Advances in Economics, Business and Management Research, volume 202
3rd International Conference on Business and Management of Technology (ICONBMT 2021)

Sentiment Analysis Based on The Aspect of Culinary and Restaurant Review Using Latent Dirichlet Allocation and Support Vector Machine to Improve the Profitability of Culinary Business and Restaurant in Surabaya

Drajad Bima Ajipangestu[1*], Riyanarto Sarno[1*]

[1] *Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia*
[*]*Corresponding author. Email: bimadrajad@gmail.com, riyanarto@if.its.ac.id*

ABSTRACT
Food is always close to us, along with the growing population, the food business will continue to grow. During the Covid-19 Pandemic, even though many people stay at home, they occasionally want to buy their favorite food outside the home, whether it's buying directly onsite or online. The aspect categorization is carried out by combining the LDA method and Semantic Similarity to categorize tweets into four culinary aspects (1) Price, (2) Taste, (3) Place, and (4) Service. The best performance of aspect categorization is by combining the LDA and TF-ICF 100% for term extension. Next, the classification stage with Word Embedding to extract features using GloVe and SVM with three-parameter modifications of the SVC method: (1) C-SVC, (2) Linear SVC, and (3) SVCnu with various kernel changes to get the best results. Then an increase in classification accuracy is carried out using SentiCircle. The results of this study show that aspects of service (4) Has a review with a high negative sentiment that reaches 10.869% compared to sentiment reviews on other aspects in percent (price: 4.348, taste: 6.522, place: 4.521) so that business owners culinary needs to make improvements to pay more attention to customer service to reduce the number of negative reviews on this service aspect. The results also show that changes in sentiment (on positive or negative sentiment) are influenced by the aspect of each review.

Keywords: *Twitter, LDA, SVM, Aspect Categorization, Sentimen Classification*

1. INTRODUCTION

The food and beverage business are indeed very interesting to study. Some of the reasons that make this business very attractive include a fairly large and highly expected profit margin, as well as a very broad market. Various outlets that offer food and beverage products in various forms have sprung up. Ranging from simple to luxurious. Many people read reviews to see how people think about the price of the food, what the menu is, whether the food is delicious, whether the place is comfortable, to how the service from the restaurant is. With the increase in the culinary and restaurant businesses, the competition has intensified. Many restaurants and culinary places are competing to win the hearts of customers in various ways.

This study conducted a sentiment analysis based on predetermined aspects of customer reviews written on Twitter so as to obtain a strategy to increase the profitability of the culinary and restaurants business in Surabaya. Every tweet about culinary and restaurant will be collected as data and then categorized into four aspects that we determine, namely: Price, Taste, Place, and Service. Before categorizing each review on four aspects of a restaurant, the researcher crawled customer review data on Twitter. Then the researchers did the preprocessing step so that the data were easier to process using tokenization, stopwords removal, convert into lowercase, remove punctuation, stemming and spelling correction.

This research was conducted to find the best aspect categorization performance combines the LDA method with STS by using data previously identified and written into the list of terms in Table 1 [1]. Latent Dirichlet Allocation (LDA) was chosen because it is a topic modeling method that is able to map hidden topics from a document. This study proposes a method that combines LDA with semantic similarity to categorize documents into four culinary aspects (Price, Taste, Place, and Service). The results of the performance of each trial will be analyzed, and the best will be selected. Then, the Word Embedding method is used to extract documents and labels into word vectors then use it as input to the customer review ranking process, using Global Vector for Word Representation (GloVe) method [2]. The classification used in this study is to use SVM with three modifications to the parameters of the SVM method, namely C-SVC, SVC Linear, SVCnu with various kernel changes to get the best results. Sentiment analysis in this study will compare sentiment classification using three methods of changing SVM and additional classification using SentiCircle. The best test performance for sentiment analysis will be applied.

2. RELATED THEORY

2.1 Previous Research

In several previous studies [3] and [4] , sentiment analysis was carried out without being divided into several aspects first, so that some ambiguous words (eg: high, cheap) would appear which would cause a decrease in the accuracy. In this study, the update that occurs is that the sentiment analysis is first broken down into several aspects and then re-examined in each aspect with various classification methods to improve accuracy. There are several studies that have used aspect-based [5], but the difference lies in the data that will be retrieved and the classification process, in addition, researchers use SentiCircle at the end of the sentiment classification to increase accuracy. Practical benefits by getting data from reviews in every aspect to be used by business people in an effort to increase restaurant profitability. The output of these two processes (Aspect Categorization and Sentiment Classification) will result in an evaluation to find out the high positive and negative reviews of culinary lovers or reviewers on the Price, Taste, Place, and Service aspects of a restaurant or restaurant review. The results also found that in conducting research sentiment was influenced by an aspect.

2.2 Culinary and Restaurant Aspect

According to Kotler [6], customer satisfaction is a person's feelings of pleasure or disappointment that arise after comparing their perceptions or impressions of what they get. If the performance is lower than expected, the customer is not satisfied, but if the performance exceeds expectations, the customer is satisfied and happy. According to Sugiyono [7], with a research entitled "Analysis of the influence of service quality, food quality, and price on customer satisfaction at Yung Ho Restaurant Surabaya." The variables in the restaurant have dimensions such as service quality, food quality, and price.

The dependent variable is customer satisfaction. At this stage, the terms obtained from previous research are added to the expanded term list with TF-ICF and the results are shown in table 1. Bold mark is an additional term resulting from the use of the Term Frequency-Inverse Document Frequency (TF-ICF) method.

Table 1. Extended aspect definition based on term list of extracted tweets

Aspect	Expanded Term related to Obtained Aspect
Place	(Decoration, furniture, restaurant atmosphere, comfort environment) *suasana, tempat, ramai, sepi, hawa, nyaman, keren, bersih, kotor, meja, estetika, fasilitas, dekorasi, cahaya, tatanan,* ***lantai, ac, lampu, santai, kursi, celah, penuh, toilet, wc, interior, sempit, basah, nongkrong***
Taste	(Menu variety, food quality, food ingredients) *enak, makan, minum, pedas, manis, segar, fresh, kuah, masak, menu, rasa, kualitas, sehat, diet, bahan,* ***salad, kue, bakso, sambal, empuk, amis, matang, bumbu, bakar, serut, potongan, goreng, kulit, asin, tebal, telur, teh, jeruk, es, buah, jus***
Service	(Parking, all of staffs, managers and employees) *layanan, lama, sebal, order, pesan, tunggu, parkir, sabar, wifi, internet, kasir, baik, membantu, pembayaran, buka, tutup,* ***jengkel, marah, sikap, jahat, kasar, ramah, bayar, resepsionis, antre, cepat, lambat***
Price	(Reasonable Price) h*arga, mahal, murah, item, nilai, wajar, uang,* ***duit, kantong, akal, porsi***

2.3 Term Frequency – Inverse Cluster Frequency (TF-ICF)

Term Frequency - Inverse Cluster Frequency (TF-ICF) is a weighting term build upon information from documents in a cluster with looks at the frequency of terms by using equation (1) to calculate ICF, and using equation (2) to calculate the TF-ICF [8].

$$ICF_i = 1 + log\frac{C}{cf_i} \quad (1)$$

$$TF - ICF_j = tf_{ji} \times ICF_i \quad (2)$$

2.4 Latent Dirichelet Alocation (LDA)

LDA is an Unsupervised Generative Model that categorizes words that appear in documents into groups which are usually referred to as topics [5]. The basic idea of Latent Direchelet Allocation (LDA) is that the document consists of several latent topics where each topic consists of a distribution of words. The relative importance of topics is captured in different weights and varies from document to document. Below equation (3) is the model equation used from Griffhs [5].

$$p(z_i = k / z_{-i}, w) \infty \frac{nxn_{-i,k}^{w_i} + \beta}{n_{-i,k} + W\beta 1} \cdot \frac{n_{-i,k}^{d_i} + \alpha}{n_{-i}^{d_i} + T\alpha 2} \quad (3)$$

$n_{-i}^{(\cdot)}$ is the number of z_i, $n_{-i,k}^{w_i} + \beta$ are a number of topics z related to the letter w_i, $n_{-i,k}^{d_i} + \alpha$ is the number of z topics associated with the document d_i , W preprocessing word, T is the number of term on topics. LDA obtained is then measured for its proximity to the terms obtained in Table 1 using Semantic Similarity, the equation is shown in equation (4) [9].

$$Similarity(w_i, w_j = \frac{\sum_{m-1}^{K} w_i^m w_j^m}{\sqrt{\sum_{m-1}^{K}(w_i^m)^2}\sqrt{\sum_{m-1}^{K}(w_j^m)^2}} \quad (4)$$

Where distance measured by similarity in word1 w_i and word2 w_j). $\sum_{m-1}^{K}$ =number of iterations m to K words.

2.5 Support Vector Machine

The concept of SVM can be simply explained as trying to find the best hyperplane as a separator between two classes in the input space. [10]. The basic principle of SVM is a linear classifier, and it is more used to deal with nonlinear problems [11]. Add the concept of kernel skills to the high-dimensional workspace. This development has stimulated research interest in the field of pattern recognition to study the potential of SVM functions in theory and application [12].

2.6 SentiCircle

SentiCircle is one way for getting sentiment build upon the relation of a set of words. Sentiment points are represented in the form of circular polar coordinates. Figure 1 term in two upper quadrants have a positive sentiment, and upper left quadrant representing stonger positive sentiment. Similary, terms in two lower quadrants have negative values with left quadrant representing stronger negative sentiment [13].

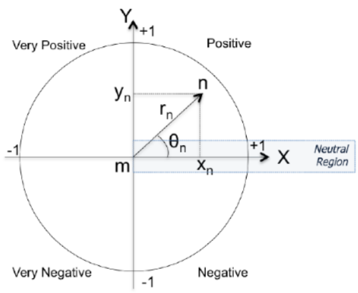

Figure 1. *SentiCircle Representation*

3. RESEARCH METHOD

The data used in this research is Twitter data crawling data taken from December 2020 to June 2021 collected 473 tweets using the Twitter API from various review account timelines and food review tags with geotags for Surabaya locations. Below are the methodological stages in this research. Figure 2 is the general process flow of the system.

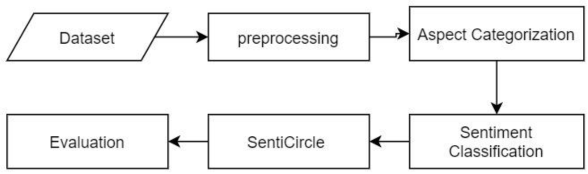

Figure 2. *System flowchart in general*

3.1 Dataset

This section describes the design of the data that will be used in the trial. The data used in the experiment is Twitter data there are obtained by crawling data with the hashtags #reviewrestoran and #reviewfood with a Surabaya geolocation filter, besides that data is also obtained from the timeline of several culinary and food accounts in Surabaya.

3.2 Preprocessing

Preprocessing is an early stage that aims to prepare text data so that it can be changed to be more structured and ensure that the data to be processed in data mining is structured and good data. Processing plays an important role. When the data has gone through the preprocessing stage, the results of research performance will also increase. In addition to the preprocessing stage, the author performs manual labeling and acts as an annotator to then serve as the ground truth of each review tweet. An example of tweet data that has passed the preprocessing stage is shown in Table 2.

Table 2. *Dataset examples*

ID	Tweet Review	Word after pre-processing	ground truth aspect label
1.	“Keren asik tempatnya enak…….!!! https://t.co/PFcCp777Iwe”	keren, asik, tempat, enak	Place
2.	"RT Pada dasarnya, Bisnis restoran banyak yang gulung tikar karena pelayanan kaku. https://t.co/QFsCp8zoIG"	dasar, bisnis, restoran, gulung, tikar, layanan, kaku	Service

3.3 Hidden Topic LDA

After doing the pre-processing stage, the next process is to categorize aspects of each review document. The main purpose of this process is to get the results of the category of each review document into four aspects that are analyzed (Price, Taste, Place, and Service). This research using LDA to get the topic hidden from the sentence. Table 3 is an example of the LDA value output for each word that has been generated from the Indonesian corpus dictionary model. This study using three scenarios of categorization aspects; AC1, AC2, AC3. with the addition of terms from the expanded term list in Table 1.

***Table 3**. Example of LDA Hidden Topic Results*

ID	Hidden Topic
1.	0.051 * “keren” + 0.032 * “asik” + 0.162 * “tempat” + 0.124 * “enak”
2.	0.271 * ”layanan” + 0.123 * ”restoran” + 0.062 * ”bisnis” + 0.212 * “dasar”+ 0.131 * “gulung” + 0.021 * “tikar” + 0.210 * “kaku”

3.4 Semantic Similarity

Semantic Similarity serves to categorize each document into four culinary aspects to be studied, namely price, taste, place, and service. The method used to categorize each document will be assisted by WordNet similarity. WordNet is a large English vocabulary database. Nouns, verbs, adjectives and adverbs are grouped into a series of synonyms called synsets. [14]. Wordnet looks for word similarity, semantic relation, and close relationship or similarity between words [15]. The output from Table 3 is then calculated for its similarity with the terms in Table 1, the output results are shown in the example Table 4.

***Table 4.** Example of aspect categorization results*

Review	Aspect				Label
	Price	Taste	Place	Service	
Keren, asik, tempat, enak	0.031	0.058	**1.215**	0.023	**place**
dasar,bisnis, restoran, gulung,tikar, layanan, kaku	0.031	0.024	0.037	**0.098**	**service**

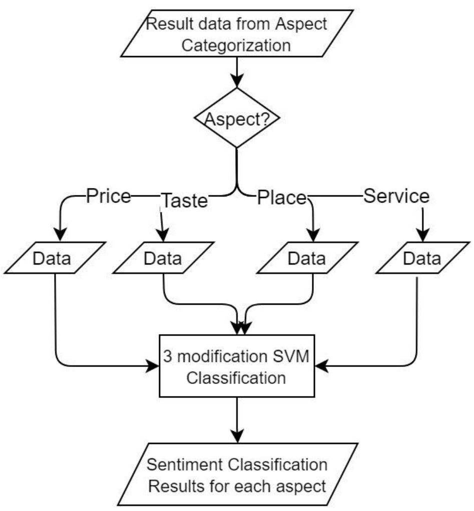

***Figure 3.** Sentiment Classification Flowchart*

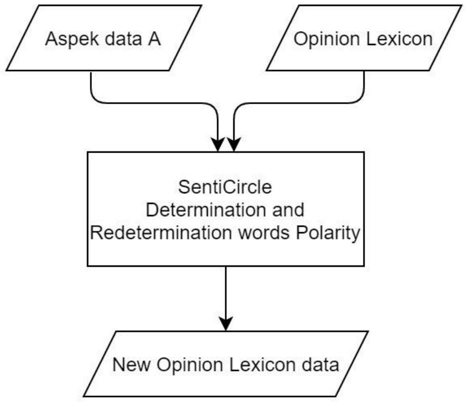

***Figure 4**. SentiCircle plot to gets new opinion Sentiment*

3.5 Sentiment Classification SVM

After passing the aspect categorization, the next stage is to extract characteristics from each tweet for classification of positive and negative sentiments. Classification will be carried out in every aspect to prevent ambiguity that will later appear in each data like shown in Figure 3. Word Embedding used in this research is Global Vector for Word Representation (GloVe). Feature data is obtained by dividing each word to get a feature value per word, then the results of the pieces are averaged to one value in each review. The classification used in this study is to use SVM with 3 modifications to the parameters of the SVM method, which is C-SVC, SVC Linear, SVCnu with various kernel changes to get the best results [10].

3.6 SentiCircle

In this study, we will use SentiCircle to reobtained the value of the polarities of each data with Opinion Lexicon on each Aspect to redefine the polarity value on new opinion word [16]. Process flow can be seen in Figure 4. The inputs at this stage are Aspect A-Data as a result of Aspect Categorization. Using SentiCircle, each word has a different polarity value depending on the aspect.

4. RESULTS AND ANALYSIS

This evaluation will be carried out to find the best performance from the Aspect Categorization (AC) and Sentiment Classification (SC) with the method used. Each performance will consist of several approaches and will be evaluated using: Precision, Recall and, F1 Measure.

4.1 Evaluation of Aspect Categorization

Based on 473 Twitter review data that were crawled and processed through preprocessing, the aspect categorization consisted of methods for finding hidden topics, expanding term lists, and categorizing using the similarity method. The performance of each trial was calculated using precision, recall, and F1 measure. The best performance is shown in Table 5. It can be seen that AC 3 with an expanded term list has the best performance.

Table 5. *AC Performance Test*

Aspect Categorization Performance		
Scenario	Methods	F1-Meassure
AC1	LDA, Semantic Similarity, predetermined termlist	0.701
AC2	LDA, Semantic Similarity, predetermined termlist + TF-ICF 50%	0.815
AC3	LDA, Semantic Similarity, predetermined termlist + TF-ICF 100%	0.914

4.2 Evaluation for Sentiment Classification

After categorizing aspects using the best categorization method, it is now possible to conduct experiments to calculate the performance of sentiment classification using SVM. 473 review tweets will be labeled with positive and negative sentiments manually by the author and will be divided into training and testing data with a ratio of 8:2.

Then after the training, it is continued with the prediction process which will produce the best system accuracy according to the method used. The SVM methods used are C-SVC, Linear and, NuSVC, with 5-fold cross-validation. Which is divided into SC1, SC2, SC3 as shown in Table 6. An example of calculating sentiment classification output is shown in Table 7.

Table 6. *Test Scenarios for Sentiment Classification*

Scenario	Evidence
SC1	AC3 + GloVe + C-SVC
SC2	AC3 + GloVe + LinearSVC
SC3	AC3 + GloVe + NuSVC

Table 7*. Classification output example with SC3*

AC	Tweet review	word	Score		Avg result	
			Pos	Neg	Pos	Neg
Service	dasar, bisnis, restoran, gulung, tikar, layanan, kaku	dasar	0.198	0.089	0.225	**0.377**
		bisnis	0.249	0.235		
		restoran	0.353	0.341		
		gulung	0.142	0.457		
		tikar	0.283	0.547		
		layanan	0.432	0.454		
		kaku	-0.084	0.513		
Place	Keren, asik, tempat, enak	Keren	0.453	-0.144	**0.389**	0.019
		asik	0.447	-0.163		
		tempat	0.097	0.495		
		enak	0.561	-0.107		

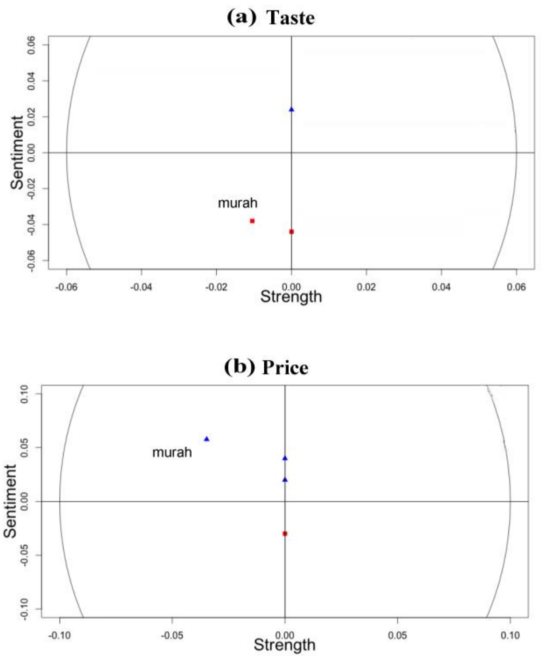

Figure 5*. SentiCircle Example on Taste and Price Aspect*

The results of the evaluation of the Sentiment Classification were obtained seen in Table 8.

Table 8*. Sentiment Analysis Test*

Sentiment Classification Performance			
Performance evaluation	Score		
	SC1	SC2	SC3
Precision	0.831	0.746	0.899
Recall	0.852	0.789	0.866
F1-Measure	0.842	0.768	0.881

4.3 Accuracy Improvement with SentiCircle

In Figure 5, we see that each word has a different value depending on the aspect in which it is categorized. The word "*murah*" in the Taste aspect has a very negative value in the quadrant III with a value of x-axis -0.012 and y-axis -0.038 while "*murah*" on price is in the quadrant II with x-axis on -0.037 and y-axis 0.058 which is very positive.

At this stage, the misclassified tweet review due to the polarity of words with multiple meanings can be corrected. The improvement using SentiCircle is shown in Table 9. The results show an increase in performance of about 0.02% or 2% with the addition of this method.

Table 9. *Improved accuracy with SentiCircle*

Fold	SC3	SC3 + SentiCircle
1	0.885	0.912
2	0.876	0.905
3	0.898	0.906
4	0.891	0.911
5	0.873	0.894

4.4 Sentiment Evaluation Results on Culinary Aspects of Surabaya

After reaching the conclusion of the method to be used, the author then applies the method to get the final result of "Sentiment Analysis based on aspects of culinary and restaurant reviews using LDA and SVM to increase the profitability of the culinary business in Surabaya". The results of the methods used for Aspect Categorization and Sentiment Classification are shown in Table 10.

Table 10. *Final Result of Sentiment Evaluation on Culinary and Restaurant Analysis Aspect*

Sentiment Result on based on every Aspect		
Aspect	**Sentiment**	***Evaluation***
Price	Positive	8.521%
	Negative	4.347%
Taste	**Positive**	**39.130%**
	Negative	6.5217%
Place	Positive	21.739%
	Negative	4.522%
Service	Positive	4.348%
	Negative	10.869%

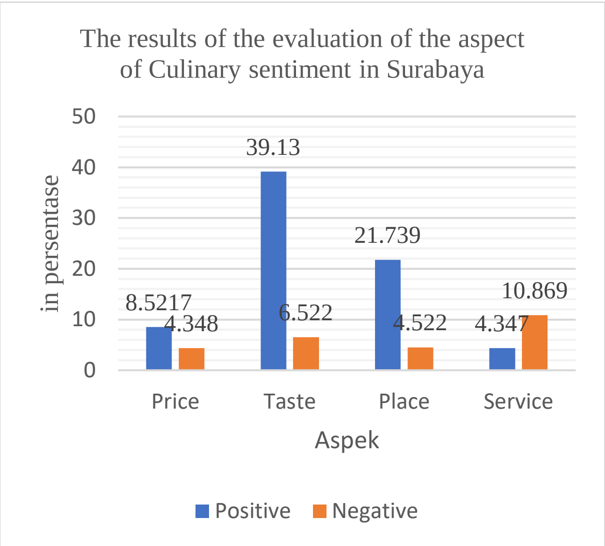

Figure 6. *Graph of Sentiment Evaluation Results on Culinary and Restaurant Analysis based on Aspects*

From Table 10 or Figure 6 it can be concluded that Twitter users and culinary lovers more interested to reviews about taste, where the evaluation results in percent reached 39.130% positive reviews and 6.522% negative reviews where this means the "Taste" aspect is very often discussed by the public in Surabaya. There are also many positive reviews on the “Place” aspect with 21.734% and negative reviews 4.521%. Slightly different from the aspect of Taste and Place, the “Price” aspect has quite a few differences, where positive reviews are 8.521% and negative reviews are 4.347%. The “Service” aspect is different from other reviews, where the negative reviews higher than the positive reviews. Positive reviews are 4.348% and negative reviews 10.869%.

This study also proves that the division of aspects is quite influential on the performance of sentiment classification because there are ambiguous words that have two different meanings in each sentence depending on the context.

5. CONCLUSION

“Taste” aspect has the highest number of reviews compared to other aspects. In the Taste aspect, it has a lot of positive sentiments of 39.130% with negative sentiments of only 6.521%, this means that peoples in Surabaya are very concerned with "Taste" compared to other aspects of culinary and restaurants in Surabaya. The lowest negative sentiment is on “Price” aspect with 4.347% which means that not many people mention how cheap or expensive a restaurant is. The thing to note is that Twitter reviews on culinary and restaurant have higher negative sentiments than positive sentiments on the “Service” aspect with 10.869%, which means that the management or culinary business owners in Surabaya should pay more attention to the services provided to customers.

AUTHORS CONTRIBUTIONS

Drajad Bima Ajipangestu: Study concept and design, data collection, writing the paper. Riyanarto Sarno: supervising.

ACKNOWLEDGMENTS

This research was funded by the Indonesian Ministry of Education and Culture under Riset Inovatif-Produktif (RISPRO) Invitation Program managed by Lembaga Pengelola Dana Pendidikan (LPDP) and under Penelitian Terapan Unggulan Perguruan Tinggi (PTUPT) Program managed by Institut Teknologi Sepuluh Nopember (ITS).

REFERENCES

[1] Y. Fu, M. Yan, X. Zhang, L. Xu, D. Yang, and J. D. Kymer, “Automated classification of software change messages by semi-supervised Latent Dirichlet Allocation,” *Inf. Softw. Technol.*, vol. 57, no. 1, pp. 369–377, 2015, doi: 10.1016/j.infsof.2014.05.017.

[2] J. Pennington, R. Socher, and C. D. Manning, “GloVe: Global vectors for word representation,” 2014, doi: 10.3115/v1/d14-1162.

[3] A. Shelar and C. Y. Huang, “Sentiment analysis of twitter data,” 2018, doi: 10.1109/CSCI46756.2018.00252.

[4] R. A. Laksono, K. R. Sungkono, R. Sarno, and C. S. Wahyuni, "Sentiment analysis of restaurant customer reviews on tripadvisor using naïve bayes," 2019, doi: 10.1109/ICTS.2019.8850982.

[5] 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.

[6] P. Kotler, "Marketing management/Philip Kotler, Kevin Lane Keller," *Pearson Educ. Int.*, vol. 817, 2012.

[7] S. Jimmy and Sugiono, "Analisis Pengaruh Service Quality, Food Quality, dan Price terhadap Kepuasan Pelanggan Restoran Yung Ho Surabaya," *J. Manaj. Pemasar. Petra*, vol. 2, no. 1, pp. 1–10, 2013.

[8] L. H. Suadaa and A. Purwarianti, "Combination of Latent Dirichlet Allocation (LDA) and Term Frequency-Inverse Cluster Frequency (TFxICF) in Indonesian text clustering with labeling," 2016, doi: 10.1109/ICoICT.2016.7571885.

[9] H. Saif, Y. He, M. Fernandez, and H. Alani, "Contextual semantics for sentiment analysis of Twitter," *Inf. Process. Manag.*, vol. 52, no. 1, pp. 5–19, 2016, doi: 10.1016/j.ipm.2015.01.005.

[10] D. B. Ajipangestu and R. Sarno, "Event classification in surabaya on twitter with support vector machine," 2020, doi: 10.1109/iSemantic50169.2020.9234205.

[11] N. Guenther and M. Schonlau, "Support vector machines," *Stata J.*, 2016, doi: 10.4018/978-1-60960-557-5.ch007.

[12] P. W. Wang and C. J. Lin, "Support vector machines," in *Data Classification: Algorithms and Applications*, 2014.

[13] F. Nurifan, R. Sarno, and K. R. Sungkono, "Aspect based sentiment analysis for restaurant reviews using hybrid ELMo-wikipedia and hybrid expanded opinion lexicon-senticircle," *Int. J. Intell. Eng. Syst.*, vol. 12, no. 6, pp. 47–58, 2019, doi: 10.22266/ijies2019.1231.05.

[14] C. Corley and R. Mihalcea, "Measuring the semantic similarity of texts," no. June, pp. 13–18, 2005, doi: 10.3115/1631862.1631865.

[15] M. Fikri and R. Sarno, "A comparative study of sentiment analysis using SVM and Senti Word Net," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 13, no. 3, 2019, doi: 10.11591/ijeecs.v13.i3.pp902-909.

[16] R. Sheik, S. S. Philip, A. Sajeev, S. Sreenivasan, and G. Jose, "Entity level contextual sentiment detection of topic sensitive influential twitterers using senticircles," in *Advances in Intelligent Systems and Computing*, 2018, vol. 542, doi: 10.1007/978-981-10-3223-3_19.