignment and updating steps until either convergence or maximum number of iterations is reached. In this research, center clusters are updated when clusters have not fulfilled one of two conditions: if the convergence criteria have not met the tolerance or if the iteration score is still under the maximum score. We set a maximum iteration score of 100 while the tolerance $\alpha=e^{-4}$. The updated *emboids* is calculated by referring equation (4), where the convergence clusters can be expressed as Equation (5):

$$e_k = \frac{1}{|c_k|}\sum_{c_i \in c_k} r_i \quad (4)$$

where r_i is the value of the *i*-th sentence's embedding, C_i is the cluster assignment for point r_i, and e_k is the updated *emboid* of c_k, while c_k is the set of averaging vectors r_i assigned to cluster *k*, and $|c_k|$ is the number of r_i in cluster *k*. Finally, we have clusters containing the sentence embeddings which were grouped by measuring the magnitude of the difference using Euclidean distance.

$$\delta = \sum_{k=1}^{K} d(e_k^{(t)} - e_k^{(t-1)}) \quad (5)$$

The δ is a convergence cluster, where $e_k^{(t)}$ is the *emboid* at iteration t, and $d(e_k^{(t)} - e_k^{(t-1)})$ is the Euclidean distance between the *current* and *new emboids*. If $\delta < \alpha$ then the algorithm has converged.

The fourth step is clusters labelling, the output of the proposed clustering technique is not only produced clusters but also assignment the label of cluster. The concept for labelling is based on the maximum similarity between the members of the cluster and the aspect category. We adopted the cosine similarity formula from the study of [28] to calculate the similarity of two vectors between the sentence embeddings and aspect category embeddings. Furthermore, calculating the average of the cluster members weight to the aspect categories refers to equation (6). The algorithm calculated T_{kj}, the average of weight for aspect category j in cluster *k*, while W_{ij} is the weight of member *i* in cluster *k* for aspect category *j*, and $|m_k|$ is the number of cluster members in the cluster *k*. Finally, the algorithm selected an appropriate label for each cluster based on maximum cluster score.

$$T_{kj} = \left(\frac{1}{|m_k|}\sum_{i=1}^{|m_k|} W_{ij}\right) \quad (6)$$

3.4 Implicit aspect categorization

After we get explicit aspect labels based on the proposed clustering technique, we predict implicit aspect labels of implicit sentences based on classification technique. In this research, we implemented several classification techniques to get the best technique for implicit aspect categorization. To predict implicit aspect categories based on train data generated from contextual embedding based clustering, we test them which are integrated with contextual embedding based clustering and then get the best technique for implicit aspect categorization. For implementing classification techniques to the implicit aspect categorization model, we used scikit-learn machine learning libraries[1] for python.

Logistic Regression (LR) is a prominent linear model used for binary and multi-class classification tasks. In text classification, the algorithm predicts the probability of a given text belonging to a specific category based on features extracted from the text [29]. LR models the conditional probability as equation (3) where *x* is the data, *y* is the class label, and $w \in R^n$ is the weight vector:

$$P_w(y = \pm 1|x) \equiv \frac{1}{1+e^{-yw^T X}} \quad (3)$$

Support Vector Machine (SVM) is a supervised technique used for text classification tasks that are effective for high-dimensional spaces[30]. SVM aims to find a hyperplane that best separates the classes in the feature space. It maximizes the margin (distance) between the hyperplane and the nearest data points of each class, known as support vectors. The formula of support vector machine can be seen in equation (4), where *x* is input feature vector, x_i is support vectors, y_i is class labels (+1 or -1), α_i is learned weights (*Lagrange multipliers*), *K* represents kernel function (e.g., linear, polynomial, RBF), and *b* is a bias term. The predicted class is given by $\hat{y} = sign(f(x))$.

$$f(x) = \sum_{i=0}^{n} \alpha_i y_i K(x_i, x) + b \quad (4)$$

[1] https://scikit-learn.org/stable/supervised_learning.html

Multinomial Naive Bayes (MNB) is a supervised-learning algorithm that is the simplest form of a Bayesian network [31]. This algorithm works based on Bayes' theorem with the "naive" assumption of conditional independence, where all attributes are independent given the value of the class variable. Multinomial NB classifier captures the term frequency of the documents [32]. This model follows Bayes' theorem which refers to equation (5), where c is a class variable and x_1 to x_n dependent feature vector in document d.

$$P(c, x_i) = \frac{P(x_i,c)\cdot P(c)}{P(x_i)} \tag{5}$$

Decision Tree (DT) is a supervised machine learning algorithm which divides the feature space into regions by recursively splitting the data based on feature values, ultimately assigning a class label to each region[33], [34]. In general, Decision Tree splits the data using entropy or Gini impurity at each node to maximize class separation. In this research, we implemented the sklearn library, so we used Gini impurity that default technique of splitting data for decision trees. The Gini impurity can be calculated using the equation (6), where p_i is the probability of a sample belonging to class i.

$$Gini = 1 - \sum p_i^2 \tag{6}$$

Random forest (RF) is a machine learning technique that generalize error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them [35]. RF operates as an ensemble method by constructing a multitude of decision trees during training and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. A Random Forest classifier makes predictions using an ensemble of decision trees. The general formula for its prediction can be calculated using equation (7), Where $f_i(x)$ is the prediction from the i-th decision tree, N is the total number of trees, and the final prediction is made by majority voting.

$$\hat{y} = majority\ vote(f_i(x), f_2(x), .., f_N(x)) \tag{7}$$

4.0 RESULT

In this research, we evaluated the proposed implicit aspect categorization model using automatic evaluation metrics to measure the performance of the model. The proposed model combines a contextual embedding-based clustering and classification technique. Therefore, we need to evaluate the results of the two essential processes of the model i.e. clustering and classification results. The description of dataset, evaluation metrics, experiment results and discussion are presented in Sub-Section 4.1, 4.2, and 4.3 respectively.

4.1 Dataset

The dataset used in the research is an Indonesian review in hotel domain, it is a secondary data which has not been published before. We obtained the data in an Excel format from the first researcher who collects the data directly from the Tripadvisor website[2]. We then annotated the data based on four aspect categories by involving three annotators. Total data used is 3680 reviews which are grouped into four aspect categories, i.e. room, service, facilities, and location. To the best of our knowledge there is no secondary data publicly available for implicit aspect categorization in Indonesian reviews. Therefore, we could not text the proposed technique on other domains or datasets to evaluate the model's generalizability.

4.2 Evaluation metrics

To measure the performance of the model, we implement some evaluation metrics which generally used to evaluate clustering and classification results. We used some python codes libraries provided by *Scikit-learn* for implementing evaluation metrics. Some common methods used for evaluation i.e. accuracy, F1-score, precision, and recall. Additionally, we used another method to asses clustering result.

Clustering is an unsupervised learning task and lacks predefined labels for evaluation, we used some common technique to evaluate clustering result, i.e. 1) Internal Evaluation Metrics, evaluating the quality of the clustering based on the data itself without requiring external labels, we used Silhouette Score, one of techniques which is used to measure the quality of the clusters [36]. The score ranges from -1 to 1, with a higher value (close to 1) indicating better-defined clusters; 2) External Evaluation Metrics, require ground truth labels to compare the clustering results, we use Adjusted Rand Index (ARI) and Fowlkes-Mallows Index (FMI). ARI is a technique to measures the similarity between the predicted clusters and the true labels, adjusting for the chance grouping. ARI ranges from -1 to 1, with 1 indicating perfect clustering[37]. Meanwhile, FMI is a technique to evaluates the similarity between the clusters and the true labels using precision and recall [38]. Values range from 0 to 1, with 1 indicating perfect clustering; and 3) Stability and Consistency Metrics, evaluating the robustness of the clustering algorithm, we use

[2] www.tripadvisor.com

Jaccard Index, a technique to measures the similarity between different sets of clusters obtained from different runs of the clustering algorithm.

4.3 Experiment Result and Discussion

In this research, we designed two experiments to measure the proposed model's performance. The first experiment, we measure the performance of the clusters result, we compare the generated clusters between the contextual embedding-based clustering technique with the baselines. The baselines include K-means algorithm, a widely recognized clustering technique. Other than that, we also use K-medoids and Agglomerative clustering for comparison. The second experiment, we combine the explicit aspect labels generated by contextual embedding-based clustering with various classification techniques to select the appropriate classification technique for the best model in implicit aspect categorization. The result of each experiment is described as follows:

Experiment #1. We evaluate the quality of clusters generated by contextual embedding based clustering and assess the result of clusters labelling automatically. We used accuracy to assess the generated labels of explicit sentences. Meanwhile, to measure the quality of the clusters, we implemented some evaluation metrics, i.e. Adjusted Rand Index (ARI), Fowlkes-Mallows Index (FMI), and Jaccard Index, and silhouette score.

For clustering experiment, the parameter setting includes maximum iteration (*max_iter*=100), convergence threshold (α=*e-4*), number of clusters (*k*=4), and sentence embedding model (we used the Bert model). Meanwhile, for sentence embedding we implemented the *Sentence Transformer* class or SBERT which used *MiniLM*, an efficient version of BERT. This model comprises 6 transformer layers, making it faster and more resource-efficient where each sentence is mapped to a 384-dimensional embedding. The text data was transformed into sentence embeddings using the *encode* method and the default pooling operation is a mean pooling.

Table 4 and 5 show the evaluation result of explicit aspect labels and clusters result respectively. We compared the results between contextual embedding-based clustering technique results and the baselines. Automatic labels measurement using contextual embedding-based clustering technique produced 71.18% accuracy, while the labeling accuracy results with K-medoids algorithm are higher at 74.60%. However, for clustering results, the contextual embedding-based clustering technique shows good performance compared to baselines referred to the silhouette score. The silhouette score ranges from -1 to 1, with a higher value (close to 1) indicating better-defined clusters. Meanwhile, the silhouette score ranges from 0 to 0.5 (0 < silhouette score < 0.5), it indicates that clusters are reasonable but could be improved, and when the silhouette score under 0 indicates the clusters are not good enough. The silhouette score of contextual embedding-based clustering presents 0.1062, it is higher than other cluster techniques. Our silhouette scores around 0 indicate overlapping clusters, the clustering is reasonable but could be improved.

Table 3: Explicit aspect labels assessment (comparison of accuracy of each technique for labelling result).

Technique	Accuracy (%)
K-means algorithm	61.67
K-medoids algorithm	**74.60**
Agglomerative algorithm	34.37
Contextual embedding-based clustering	**71.18**

Table 4: Clusters assessment (Performance comparison between contextual embedding-based clustering and the baselines based on silhouette score).

Technique	Silhouette
K-means algorithm	0.0606
K-medoids algorithm	0.0498
Agglomerative algorithm	-1
Contextual embedding-based clustering	0.1062

We also measured the performance of clusters using alternative metrics besides silhouette score. The other evaluation metrics used are external evaluation metrics (i.e. Adjusted Rand Index (ARI), Fowlkes-Mallows Index (FMI)), external evaluation metrics (i.e. Silhouette score), and stability and consistency metrics (i.e. Jaccard Index). Referring to Table 5 which displays the cluster quality measurement result, the cluster quality generated by contextual embedding-based clustering technique gets the score of ARI, FMI, and Jaccard Index in 0.5170,

0.6549, and 0.5201 respectively. It proves that our clustering technique can produce good clusters proven by a metric score above 0.5, except the result of silhouette score is under 0.5 which represents the clusters is reasonable but could be improved.

Table 5: Cluster quality measurement results

Metrics		Score	Representation
External Evaluation	ARI	0.5170	The clusters are good
	FMI	0.6549	The clusters are good
Internal Evaluation	Silhouette	0.1062	The clusters are reasonable but could be improved
Stability and Consistency	Jaccard Index	0.5201	The clusters are good

The final result of the clustering using a contextual embedding-based clustering technique is labelled clusters based on aspect categories. The clusters are generated from explicit sentences which contain 1315 sentences. The generated labelled clusters will be used as train data on the classification process. The distribution of the aspect categorization labels generated by contextual embedding-based clustering technique shown in Fig 3.

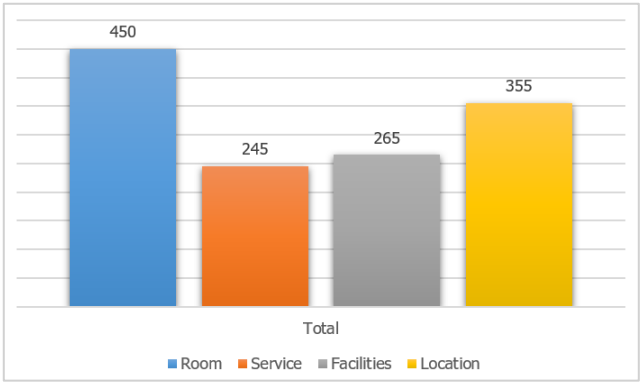

Fig. 3: The distribution of labelled clusters generated by contextual embedding-based clustering technique.

Experiment #2. We experimented by trying several classification techniques in the proposed model. It aims to get the best combination of contextual embedding-based clustering and appropriate classification technique. The classification techniques experimented in this research are Logistic Regression (LR), Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (DT), and Random Forest (RF). Furthermore, we measure the performance of the model for each combination classification technique. The evaluation metrics used in this experiment are accuracy, F1 score, recall, and precision.

Table 6 displays the experiment result of the combination of contextual embedding-based clustering and various classification techniques. From the table, we can see that Random Forest (RF) algorithm produces accuracy in 72.04%, it is higher than other classification techniques. This result is in line with the F1 score, recall, and precision of RF, which is also higher than other classification techniques, with the score is in 0.6788, 0.6634, and 0.7022 respectively. Based on this experiment, we got the best version of the proposed implicit aspect categorization model by combining contextual embedding-based clustering and Random Forest classification.

Table 6: Comparison of classification techniques for implicit aspects categorization

Clustering	Classification	Accuracy (%)	F1	Recall	Precision
Contextual embedding-based clustering	LR	52.85	0.5382	0.5962	0.6049
	SVM	54.57	0.5509	0.5986	0.6005
	NB	55.41	0.5438	0.5877	0.5864
	DT	49.46	0.5340	0.5329	0.6848
	RF	**72.04**	**0.6788**	**0.6634**	**0.7022**

Note: The best results are in bold

Furthermore, we experimented with clustering techniques which are combined with Random Forest technique. Experimental results demonstrate that our model outperforms other approaches that combine clustering techniques with Random Forest (RF). The complete results of the experiment can be seen in Table 7.

Table 7: Experiment result of the proposed model compared to other approaches (various clustering techniques combined with RF)

Clustering		Classification	Accuracy (%)	F1	Recall	Precision
Baseline	K-means	RF	39.76	0.3750	0.4178	0.5402
	K-medoids		61.41	0.6045	0.6418	0.6257
	Agglomerative		15.46	0.1062	0.2099	0.3830
Proposed	Contextual embedding-based clustering		**72.04**	**0.6788**	**0.6634**	**0.7022**

Note: The best results are in bold

Next, we display the results of the aspect categorization using the proposed model. We get the best model of implicit aspect categorization based on experiment #2 using hybrid contextual embedding-based clustering and Random Forest (RF) classification technique. Furthermore, using the proposed model we get prediction result of implicit aspect categorization. We present 10 samples of the implicit aspect categorization predicted by the proposed model compare to the experts' labels which displays on the Table 8. Sentence number 1, 3, 4, 6, and 7 show the correct result of the implicit aspect category labels generated by the model compare to the experts' labels. Refered to the sentence 1 and 4, the model can correctly categorized implicit aspects "*tempat*" (place) and "*parkir valet*" to the *service* and *location* labels respectively. Likewise, the sentence 7 which has two aspects "*kolam renang*" (pool) and "*sky lounge*" can also be predicted correctly to the *facilities* label. The proposed model also demonstrated its ability to correctly predict categories for sentences with no aspect, as shown by sentence 3 with the aspect category *room* and sentence 6 with the aspect category *location*.

Table 8: Sample of implicit aspect categorization result (expert vs system)

No	Sentence	Implicit aspect	Aspect category	
			By Expert	By system
1	"***tempat*** *strategis*" (in English: strategic location)	*Tempat* (in English: place)	location	location
2	*"dan* ***sarapan*** *yang variatif dan enak"* (in English: and the breakfast is variation and tasty)	*Sarapan* (in English: breakfast)	service	facilities
3	"*berisik tidak ada kedap suara nya*" (in English: it's noisy, no soundproofing)	No aspect	room	room
4	"***parkir valet*** *nya yang tidak siap selalu harus disamperin di pos security*" (in English: Parking valet is not ready, always must be ampererized at the security post)	Parkir valet	service	service
5	"***sky longue live*** *music dengan panorama yang mengesankan*" (in English: Sky Longue Live Music with impressive panoramas)	Sky lounge	facilities	service
6	"*tinggal nyebrang dari stasiun saja*" (in English: Just cross from the station)	No aspect	location	location
7	"***kolam renang*** *juga bagus dan* ***Sky Lounge*** *juga sangat bagus*" (in English: the swimming pool is also nice, and the Sky Lounge is also very nice)	*Kolam renang*, sky lounge (in English: pool, sky lounge)	facilities	facilities
8	"*tapi sayang nya* ***dispenser*** *nya agak jauh dari kamar"* (in English: But unfortunately, the dispenser is a bit far from the room)	Dispenser	facilities	room

9	"*tidak disediakan* ***minuman kemasan botol*** *plastik dalam kamar*" (in English: No plastic bottled drinks are provided in the room)	*Minuman kemasan* (in English: bottled water)	facilities	room
10	"***letak*** *geografisnya sempurna*" (in English: its geographical location is perfect)	*Letak* (in English: location)	location	service

We reported there are misclassifications in implicit aspect category labeling. Several sentences can not be categorized correctly by the model, such as in the sentence 2, 5, 8, 9, and 10, include predicting the label *facilities* as *service*, *room* as *facilities*, and *service* as *location*. Refer to the confusion matrix (see Fig. 4), the largest number of errors occurred in predicting the aspect category *service* for the true *facilities* label, this error is represented at sentence 4 in Table 8. Total error for *facilities* as *service* labels is 352 of 3680 data (9.6%), it is an interesting point for further analysis.

In sentence 5, the implicit aspect is *"sky lounge"* which should be included in the *facilities* label but the model are incorrectly grouped into the *service* label. We analyzed the majority of the sentence with "*sky lounge*" aspect which tends to be predicted as a *service* label than other labels. However, this statement does not apply to certain sentences, such as sentence 7. Although this sentence contains the aspect "*sky lounge*", the model predicts it under the *facilities* label. Our analysis suggests that this happens because the sentence also includes another aspect, "*kolam renang*" (pool), in addition to "*sky lounge*". It seems that the aspect "kolam renang" has a stronger relationship with the *facilities* compared to "*sky lounge*", which is related to the *service*. As a result, the model ultimately predicts this sentence under the *facilities* label rather than the *service* label.

On the other hand, in the sentence 2, with the implicit aspect "*sarapan*" (breakfast), was predicted opposite to sentence 4; this sentence was classified under the *service* label instead of the true *facilities* label. This error occurred in 182 sentences (5% of the total sentences). Another interesting case is in the sentence 8 and 9, both sentences are missclassified to the *room* label instead of the true *facilities* label. We argue the error comes when the sentences contain explicit term "*kamar*" (room) then the model tends to classify sentences to the *room* label. Whereas, the true aspects of the sentences are "*dispenser*" and "*minuman kemasan*" (bottled water) respectively that represent facilities category. Meanwhile, we assess for misclassification of the sentence 10 which is labeled as *service* from true *location* label due to the sentence contains "*sempurna*" (perfect) term. This term is stronger in representing *service* than *location*. So the model classifies this sentence that have implicit aspect "*letak*" (location) into the label *service* rather than *location*.

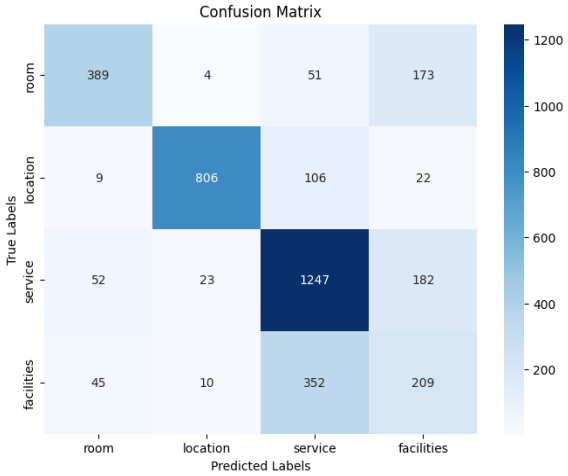

Fig. 4: Confusion matrix of the proposed implicit aspect categorization model

Our model succeeded in categorizing implicit aspects for Indonesian reviews without labelled data by combining contextual embedding-based clustering and Random Forest technique. However, it still needs to increase the silhouette score, a general evaluation matrix in clustering, to produce robust train data. Other than that, in classification process, in some cases, the model is still misclassified the sentences into the true category labels. We assume the model is still not able to recognize important terms that are considered aspects of sentences (see sentence 8 and 9 in Table 5). Therefore, to improve clustering performance, we recommend exploring embedding techniques to better recognize the context of sentences, so that the resulting clusters can be probably improved significantly. Meanwhile, to minimize misclassification aspect category labels, we suggest exploring classification

techniques to recognize sentence context. Deep learning such as BERT is recommended for future work to increase the accuracy of the model.

To the best of our knowledge, there are no automated methods for this task (implicit aspect categorization) that can be used to enhance clarity and provide a comprehensive evaluation of the model's performance. Therefore, we cannot compare the proposed technique result with the automated methods result.

5.0 CONCLUSION

Implicit aspect categorization model is proposed through a hybrid contextual embedding-based clustering and classification technique. We developed the model using an unsupervised learning approach so that there is no need to labelled data in training. Contextual embedding-based clustering generated train data from explicit sentences which will be used to classify implicit aspect categorization. We experiment with several classification techniques, i.e. Logistic Regression (LR), Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (DT), and Random Forest (RF), to get the best combination of the proposed contextual embedding based clustering and appropriate classification technique. Based on the experiment, the combination of contextual embedding based clustering and RF algorithm produces higher accuracy than other classification techniques, with accuracy tent to 72.04%. However, the model is still misclassified the sentences into the true category labels, it is still not able to recognize important terms that are considered aspects of sentences. For future work, exploring classification techniques is needed to minimize misclassification aspect category labels. Implementing deep learning such as BERT is recommended to increase the accuracy of the model.

ACKNOWLEDGEMENT

This work is supported by: Kementerian Pengajian Tinggi Malaysia, Fundamental Research Grant Scheme (FRGS) by code number FRGS/1/2020/ICT02/UMS/02/2; Language Engineering and Application Development (LEAD) research group of Faculty of Computing and Informatics, Universiti Malaysia Sabah; and Lembaga Pengembangan Publikasi Ilmiah (LPPI) University of Muhammadiyah Malang (UMM), Indonesia.

REFERENCES

[1] H. Kim and K. Ganesan, "Comprehensive review of opinion summarization," *Illinois Environment*, pp. 1–30, 2011.

[2] W. Wang and S. J. Pan, "Syntactically Meaningful and Transferable Recursive Neural Networks for Aspect and Opinion Extraction," *Comput. Linguist.*, vol. 45, no. 4, pp. 705–736, Jan. 2020, doi: 10.1162/coli_a_00362.

[3] H. Cai, R. Xia, and J. Yu, "Aspect-Category-Opinion-Sentiment Quadruple Extraction with Implicit Aspects and Opinions," in *Proceedings ofthe 59th Annual Meeting ofthe Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Association for Computational Linguistics, 2021, pp. 340–350. doi: 10.18653/v1/2021.acl-long.29.

[4] J. Z. Maitama, N. Idris, A. Abdi, L. Shuib, and R. Fauzi, "A systematic review on implicit and explicit aspect extraction in sentiment analysis," *IEEE Access*, vol. 8, no. November, pp. 194166–194191, 2020, doi: 10.1109/ACCESS.2020.3031217.

[5] M. D. S. Nandhini and G. Pradeep, "A Hybrid Co-occurrence and Ranking-based Approach for Detection of Implicit Aspects in Aspect-Based Sentiment Analysis," *SN Comput Sci*, vol. 1, no. 3, p. 128, May 2020, doi: 10.1007/s42979-020-00138-7.

[6] T. A. Rana, Y.-N. Cheah, and T. Rana, "Multi-level knowledge-based approach for implicit aspect identification," *Applied Intelligence*, vol. 50, no. 12, pp. 4616–4630, 2020, doi: 10.1007/s10489-020-01817-x.

[7] O. M. AL-Janabi, N. H. Ahamed Hassain Malim, and Y. N. Cheah, "Unsupervised model for aspect categorization and implicit aspect extraction," *Knowl Inf Syst*, vol. 64, no. 6, pp. 1625–1651, 2022, doi: 10.1007/s10115-022-01678-5.

[8] J. Z. Maitama, N. Idris, A. Abdi, L. Shuib, and R. Fauzi, "A systematic review on implicit and explicit aspect extraction in sentiment analysis," *IEEE Access*, vol. 8, no. November, pp. 194166–194191, 2020, doi: 10.1109/ACCESS.2020.3031217.

[9] T. A. Rana, Y.-N. Cheah, and T. Rana, "Multi-level knowledge-based approach for implicit aspect identification," *Applied Intelligence*, vol. 50, no. 12, pp. 4616–4630, 2020, doi: 10.1007/s10489-020-01817-x.

[10] Y. Setiowati, A. Djunaidy, and D. O. Siahaan, "Pair Extraction of Aspect and Implicit Opinion Word based on its Co-occurrence in Corpus of Bahasa Indonesia," in *2019 2nd International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2019*, IEEE, 2019, pp. 73–78. doi: 10.1109/ISRITI48646.2019.9034672.

[11] Y. Setiowati, F. Setyorini, and A. Helen, "Penentuan Aspek Implisit dengan Ekstraksi Knowledge Berbasis Rule pada Ulasan Bahasa Indonesia (Determination of Implicit Aspects with Rule Based Knowledge Extraction in Indonesian Reviews)," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 9, no. 1, pp. 35–44, 2020, doi: 10.22146/jnteti.v9i1.145.

[12] Z. Fang, Q. Zhang, X. Tang, A. Wang, and C. Baron, "An implicit opinion analysis model based on feature-based implicit opinion patterns," *Artif Intell Rev*, vol. 53, no. 6, pp. 4547–4574, 2020, doi: 10.1007/s10462-019-09801-9.

[13] Q. Xu, L. Zhu, T. Dai, L. Guo, and S. Cao, "Non-negative matrix factorization for implicit aspect identification," *J Ambient Intell Humaniz Comput*, vol. 11, no. 7, pp. 2683–2699, 2020, doi: 10.1007/s12652-019-01328-9.

[14] A. N. Azhar, M. L. Khodra, and A. P. Sutiono, "Multi-label Aspect Categorization with Convolutional Neural Networks and Extreme Gradient Boosting," *Proceedings of the International Conference on Electrical Engineering and Informatics*, vol. 2019-July, no. July, pp. 35–40, 2019, doi: 10.1109/ICEEI47359.2019.8988898.

[15] Y. Setiowati, "Service Extraction and Sentiment Analysis to Indicate Hotel Service Quality in Yogyakarta based on User Opinion," *2018 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, pp. 427–432, 2018.

[16] Y. Setiowati, F. Setyorini, and A. Helen, "Aspect and Opinion Word Extraction on Opinion Sentences in Bahasa Indonesia using RUle Based Generated from Regular Expression," *Internatioal Conference on Information Technology, Information Systems and Elctrical Engineering (ICITISEE)*, vol. 01, no. 01, pp. 1689–1699, 2018.

[17] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *J R Stat Soc Ser C Appl Stat*, vol. 28, no. 1, pp. 100–108, 1979.

[18] K. L. and R. P., "Clustering by means of Medoids," 1987.

[19] A. I. Kadhim, "An Evaluation of Preprocessing Techniques for Text Classification," *International Journal of Computer Science and Information Security*, vol. 16, no. 6, pp. 22–32, 2018, [Online]. Available: https://sites.google.com/site/ijcsis/

[20] S. N. Aliyah, W. Y. Ardhito, S. A. Akbar, and J. Ade, "Colloquial Indonesian Lexicon," in *2018 International Conference on Asian Language Processing (IALP)*, IEEE, 2018. Accessed: Nov. 21, 2024. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8629151

[21] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020. [Online]. Available: https://nlp.stanford.edu/pubs/qi2020stanza.pdf

[22] M. S. I. Malik, U. Cheema, and D. I. Ignatov, "Contextual Embeddings based on Fine-tuned Urdu-BERT for Urdu threatening content and target identification," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 7, p. 101606, 2023, doi: 10.1016/j.jksuci.2023.101606.

[23] N. Reimers, B. Schiller, T. Beck, J. Daxenberger, C. Stab, and I. Gurevych, "Classification and Clustering of Arguments with Contextualized Word Embeddings," 2018.

[24] Y. Li, J. Cai, and J. Wang, “A Text Document Clustering Method Based on Weighted BERT Model,” no. Itnec, pp. 1426–1430, 2020.

[25] A. Subakti, H. Murfi, and N. Hariadi, “The performance of BERT as data representation of text clustering,” *J Big Data*, vol. 9, no. 1, 2022, doi: 10.1186/s40537-022-00564-9.

[26] L. George and P. Sumathy, “An integrated clustering and BERT framework for improved topic modeling,” *International Journal of Information Technology (Singapore)*, vol. 15, no. 4, pp. 2187–2195, 2023, doi: 10.1007/s41870-023-01268-w.

[27] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using siamese BERT-networks,” *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pp. 3982–3992, 2019, doi: 10.18653/v1/d19-1410.

[28] F. Rahutomo, T. Kitasuka, and M. Aritsugi, “Semantic Cosine Similarity,” *The 7th international student conference on advanced science and technology ICAST*, vol. 4, no. 1, pp. 4–5, 2012.

[29] H. F. Yu, F. L. Huang, and C. J. Lin, “Dual coordinate descent methods for logistic regression and maximum entropy models,” *Mach Learn*, vol. 85, no. 1–2, pp. 41–75, Oct. 2011, doi: 10.1007/S10994-010-5221-8/METRICS.

[30] T. Joachims, “Text categorization with Support Vector Machines: Learning with many relevant features,” pp. 137–142, 1998, doi: 10.1007/BFB0026683.

[31] J. Chen, H. Huang, S. Tian, and Y. Qu, “Feature selection for text classification with Naïve Bayes,” *Expert Syst Appl*, vol. 36, no. 3 PART 1, pp. 5432–5435, 2009, doi: 10.1016/j.eswa.2008.06.054.

[32] D. H. Abd, A. T. Sadiq, and A. R. Abbas, *Political Articles Categorization Based on Different Naïve Bayes Models*, vol. 1174 CCIS. Springer International Publishing, 2020. doi: 10.1007/978-3-030-38752-5_23.

[33] J. R. Quinlan, “Induction of decision trees,” *Machine Learning 1986 1:1*, vol. 1, no. 1, pp. 81–106, Mar. 1986, doi: 10.1007/BF00116251.

[34] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, “Classification and regression trees,” *Classification and Regression Trees*, pp. 1–358, Jan. 2017, doi: 10.1201/9781315139470/CLASSIFICATION-REGRESSION-TREES-LEO-BREIMAN-JEROME-FRIEDMAN-OLSHEN-CHARLES-STONE/ACCESSIBILITY-INFORMATION.

[35] L. Breiman, “Random forests,” *Mach Learn*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324/METRICS.

[36] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *J Comput Appl Math*, vol. 20, no. C, pp. 53–65, 1987, doi: 10.1016/0377-0427(87)90125-7.

[37] L. Hubert and P. Arabie, “Comparing partitions,” *J Classif*, vol. 2, pp. 193–218, 1985, Accessed: Dec. 11, 2024. [Online]. Available: https://sci-hub.ru/10.1007/bf01908075

[38] E. B. Fowlkes and C. L. Mallows, “A method for comparing two hierarchical clusterings,” *J Am Stat Assoc*, vol. 78, no. 383, pp. 553–569, 1983, doi: 10.1080/01621459.1983.10478008.