Accredited by National Journal Accreditation (ARJUNA) Managed by
Ministry of Research, Technology, and Higher Education, Republic Indonesia with Second Grade (Peringkat 2, Sinta 2)
since year 2017 to 2021 according to the decree No. 10/E/KPT/2019

Published online on the journal's webpage: **http://jurnal.iaii.or.id**

RESTI JOURNAL
(System Engineering and Information Technology)
Vol. 5 No. 3 (2021) 520 - 526 ISSN Electronic Media: 2580-0760

Multi Aspect Sentiment of Beauty Product Reviews using SVM and Semantic Similarity

Irbah Salsabila[1], Yuliant Sibaroni[2]
[1,2]School of Computing, Telkom University
[1]irbahsalss@student.telkomuniversity.ac.id, [2]yuliant@telkomuniversity.ac.id

Abstract

Beauty products are an important requirement for people, especially women. But, not all beauty products give the expected results. A review in the form of opinion can help the consumers to know the overview of the product. The reviews were analyzed using a multi-aspect-based approach to determine the aspects of the beauty category based on the reviews written on femaledaily.com. First, the review goes through the preprocessing stage to make it easier to be processed, and then it used the Support Vector Machine (SVM) method with the addition of Semantic Similarity and TF-IDF weighting. From the test result using semantic, get an accuracy of 93% on the price aspect, 92% on the packaging aspect, and 86% on the scent aspect.

Keywords: support vector machine, semantic similarity, TF-IDF

1. Introduction

Nowadays, there are many types of beauty products on the market, but not all beauty products give the expected results because it depends on the consumers' condition. Therefore, the consumers need to do research before buying beauty products that suit to the consumers' condition. Female Daily is one of the largest community platforms in Indonesia that can be used to search for a beauty product review, not only about skincare but also about makeup, perfume, nail, hair, and body.

The reviews of a product can help consumers to know the overview of the product[1]. The existence of the reviews can be referenced for consumers, for example, there is a Female Daily site which aims to find reviews of beauty product, that has thousands of reviews on each category[2]. When the consumers read the reviews of each product, usually they want to know the sentiment written in the reviews. By reading these reviews in total takes a long time, so to overcome the problem, the sentiment analysis method in mining text is used to find the sentiment(positive, negative, and neutral) in each review[3].

Sentiment analysis is a technique to collect and analyze public opinions, attitudes, and emotions towards a product. Sentiment analysis is also a reading contextual mining process that can identify and extract subjective data in the form of review, either opinion or review[4][5]. Sentiment analysis has three levels, namely document, sentence, and aspect. This is an example of a review on one of the beauty review on the femaledaily.com site "Lipstik ini packgingnya bagus harganya murah dan warnanya sesuai sama yang aku, ketahananya juga bagus", from the review, there are several general aspects such as the packaging, price, color, and the resistance. This study used aspect level in analyzing the sentiment of the review on female daily site.

A study conducted by Prayogo [6] entitled "Aspect Based Sentiment Analysis Terhadap Ulasan Hotel Berbahasa Indonesia" discussed hotel booking site reviews on the Tripadvisor site using the Naïve Bayes classification method with the aspect level. There are five aspects, namely location, food, sanitation, service, and convenience. The final result of the study gets an optimal result, namely F1-Measure of 70 % and accuracy value of 70%. A Similar study that used aspect level was done by Rahmawati [7] using the Support Vector Machine method and Particle Swarm Optimizing got accuracy value in each category aspect of 80,68% attractiveness, 51.80% accessibility, 73.68% accommodation, 86.24% price, 74.60% infrastructure, and 83.97% service.

Accepted by editor: 12-04-2021 | Final Revision: 12-06-2021 | Online Publication: 20-06-2021

In the study conducted by Wijayanti used Support Vector Machine as the classification with TF-IDF weighing using data found on the marketplace got an accuracy value of 91.42% [8]. A study conducted by Prasanti used N-Gram feature extraction [9] with feature weighting using TF-IDF, got a precision value of 77.85%, recall 74.18%, and f-measure 75.25% which is optimal by using unigram, compared to using bigram, because bigram categorized the feature by two words and it caused many word features are not in the training data.

A study conducted by Gautman [10] proved that the addition of Semantic Similarity can enhance the accuracy, the final result showed an increase of 1.7%, from 88.2% to 89.9%, which used dataset product review from Twitter. A similar study [11] using the Support Vector Machine method with the addition of Semantic Similarity in the testing stage obtained an increase of accuracy of 7% from 67% to 74%. The Semantic Similarity used comes from corpus knowledge in the form of similarity between the suitable word to the word used in the study. The advantage of Semantic Similarity can be used to solve retrieval information such as searching problems, query suggestion, and automatic summarization, besides helping the engine solve the ambiguity problem[12][13].

Based on the previous research can be concluded that the Support Vector Machine method and the addition of Semantic Similarity to the testing process can enhance the accuracy [11]. Support Vector Machine learning method is one of the methods that functions to analyze data, which is used to classify and analyze regression, the use of Support Vector Machine can produce better accuracy values that can be applied to the case of data classification [14]. This case study uses N-gram as feature extraction with TF-IDF weighting, there are several features on N-gram, one of them is the Unigram feature which gives better results when compared to Bigram [9]. Sentiment analysis is carried out using the aspect level because the aspect level is more specific in determining sentiment and the focus on an aspect that will be assessed from the entire document. The difference of this study from the previous research is that the Semantic Similarity process is not only in the testing process, but also the Semantic Similarity is carried out in the training and testing process and uses the aspect level as the sentiment analysis. This study aims to build an aspect-level sentiment analysis system using female daily data and the Support Vector Machine method with the addition of Semantic Similarity, to know the best implementation of addition Semantic Similarity and shows the most dominant aspects discussed in this study.

2. Research Method

This study proposes a system design that can perform classification on the beauty product review in the Indonesian language using Support Vector Machine and Semantic Similarity at the aspect level, system design drawing is shown in Figure 1 through several stages, starting with collecting data that will be preprocessed, after that the data will go through the feature extraction stage using N-Gram, feature weighting uses TF-IDF method, furthermore the additional work in finding the word similarity with semantic calculation approach using corpus basis by building Word2Vec dictionary. The Support Vector Machine method's learning stage is the last stage by giving positive, negative, and neutral class results.

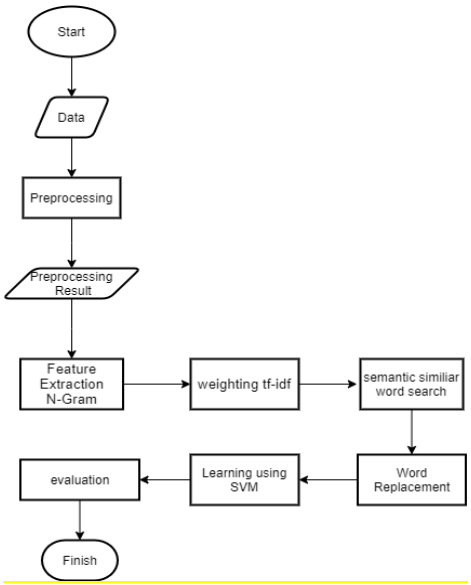

Figure 1. The stages of research process

2.1. Data set

Using 5054 reviews on the femaledaily.com site. 80% data used for building models, 20% testing. The product categories that are used are serum & essence, toner, scrub & exfoliating, and sunscreen. Four annotators manually labeled, this study using three general aspects that have a relationship with the product, namely price, packaging, and scent. This is an example of a beauty product review dataset on the female daily site shown in Table 1.

After determining the aspect discussed, it will be classified into three polarity classes, negative class is assigned '-1', neutral class is assigned '0', and positive class is assigned '1'. Aspect that are not discussed in a review will be assigned '0'. The example of labeling can be seen in Table 2.

DOI: https://doi.org/10.29207/resti.v5i3.3078

Table 1. Example of product review on femaledaily.com

Category product	Review
Sunscreen	suncreen yang. cocok untuk kulit saya yang kering.
Toner	toner yang bekerja paling baik untuk kulit wajah saya,membantu untuk tidak memperburuk jerawat
Serum & essence	serum yang menghasilkan rasa halus pada kulit secara instan, tetapi mungkin karena
Scrub & Exfoliate	efek silikonnya. bukan serum favoritku Butiran scrubnya halus tidak bikin wajah sakit

Table 2. Labeling dataset

No	Review	Price	Packaging	Scent
1	Kemasan sabun cuci mukaa ini sangat bagus. Harganya terbilang mahal. Texturenya enak seperti gel	-1	1	0
2	Mendapatkan toner yang, meskipun kemasannya jelek, karena suka bocor tetapi harganya murah. Teksturnya sendiri cair, aromanya juga wangi seperti wangi bunga	1	-1	1
3	Sunscreen termahal yang pernah dibeli, teksturnya gaenak. Nyesel beli ini	-1	0	0
4	Scrub kesukaan temen-temen, tapi ternyata malah bikin jerawatan. Aromanya enak aja wangi bunga, kemasannya tidak aman	0	-1	-1

2.2. Preprocessing

Remove Symbol, Punctuation and Case Folding

This process changes the capital letters to lowercase letters to form a standard text and then removes symbols, punctuation marks, and usernames. Table 3 is an example of remove symbol, punctuation, and case folding.

Table 3. Remove Symbol, Punctuation, and Case Folding

Before	After
“Toner wardah murah, tapi gk cocok sm muka. Bikin jerawatan☹”	“toner wardah murah tapi gk cocok sm muka bikin jerawatan“

Tokenizing

The process to convert words into tokens, in this study using a function available in NLTK, namely word_tokenize (). Table 4 is an example of the tokenizing stages.

Table 4. Tokenizing

Before	After
“toner wardah murah tapi gk cocok sm muka bikin jerawat”	‘toner’,’wardah’,’murah’,’tapi’,’gk,’cocok’,’sm’,’muka’,’bikin’,’jerawat’

Word Normalization

Non-standard words will be changed into familiar words, changing abbreviated words into original words, typos. Researchers created 6820 normalized words. Table 5 is an example of the word normalization stages.

Table 5. Word Normalization

Before	After
‘toner’,’wardah’,’murah’,‘tapi’,’gk’,’cocok’,’sm’,’muka’,’bikin’,’jerawatan’	‘toner’,’wardah’,’murah’,’tapi’,’tidak’,’cocok’,’sama’,’muka’,’membuat’,’jerawatan’

Stopword Removal

The process to remove words that do not affect sentiment analysis by using the NLTK library to get an Indonesian stopword list. Table 6 is an example of the stopword removal stages.

Table 6. Stopword Removal

Before	After
‘toner’,’murah’,’tapi’,’tidak’,’cocok’,’sama’,’muka’,’membuat’,’jerawatan’	’toner’,’wardah’,’murah’,’cocok’,’muka’,’membuat’,’jerawatan’.

Stemming

Stemmer functions to remove affixes to words, to perform stemming uses stemmer sastrawi by adding the swifter library to accelerate the steaming process. Table 7 is an example of the stemming stages.

Table 7. Stemming

Before	After
‘toner’,’wardah’,’murah’,’cocok’,’muka’,’membuat’,’jerawatan’	toner’,’wardah’,’murah’,’cocok’,’muka’,buat’,’jerawat’.

2.3. Feature Extraction N-Gram

This study uses N-Gram as the feature extraction. N-Gram is a combination of adjectives that frequently appears to show a sentiment [15]. There are several N-Gram features, namely n = 1 (unigram), n = 2 (bigram), and n = 3 (trigram). Unigram is the feature extraction by making every word in a sentence be a feature, bigram makes every two words in a sentence to be a feature, and trigram makes every three words in a sentence to be a feature. This study will use the unigram feature because bigram categorized the feature by two words, and it caused many word features are not in the training data.

2.4. TF-IDF

Weighting is used TF-IDF to calculate weight in each word. The data will be processed on weighting for each

DOI: https://doi.org/10.29207/resti.v5i3.3078

word. The aim is to find word features then calculate the frequency of the words (TF). After that, it will be calculated the frequency weight value in each term that has been obtained (Wtf). Finally, the calculation of the document that contains several words (DF) will be carried out and inverse the DF Value (idtf). There are two libraries to implement TF-IDF, namely Tfidfvectorizer and Tfidftransformer, this study uses Tfidfvectorizer to finish it.

2.5. Word2Vec

Word2Vec is two layers of neural network that process the text. There are two Word2Vec algorithms, namely continuous bag-of-word and continuous skip-gram. With the continuous bag-of-words algorithm, the order of the sentences in the history does not affect the projection. This algorithm predicts the current words based on the context. While the skip-gram algorithm predicts the words around a word [16]. Word2Vec is used to map every feature word into vector form to find the Semantic Similarity among words by looking at the proximity of the results between vectors. The corpus was built using all sentence review_data, as many as 5054 corpus. Figure 2 is an overview of the system for building the Word2Vec formula.

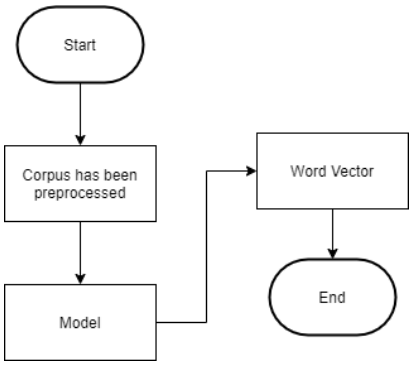

Figure 2. Build a Word2Vec dictionary

2.6. The addition of semantic information

The addition of semantic information is a measurement to find values that indicate the level of Semantic Similarity or closeness, texts or sentences. The calculation is done by building the Word2Vec model using the insight dictionary contained in the dataset. The system will read the two-word vector then do the Semantic Similarity calculation by using cosine similarity formula and taking ten top word features using most_similiar function. For example, a sentence " Saya suka memakai lipstick Maybelline " the word "suka" becomes the selected adjective. Let's assume two words; 'suka' and 'puas' tend to be similar to the word 'senang', can be seen from the word meaning closeness, by searching Semantic Similarity based on the value of cosine-similarity, finding the word "senang", that has the same meaning to the word "suka". The addition of Semantic Similarity aims to classify the unrecognized data by the model that has been made. In general, the addition of information is written in Algorithm 1[11].

```
Algoritma Semantic Similarity
For every feature in sentence do
  If feature not in TFIDF vocab then
     Search top ten similar word in Word2Vec
  dictionary()
     For each word in ten word do
        if word in TFIDF vocab then
              Replace old feature with new
  word in TFIDF
```

2.7. Support Vector Machine

After the preprocessing stage and the feature extraction, the next step is to execute the training data by classification method using Support Vector Machine algorithm. Support Vector Machine searches hyperplane to create maximum separation for all classes used. Figure 3 is the general overview of Support Vector Machine.

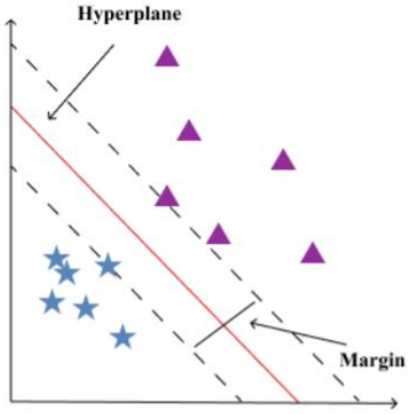

Figure 3. Support Vector Machine

The output generated from this classification stage is the sentiment polarity from every aspect resulting form the previous stage, namely the feature extraction stage using a pre-shaped TF-IDF vector. To classify the test data, the result of training data is needed. The classification will use library Support Vector Machine (libSVM).

2.8. Evaluation

This evaluation stage will be conducted by using accuracy that can be measured through prediction results. To calculate the evaluation method required true positive, true negative, false positive, and false negative from confusion matrix. There are four types of classification results that are used to determine accuracy, recall, precision, and F1-score can be seen from the similarity 1, 2, 3, and 4.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

DOI: https://doi.org/10.29207/resti.v5i3.3078

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - Score = \frac{2 \times precision \times recall}{precision + recall} \quad (4)$$

TP is true positive which means the result of classification class is predicted correct, and the fact is correct. TN is true negative which means the result of classification class is the wrong prediction, and the fact is correct. FP is false positive which means the result of the classification class is predicted correct, and the fact is wrong. FN is false negative which means that the classification class is a wrong prediction, and the fact is right [17].

3. Result and Discussion

In this study, 3 test combinations were carried out, namely Support Vector Machine with the addition of Semantic Similarity , Support Vector Machine with the addition of Semantic Similarity to the testing process, and Support Vector Machine. The calculation that is used to know the effect of the addition of Semantic Similarity in this study is used accuracy result.

3.1. Result Evaluation

Testing was carried out using the Support Vector Machine method with the addition of Semantic Similarity. The Support Vector Machine method with the addition of Semantic Similarity to the testing process and the Support Vector Machine method without the addition of Semantic Similarity gave results as table 8.

Table 8. Result Evaluation

Method	Aspect Accuracy		
	Price	Packaging	Scent
SVM + SS	93%	92%	86%
SVM + SS (Testing)	88%	90%	81%
SVM	93%	92%	86%

The results of the test show that with the addition of semantics to the training and testing process, the accuracy is good, there is an increase in accuracy of 5% in the scent aspect. Furthermore, there is an increase in accuracy on the packaging aspect by 2%, and in scent accuracy, there is an increase in accuracy of 5%. By using the optimal parameter of the tuning parameter, namely the Linear Kerner, C = 1, and the Gamma = 1, in making the Word2Vec model using a parameter with a value of count one and window three and using the skip-gram algorithm. This is because the sentence used is a sentence that has been trained, so when predicting, the existing model is familiar with the words contained in the sentence. Addition of Semantic Similarity has been shown to improve accuracy significantly.

3.2. Data Testing Confusion Matrix

The final step in this research is to test the accuracy of the model Support Vector Machine with the addition of Semantic Similarity using the Confusion Matrix for aspect price. The result of the test can be seen in Table 9.

Table 9. Confusion Matrix Price

Actual	Prediction		
	Positive	Negative	Neutral
Positive	(TP)152	(FN)29	(FNet)5
Negative	(FP)6	(TN)588	(FNet)12
Neutral	(FP)7	(FN)19	(TNet)263

In predicting the price aspect sentiment results, there were 152 documents that correctly predicted positive, then 588 documents correctly predicted negative, and 263 documents correctly predicted neutral.

Result confusion matrix for aspect packaging can be seen in Table 10.

Table 10. Confusion Matrix Packaging

Actual	Prediction		
	Positive	Negative	Neutral
Positive	(TP)17	(FN)23	(FNet)16
Negative	(FP)2	(TN)905	(FNet)8
Neutral	(FP)3	(FN)33	(TNet)74

In predicting the price aspect sentiment results, there were 17 documents that correctly predicted positive, then 905 documents correctly predicted negative, and 74 documents correctly predicted neutral.

Result confusion matrix for aspect scent can be seen in Table 11.

Table 11. Confusion Matrix Scent

Actual	Prediction		
	Positive	Negative	Neutral
Positive	(TP)8	(FN)19	(FNet)24
Negative	(FP)1	(TN)715	(FNet)39
Neutral	(FP)5	(FN)61	(TNet)209

In predicting the price aspect sentiment results, there were 8 documents that correctly predicted positive, then 715 documents correctly predicted negative, and 209 documents correctly predicted neutral.

3.4. Discussion

It can be seen from table 8 that the accuracy of the price aspect gets a high accuracy value compared to other aspects, because in the price aspect the results of the dataset labeling are more balanced, on the packaging and scent aspects, there is a neutral class being the majority, the number of reviews that do not discuss packaging and

DOI: https://doi.org/10.29207/resti.v5i3.3078

scent, resulting in unbalanced data, which resulted in the model being unable to predict optimally. Improved accuracy also depends on the dictionary being built. This is due to the vector that has been plotted according to the richness of the corpus topic which affects the top ten word lists obtained from the Word2Vec that have been made. In this study, the Word2Vec corpus was obtained from the preprocessing results of 5054 beauty product reviews, which means there is still little to determine the meaning according to these data. In addition, the length of the review sentence used to build the Word2Vec dictionary is not uniform, some queries only consist of one to two words after going through the preprocessing stage. While the value of the word vector is obtained from the calculation of the proximity or proximity of the words around it, it can be seen from the window value used.

In general, system testing is used to know the success of the built system, besides knowing how much the effect of the addition of Semantic Similarity. The Substitution of words can be seen in the form of words that are similar based on the closeness meanings, opposite meanings, the same context, even the form of the word. If there is the word "murah" replaced with the word "jangkau" because the word "jangkau" is a positive meaning of the word. It can be seen based on the top ten words in figure 4 selected from the Word2Vec dictionary, for the word cheap is the one with the closest word meaning. Another example of the word "aroma" has a close relationship with the word "wangi" because it is still in the same context.

```
[('jangkau', 0.9753187894821167),
 ('kualitas', 0.945315957069397),
 ('isi', 0.9377820491790771),
 ('segi', 0.9314621686935425),
 ('lumayan', 0.927365243434906),
 ('padan', 0.9269727468490601),
 ('susah', 0.9256956577301025),
 ('relatif', 0.9197761416435242),
 ('sayang', 0.9181742668151855),
 ('lokal', 0.9176479578018188)]
```

Figure 4. Selected word in Word2Vec dictionary

In the sentences of beauty product reviews on the female daily site, there are words containing many types of writing that cause the actual words in the test to be replaced first. For example, there are cheap words that have more than one type of writing such as “muraahhh”, “muerah” and others that repeat each character but have the same meaning.

The aspects that are often discussed in this study can be seen in Table 12, it can be concluded that the price aspect is the most frequently discussed aspect because there are 1306 positive reviews and 868 negative reviews, among other aspects which more dominating neutral reviews.

Table 12. Total review aspects

Aspect	-1	0	1
Price	868	2880	1306
Packaging	226	4322	506
Scent	247	3589	1218

4. Conclusion

Based on the testing and analysis that has been done, it can be concluded. The quality built-in mapping of each word feature affects the test results. The addition of Semantic Similarity result is higher than the Semantic Similarity which is only in-process testing, because the sentence used is a sentence that has been trained, when the prediction of the existing model is familiar with the words contained in the sentence. Addition of Semantic Similarity has been shown to improve accuracy significantly. The optimal accuracy obtained in this study is 93% accuracy on the price aspect, 92% on the packaging aspect, and 86% on the scent aspect. The price aspect is an aspect that is frequently discussed in this study. There are differences in the results obtained by the sentiment analysis method using data labeling, preprocessing, Support Vector Machine, and Semantic Similarity with previous studies. Because this study using the aspect level as the level of sentiment analysis and implementing Semantic Similarity in the training and testing process. Applying the sentiment analysis that classifies positive, negative, and neutral reviews can make it easier for consumers who are looking for reviews of a product because consumers do not need to read the whole reviews. The model derived from this research predicts sentiment on unlabelled data.

Future research can improve performance by using the corpus, which is used to build a broader and more varied dictionary. By doing the addition of preprocessing stage according to the data, the further researcher can add a lemmatization stage, convert words into important words from the data set, and use models and other methods as a comparison in creating the classifications and predictions for the better version.

References

[1] K. T. Kamila, Suharyono, and I. Perwangsa Nuralam, “Pengaruh Online Consumer Review Terhadap Keputusan Pembelian (Survei pada Mahasiswa Universitas Brawijaya TA 2015 / 2016 – 2018 / 2019 yang Pernah Membeli dan Menggunakan Xiaomi Smartphone),” *J. Adm. Bisnis*, vol. 72, no. 1, 2019.

[2] Z. Salsabil and M. Arfa, “Efektifitas website Femaledaily.com dalam memenuhi kebutuhan informasi pengguna,” *J. Ilmu Perpust. - Univ. Diponegoro*, vol. 8, no. 2, pp. 199–210, 2018.

[3] E. Indrayuni, “Analisa Sentimen Review Hotel Menggunakan Algoritma Support Vector Machine Berbasis Particle Swarm Optimization,” *J. Evolusi Vol. 4 Nomor 2 - 2016*, vol. 4, no. 2, 2016.

DOI: https://doi.org/10.29207/resti.v5i3.3078

[4] C. G. Kencana and Y. Sibaroni, "Klasifikasi Sentiment Analysis pada Review Buku Novel Berbahasa Inggris dengan Menggunakan Metode Support Vector Machine (SVM)," vol. 6, no. 3, pp. 10451–10462, 2019.

[5] D. Rolliawati, K. Khalid, and I. S. Rozas, "Teknologi Opinion Mining untuk Mendukung Strategic Planning," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 2, p. 293, 2020, doi: 10.25126/jtiik.2020721685.

[6] S. Prayogo, "Aspect Based Sentiment Analysis Terhadap Ulasan Hotel Berbahasa Indonesia," 2018.

[7] 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."

[8] R. Wijayanti and A. Arisal, "Ensemble approach for sentiment polarity analysis in user-generated Indonesian text," *Proc. - 2017 Int. Conf. Comput. Control. Informatics its Appl. Emerg. Trends Comput. Sci. Eng. IC3INA 2017*, vol. 2018-Janua, pp. 158–163, 2017, doi: 10.1109/IC3INA.2017.8251759.

[9] A. A. Prasanti, M. A. Fauzi, and M. T. Furqon, "Klasifikasi Teks Pengaduan Pada Sambat Online Menggunakan Metode N-Gram dan Neighbor Weighted K-Nearest Neighbor (NW-KNN)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya*, vol. Vol. 2, no. 2, pp. 594–601, 2018.

[10] G. Gautam and D. Yadav, "Sentiment analysis of twitter data using machine learning approaches and semantic analysis," *2014 7th Int. Conf. Contemp. Comput. IC3 2014*, pp. 437–442, 2014, doi: 10.1109/IC3.2014.6897213.

[11] L. Afina, H. Raudhoti, A. Herdiani, Romadhony, and Ade, "Identifikasi Cyberbullying pada Kolom Komentar Instagram dengan Metode Support Vector Machine dan Semantic Similarity (Cyberbullying Identification on Instagram Comment Using Support Vector Machine and Semantic Similarity)," vol. 4, no. 1, pp. 1–8, 2020, [Online]. Available: http://jcosine.if.unram.ac.id/.

[12] P. Resnik, "Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language," *J. Artif. Intell. Res.*, vol. 11, 1999, doi: 10.1613/jair.514.

[13] T. Kenter and M. de Rijke, "Short Text Similarity with Word Embeddings Categories and Subject Descriptors," *Proc. 24th ACM Int. Conf. Inf. Knowl. Manag. (CIKM 2015)*, 2015.

[14] S. Chidambaram and K. G. Srinivasagan, "Performance evaluation of support vector machine classification approaches in data mining," *Cluster Comput.*, vol. 22, 2019, doi: 10.1007/s10586-018-2036-z.

[15] W. C. Indhiarta, "Penggunaan N-Gram Pada Analisa Sentimen Pemilihan Kepala Daerah Jakarta Menggunakan Algoritma Naïve Bayes," pp. 1–18, 2017.

[16] J. Lilleberg, Y. Zhu, and Y. Zhang, "Support vector machines and Word2vec for text classification with semantic features," 2015, doi: 10.1109/ICCI-CC.2015.7259377.

[17] S. Agarwal, *Data mining: Data mining concepts and techniques*. 2014.

DOI: https://doi.org/10.29207/resti.v5i3.3078