OPEN ACCESS
J DATA SCI APPL, VOL. 3, NO. 2, PP.067-077, JULY 2020

E-ISSN 2614-7408
DOI:. 10.34818/JDSA.2020.3.58

JOURNAL OF DATA SCIENCE AND ITS APPLICATIONS

Aspect Based Sentiment Analysis on Beauty Product Review Using Random Forest

Anggitha Yohana Clara[1], Adiwijaya[2], Mahendra Dwifebri Purbolaksono[3]

School of Computing, Telkom University
Bandung, Indonesia

*adiwijaya@telkomuniversity.ac.id

Received on 11-02-2021, revised on 27-06-2021, accepted on 24-06-2021

Abstract

Cosmetics and beauty products (including skincare) are the products used as body care or face care and used to accentuate the body alure. A product could give diverse sentiment to the consumers including positive and negative sentiment. Many consumers of beauty products are sharing their reviews to help other consumers to find the right products to buy and to give feedback to the brand of the beauty product itself. The number of reviews is inversely proportional to the lack of opinion identification towards product's aspects. Hence, a study has been conducted to analyze beauty products reviews as toner, serum, sun protection, and exfoliator. The analysis process is conducted aspect based to determine sentiment towards aspect of beauty products based on the reviews. The result is addressed to people using skincare and beauty product brands in deducting consumer's opinion. The solution to this problem is by using Random Forest with hyperparameters tuning as classification method, and TF-IDF and n-gram as feature extraction methods. The multi-aspect sentiment analysis in this study obtained highest accuracy for 90.48%, precision for 87.27%, recall for 70.13%, and F1-Score for 71.77%.

Keywords: beauty products, multi-aspect sentiment analysis, TF-IDF, n-gram, random forest

I. INTRODUCTION

COSMETICS or beauty products are products used to care for our body and accentuate the body allure [1]. The use of beauty products is now becoming a lifestyle for many people these days. It has an impact on the growth of the number of consumers giving their review towards the product that they use. The review helps other consumers who consider buying a product by giving a perspective from the people who already used the product. Nowaday, there are plenty of reviews available from diverse products and brands. These reviews are spread all over the internet in many local or non-local platforms. Female Daily is a local platform for beauty product consumers sharing their experiences of using beauty products. The review can be given in a text review or 1 to 5 rating scale. However, this research conducts sentiment analysis only for the text review.

Sentiment is opinion or perspective toward situation or event. Sentiment analysis or also known as opinion mining is a research related to sentiment towards an entity [2]. In the last few years, the number of discussion forums and sites for review sharing is increasing significantly [3]. The number of beauty product reviews is increasing along with the growth of the cosmetics and beauty product industry. There are beauty product reviews in a huge number available online, but that number is inversely proportional to the lack of aspect based sentiment analysis research towards the aspect of the product. The research related to beauty products was conducted to classify a review into positive or negative sentiment [4]. Another previous research was conducted to predict the rating of product using text review from Female Daily [5].

The purpose of this research is to analyze the implicit sentiment in a review. Each review is extracted to acquire the sentiment towards four different aspects in a product. Furthermore, this research builds a classification model to do multi aspect sentiment analysis on beauty product reviews. The reviews in the dataset are translated to Indonesian because most of the reviews are written in code-mixed (bilingual, mixed of Indonesian and English). The dataset is also translated because based on previous research, a model with one language dataset has better performance [6]. This research uses Term Frequency-Inverse Document Frequency (TF-IDF) and n-gram as feature extraction methods to solve the problem that comes with the large size of the dataset. They are needed because it takes high computational time and complexity to process every word in the datasets without being extracted [7]. Random Forest is the classifier method used in this research. The Random Forest classification is conducted by using hyperparameters tuning. Hyperparameters tuning makes the model come with better performance. The results of parameters combinations are compared to find the combination with highest performance.

II. LITERATURE REVIEW

A. Multi Aspect Sentiment Analysis

Sentiment analysis or also known as opinion mining is a development of data mining on data from microblogs [8]. The beginning of sentiment learning was the analysis of public opinion in the early 20th century. Sentiment can be categorized as positive and negative, or into n number categories [9]. The sentiment contained in a review can provide an important indicator for various purposes. According to Bright Local, 82% of consumers read online reviews. Online reviews impact perception of a brand and the ability to attract and convert new consumers. It has been examined that online review affects new consumer's intention to buy cosmetic products [10].

Aspect based sentiment analysis is the development of sentiment analysis. The purpose of aspect based sentiment analysis is to identify aspect and sentiment of a review [11]. It is commonly used in business industry. Aspect based sentiment analysis reveals more essence than the usual sentiment analysis. Therefore, aspect based sentiment analysis is needed to study user's input in more detail. If we are looking from a business perspective, aspect based sentiment analysis enables the model to know consumer's opinion to spesific aspect and understand consumer's preferences [12]. The ability to know an aspect in more detail makes aspect based sentiment analysis used for product or service analysis. Aspect based sentiment analysis implemented in a problem with more than one aspect is known as multi aspect sentiment analysis. Multi aspect sentiment analysis is implemented in this research for more detail analysis between the aspects.

Mubarok, et al. [13] developed multi aspect sentiment analysis on restaurant reviews using Naive Bayes Classifier and obtained 88.13% F1-Measure value for the aspect classification and 75% for the sentiment classification. Similar research was conducted by Dina and Juniarta on employee's experience reviews using reviews from Glassdoor [14]. This research obtained highest precision and recall in Google Company aspects of company benefit and culture value.

B. Term Frequency – Inverse Document Frequency (TF-IDF)

TF-IDF is a weighting method in numerical statistics to denote the importance of a word [15]. TF-IDF comes from two different methods, namely Term Frequency (TF) and Inverse Document Frequency (IDF). TF is the frequency at which a word occurs in a corpus or document. The ratio of the TF value is directly proportional to the weight. The greater the TF value, the greater the weight value that is resulted. IDF shows the distribution of occurence of words in the corpus as a whole. The IDF calculation uses the logarithm of the total document divided by the number of documents that have the search word in it. The equations below are the equation of TF-IDF that previously explained:

$$tf_{(i,j)} = \frac{Term\ (i)\ frequency\ in\ document\ (j)}{Total\ words\ in\ document} \tag{1}$$

$$idf_{(i,N)} = \log\left(\frac{total\ number\ of\ documents}{df_{(i)}}\right) \tag{2}$$

$$w_{(i,j)} = tf_{(i,j)} \text{ x } idf_{(i,N)} \tag{3}$$

C. N- Gram

N-gram is a feature extraction method developed from Bag of Words (BoW). BoW and n-gram map the words in corpus differently. The words in BoW are mapped word by word, but the words in n-gram are mapped into n tokens of consecutive words (words) [16]. Suppose the following sentence is given, "I love working out." BoW extracts the words into ['I', 'love', 'working', 'out']. N-gram with n = 1 (unigram) extracts the word the exact same way with BoW, but the use of n = 2 (bigram) and n = 3 (trigram) give different results. Bigram extracts the words into ['I love', 'love working', and 'working out'] and trigram extracts the words into ['I love working', 'love working out'].

Mittal and Dhyani in a previous research raised the problem of multilingual text classification using TF-IDF and n-gram [17]. The problem is solved by building a classification system with the combination of TF-IDF and n-gram. The system efficiency increases from 65% to 96% by applying n-gram with n = 4. Riswanda and Ilham in 2019 also conducted a study related to sentiment analysis on hotel reviews using n-gram and TF-IDF [18]. This study was conducted by using Naive Bayes and three n-gram scenarios which are unigram, bigram, and trigram. The classification model using unigram obtained best result with precision for 100% and error rate for 0%.

D. Random Forest

Random Forest is a machine learning classification method introduced by Leo Breiman. According to Breiman, Random Forest is an ensemble model using decision tree as the individual model and bagging as the ensemble method [19]. Ensemble Learning works by combining diverse set of learners to improvise the on the quality of the model. Random Forest is a combination of decision trees with independent random samples of vectors and has the same distribution for each tree. The decision tree tends to classify based on what it has learned rather than doing new learning. Meanwhile, Random Forest is capable of doing new learning and avoiding overfit. According to Bedy Purnama in a previous study, Random Forest is a $\{h(x, \theta\ k), k = 1, \dots\}$ tree-shaped classifier, which $\theta\ k$ is a random vector that is independently distributed. Each tree in the forest defines popular classes according to the x input [20].

While Random Forest is a collection of decision trees, there are some differences between Random Forest and any Decision Tree classification model. In the application of Random Forest algorithm, as the trees are growing, some randomness will be added to make the forest diverse. As the forest gets diverse, the model works by splitting the features into random subsets and finding the best feature in that subset. This action results in a wide diversity that is generally better. Being different from Decision Tree algorithm that works by generating prediction using some set of fixed rules, Random Forest algorithm randomly select some features to perform the ensemble learning.

Optimizing hyperparameters tuning in Random Forest is a key step in making accurate predictions. The hyperparameters in Random Forest are used to increase the prediction performance or to build a faster model. Parmar, et al. [21] developed a Random Forest classification model for movie review using hyperparameters tuning. The tuned hyperparameters in this study were number of trees, number of features, and depth of each tree. This study obtained highest accuracy in the first dataset with accuracy value for 91%.

E. Evaluation Metrics

Confusion Matrix is an approach to evaluate classification models. This approach can be useful to compare actual and prediction data. The evaluation process is done by comparing positive and negative labels. The Confusion Matrix itself is filled by the number of True Positive, False Positive, False Negative, True Negative. Table I below is the Confusion Matrix table.

TABLE I
CONFUSION MATRIX

Class		**Actual**	
		Yes	**No**
Prediction	**Yes**	True Positive (TP)	False Positive (FP)
	No	False Negative (FN)	True Negative (TN)

This matrix is capable of counting accuracy, precision, recall, and F1-Score which are used to evaluate the performance. Accuracy is the most-used metric to evaluate classification models. The prediction labels are compared to the train labels, accuracy is then calculated as the ratio of correct prediction in the test set divided by total predictions. In contrast to accuracy, precision is a good measure to quantify the number of positive predictions that are actually positive and the costs of False Positive. In comparison to precision, recall is the model metric to get the proportion of actual positives that is correctly identified. In addition, F1-Score provides a single average score that balances precision and recall. These formulas below show how accuracy, precision, recall, and F1-Score is calculated:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \times 100\% \tag{4}$$

$$Precision = \frac{TP}{TP+FP} \times 100\% \tag{5}$$

$$Recall = \frac{TP}{TP+FN} \times 100\% \tag{6}$$

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{7}$$

Among the four ways to evaluate performance, precision, recall, and F-1 Score evaluate imbalanced data better than accuracy. In a training dataset of which the distribution across the classes is equal, accuracy can be useful. It is because the class distribution is not or just slightly skewed. However, when the skew in class distribution is severe, accuracy is unreliable to evaluate the model performance. to evaluate a model with imbalanced data. Galar, et al. [21] in 2014 conducted that accuracy is not a proper evaluation for imbalanced dataset. The reason is because it does not identify the numbers of correctly predicted examples of different classes.

III. RESEARCH METHOD

A. System Description

This research aims to establish a classification model to do sentiment analysis. The general description of how the system works can be seen in Figure 1. To begin with, this process is started by labeling the dataset. The labeled dataset is then translated to Indonesian. The following step is preprocessing the dataset before going to the following step which is splitting the dataset into train set and test set. The further step is extracting the features and establishing the classification model. The final step is evaluating the system that is already established.

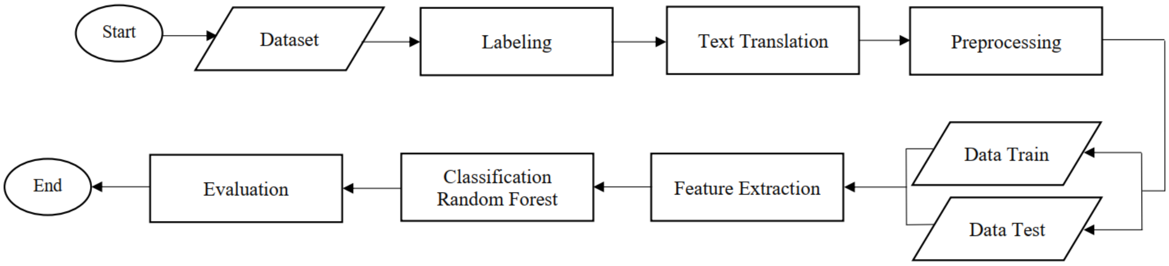

Fig. 1. Flowchart of the System

B. Dataset

This research uses review products dataset from Female Daily that is already available as the form of collaboration between Female Daily and Telkom University. This dataset can be accessed in femaledaily.com in each product category. There are 5053 reviews from 4 product categories. User's opinion of the product found in text review. The reviews are not entirely written in one language because most of the corpus contains code-mixed. Table II given below lists the product categories used in this research:

TABLE II
DESCRIPTION OF DATASET

Product Category	Number of Corpus
Toner	1456
Serum & Essence	1475
Scrub & Exfoliator	1456
Sun Protection	1457

C. Multi Aspect Labeling

The processes of both specifying aspects and labeling sentiment are done manually. There are three aspects used in this research which are price, packaging, and fragrance. Each aspect is divided into three types of sentiment namely positive, neutral, and negative. Neutral is the default value of sentiments. Every document that does not have any opinion on an aspect will be labeled as neutral. The numbers of each class of every aspect can be seen in these figures below.

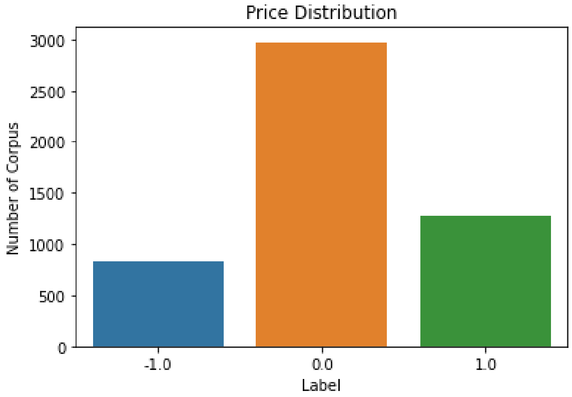

Fig. 2. Number of each label in aspect 'Price'

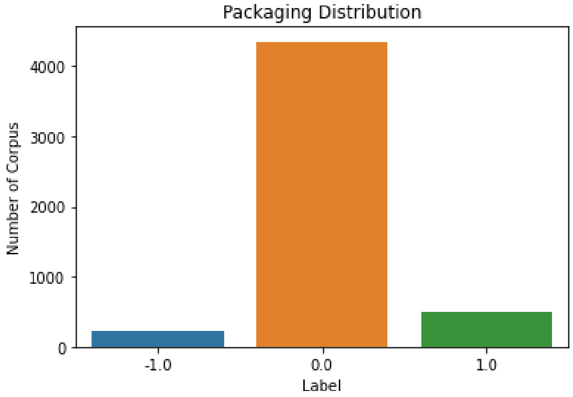

Fig. 3. Number of each label in aspect 'Packaging'

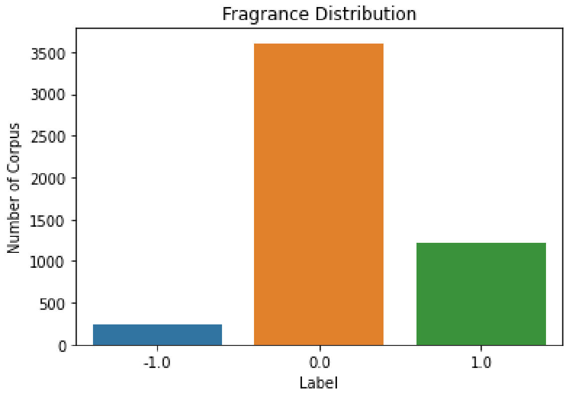

Fig. 4. Number of each label in aspect ‘Fragrance’

D. Text Translation

The research translates the dataset to minimize code-mixed which are spread among the reviews. It is because unilingual documents tend to create a model with better performance. Every corpus is translated to Indonesian using Google Translate. However, translation using Google Translate is also not fully guaranteed that the dataset is going to be translated fully and correctly.

E. Preprocessing

Textual data, just like non-textual data, has inconsistency in the data itself. However, the right preprocessing technique can improve the quality of a Machine Learning model. The output of this section is to retrieve clean words for the following step. The steps of preprocessing in this research are case folding, normalization, stopwords removal, and stemming. Figure 5 below describes the preprocessing flow.

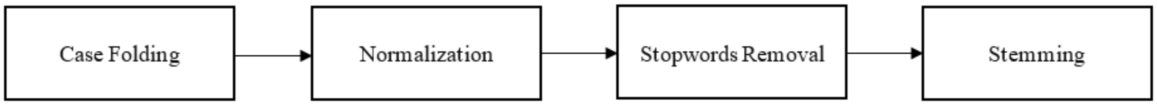

Fig. 5. Flowchart of Preprocessing

1) Case Folding

Case folding is a process of converting every letter to lowercase letter and removing non-letter characters such as numbers, punctual, and special characters. The output of this step is to equate the letters in the corpus. Given an example of a corpus with this sentence, “Produk ini berhasil membuat gue jadi mau pakai sunscreen! Suka banget pokoknya.” This sentence is processed to be, “produk ini berhasil membuat gue jadi mau pakai sunscreen suka banget pokoknya”.

2) Normalization

Normalization is a process of retrieving a word in the corpus into its actual word such as [‘bagusss’] into [‘bagus’]. The list of words that are normalized is manually written to help improving the deficiency of translation using Google Translate. This action can be done by selecting the right keywords related to each aspect and its sentiment such as the word [‘pricey’] into [‘mahal’]. Besides that, normalization can also minimize keywords such as example [‘terjangkau’] into [‘murah’].

3) Stopwords Removal

Stopwords removal, also known as filtering, is a process of eliminating unimportant words. This research eliminates stopwords in Indonesian. The list of stopwords in Indonesian is available online. The example of stopwords removal is “tapi bagus sih dan akan beli lagi” turned to be “bagus beli”.

4) Stemming

Stemming is a process of reducing words to the root form of the words. Stemming aims to extract meaningful information in a large size of data. This research uses Sastrawi stemmer that can be used for documents in Indonesian. The example of stemming is “tabir surya termahal” reduced to be “tabir surya mahal”.

F. Feature Extraction

The dataset is then processed to feature extraction after getting preprocessed. The feature extraction used in this research are TF-IDF and n-gram with n = 1 (unigram) and n = 2 (bigram). The results of unigram and bigram are compared thereafter. Given this example of consecutive words “tabir surya mahal”, using unigram the words list is turned to be [‘tabir’, ‘surya’, ‘mahal’]. On the other hand, by using bigram the words list is turned to be [‘tabir surya’, ‘surya mahal’]. The TF-IDF process is done by extracting 1000 features among the entire documents.

G. Classification

The following step of this research is constructing a classification model to train and test the dataset. The classification model is constructed to know the impact of preprocessing, n-gram in feature extraction, and hyperparameters tuning on Random Forest in sentiment classification of product’s aspects. The parameters used in this research are number of trees, maximum features, maximum depth, and criterion. The distribution of the values is then compared to find the best value for parameters that generates the best performance. The list of parameters distribution can be seen in Table III below:

TABLE III
PARAMETERS DISTRIBUTION FOR RANDOM FOREST

Hyperparameters	**Value**
Number of trees	150, 250, 350
Maximum Features	‘auto’, ‘sqrt’, ‘log2’
Maximum Depth	70, 90, 110, 150
Criterion	‘entropy’, ‘gini’

H. Evaluation Model

The final step of this research is by evaluating performance of constructed models. The process of evaluating them is done by comparing F1-Score. Model with the highest F1-Score is the model with best performance quality. However, there is still analysis to compare accuracy, precision, and recall.

IV. RESULTS AND DISCUSSION

This research is performed using three scenarios which each scenario is done consecutively from the first scenario. The first scenario shows different steps that can be performed in preprocessing. The second stop shows the comparison of unigram and bigram implementation in feature extraction. The third scenario works to find the best model using hyperparameters tuning.

1) Preprocessing

In this scenario, testing is conducted to discover the impact of using stemming and stopwords removal. There are four results of this scenario. The first result is a model using both stemming and stopwords removal. The second result is a model without using stemming, on the contrary to the third result which is a model without using stopwords removal. The fourth result is a model without using both stemming and stopwords removal. The performance of these four models then being compared as seen in Table IV.

TABLE IV
PERFORMANCE OF DIFFERENT PREPROCESSING METHODS

Preprocessing	Akurasi	Precision	Recall	F1-Score
Stemming and Stopwords Removal	90.48%	87.27%	70.13%	71.77%
Without Stemming	90.22%	83.68%	69.89%	70.96%
Without Stopwords Removal	89.91%	87.03%	67.72%	68.95%
Without Stemming and Stopwords Removal	90.25%	83.99%	69.06%	69.34%

Based on the results in Table IV, the best result is obtained from the first technique. This technique is using both stemming and stopwords removal. This technique obtained F1-Score for 71.77%, with the highest value of any metrics from accuracy to F1-Score. This result shows that in this research, it is best for this study to use stemming and stopwords removal to get better results.

2) Feature Extraction

From the previous scenario, it is obtained that the best preprocessing technique is using both stemming and stopwords removal. The dataset that has been preprocessed using the first technique, later then processed to test the second scenario. This technique is then continued by comparing the performance of using unigram and bigram. The results of unigram and bigram implementation can be seen in Table V.

TABLE V
PERFORMANCE OF DIFFERENT N-GRAM

N-gram	Accuracy	Precision	Recall	F1-Score
Unigram	90.48%	87.27%	70.13%	71.77%
Bigram	86.19%	80.26%	64.63%	68.91%

Based on scenario 2, the model using unigram obtains better performance in accuracy, precision, recall, and F1-Score. It can be seen from Table V that the major difference is in precision because the difference of both unigram and bigram precision values are about 7%. Unigram obtains better performances because of the concept of single word that also has single meaning. It is a bit different from bigram which consists of two words.

3) Hyperparameters Tuning

The third scenario implements hyperparameters tuning. Moreover, it also implements feature extraction using unigram with maximum features value is 1000. It also implements stemming and stopwords removal in the preprocessing method. The hyperparameters tuned in this scenario can be seen in Table III as mentioned before. The results of five models resulting in the highest F1-Score with tuned hyperparameters can be seen in Table VI.

TABLE VI
PERFORMANCE OF HYPERPARAMETERS TUNING

Number of Trees	Maximum Features	Maximum Depth	Criterion	Accuracy	Precision	Recall	F1-Score
150	sqrt	70	gini	90.25%	85.94%	69.29%	70.50%
150	sqrt	110	gini	90.38%	83.46%	69.54%	70.89%
150	sqrt	150	gini	90.48%	87.27%	70.13%	71.77%
250	sqrt	90	gini	90.12%	86.88%	68.73%	70.47%
350	sqrt	110	gini	90.17%	86.34%	68.77%	70.13%

Based on the results in Table VI above, the third model with 150 number of trees, ‘sqrt’ maximum features, 150 maximum depth, and ‘gini’ criterion obtains the highest performance in all metrics. This model comes with 71.77% for F1-Score. From all the five models above, the accuracy and recall values differ greatly. These values tell that the data is imbalanced. For further details of the performance of the third model with, Table VII below shows the performance of each aspect.

TABLE VII
PERFORMANCE OF EACH ASPECT OF THE BEST MODEL

Aspect	Accuracy	Precision	Recall	F1-Score
Price	93.67%	94.28%	91.55%	92.82%
Packaging	90.36%	78.55%	56.26%	60.19%
Fragrance	87.43%	88.97%	62.59%	62.30%
Average	90.48%	87.27%	70.13%	71.77%

It can be concluded from Table VII above that aspect ‘Price’ comes out with highest performance. Aspect ‘Price’ values for precision especially recall differ greatly from the other two aspects. It can be caused by several factors. In Figures 2 to 4, it is shown that the data in ‘Price’ is not as imbalanced as the other two aspects. Aspect ‘Packaging’ has the lowest precision and recall values. Figures 3 shows that the data in this aspect is the most imbalanced compared to ‘Fragrance’ and ‘Price’. The imbalanced data can affect the performance, particularly for ‘Packaging’ and ‘Fragrance’ which their most labeled class is neutral. This neutral value comes from every corpus that does not talk about packaging and fragrance, so it is labeled as the default value. This lack of labels makes the model not able to work well because the sample of the positive or negative is not comparable to the neutral value. The low precision values of ‘Packaging’ and ‘Fragrance’ means that the model fails to predict a class which in fact does not belong to a class but predicted as belonging there. Meanwhile the low recall values of ‘Packaging’ and ‘Fragrance’ means that the model misses to predict a class that does actually belong in a class but then predicted as not a part of that class.

V. CONCLUSION

1) Conclusion

It can be concluded in this research that the labels for aspect ‘Packaging’ and ‘Fragrance’ are imbalanced. The imbalanced data affects the performance of the model. The choice and combination of preprocessing method, feature extraction, and hyperparameters tuning does affect the performance of aspect-based sentiment analysis. Preprocessing method using stemming and stopwords removal obtains the best performance among all preprocessing methods. Feature extraction using unigram obtains better performance than bigram. Classification model with highest performance is tuned to 150 number of trees, 150 maximum depth, 'sqrt' maximum features, and 'gini' criterion. This model obtains the highest F1-Score for 71.77%. Accuracy values of all aspects are high, but it does not necessarily imply the performance of the built model. The accuracy is not capable of capturing characteristics of the class compared to precision and recall that are capable of capturing characteristics of each class. Hence, F1-Score is used as the value to measure the performance.

2) Recommendation

As the recommendation for future research, the aspect selection can be improved by choosing aspect that could make less bias and more specific to define the adjectives. Aspect 'Product' is excessively general compared to 'Price' that can be very specific whether it is cheap or expensive. Besides that, a review with negation statement is even more difficult to classify the sentiment of the review itself. The unification of several n value for n-gram can be used to handle the negative impact of negation statement.

ACKNOWLEDGMENT

The author would like to thank the Lord Almighty, my family who supports me endlessly, and the people involved in this research such as lecturers and friends. The author would also thank all authors whose papers are used as references in this paper.

REFERENCES

[1] P. R. Iswari and F. Latifah, *Buku Pegangan Ilmu Pengetahuan Kosmetik*. Gramedia Pustaka Utama, 2013.

[2] X. Fang and J. Zhan, "Sentiment analysis using product review data," *J. Big Data*, vol. 2, no. 1, 2015, doi: 10.1186/s40537-015-0015-2.

[3] S. J. Lewis, "Thumbs up," *Am. J. Orthod. Oral Surg.*, vol. 31, no. 9, pp. 481–482, 1945, doi: 10.1016/0096-6347(45)90048-2.

[4] H. Ardian and S. Kosasi, "Analisis Sentimen Pada Review Produk Kosmetik Bahasa Indonesia Dengan Metode Naive Bayes," *J. ENTER*, vol. 2, no. 1, pp. 306–320, 2019.

[5] F. Rosi, M. A. Fauzi, and R. S. Perdana, "Prediksi Rating Pada Review Produk Kecantikan Menggunakan Metode Naïve Bayes dan Categorical Proportional Difference (CPD)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya*, vol. 2, no. 5, pp. 1991–1997, 2018.

[6] T. S. Moh and Z. Zhang, "Cross-lingual text classification with model translation and document translation," *Proc. Annu. Southeast Conf.*, pp. 71–76, 2012, doi: 10.1145/2184512.2184530.

[7] S. Mardianti, M. Zidny, and I. Hidayatulloh, "Ekstraksi Tf-Idf N-Gram Dari Komentar Pelanggan Produk Smartphone Pada Website E-Commerce," pp. 79–84, 2018.

[8] Suyanto, *Data Mining untuk Klasifikasi dan Klasterisasi Data*. Informatika Bandung, 2017.

[9] R. Prabowo and M. Thelwall, "Sentiment analysis: A combined approach," *J. Informetr.*, vol. 3, no. 2, pp. 143–157, 2009, doi: 10.1016/j.joi.2009.01.003.

[10] M. A. Sutanto and A. Aprianingsih, "He Effect of Online Consumer Review Toward Purchase Intention: a Study in Premiumcosmetic in Indonesia," *Int. Conf. Ethics ofBusiness, Econ. Soc. Sci.*, vol. 53, no. 2, pp. 1689–1699, 2016, [Online]. Available: http://journal.unhas.ac.id/index.php/kareba/article/view/346%0Ahttp://administrasibisnis.studentjournal.ub.ac.id/index.php/jab/article/view/2548%0Ahttp://teknonisme.com.

[11] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, and I. Androutsopoulos, "SemEval-2015 Task 12: Aspect Based Sentiment Analysis," pp. 486–495, 2015, doi: 10.18653/v1/s15-2082.

[12] S. Guha, A. Joshi, and V. Varma, "SIEL: Aspect Based Sentiment Analysis in Reviews," no. SemEval, pp. 759–766, 2015, doi: 10.18653/v1/s15-2129.

[13] "Proceeding AIP2017 mubarok Adiwijaya Aspect-based sentiment analysis to review products using Naïve Bayes.en.id.pdf." .

[14] N. Z. Dina and N. Juniarta, "Aspect based Sentiment Analysis of Employee's Review Experience," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 6, no. 1, p. 79, 2020, doi: 10.20473/jisebi.6.1.79-88.

[15] D. D. Mehare, "Introduction to TF-IDF: To Represent Importance of Keyword within whole Dataset," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 6, no. 3, pp. 2321–2323, 2018, doi: 10.22214/ijraset.2018.3369.

[16] D. Jurafsky and J. H. Martin, "Chapter 3: N-Gram Language Models N-Gram Language Models," *Speech Lang. Process.*, 2019.

[17] Prof. Praveen Dhyani and Sonam Mittal, "Multilingual Text Classification," *Int. J. Eng. Res.*, vol. V4, no. 03, pp. 99–101, 2015, doi: 10.17577/ijertv4is030032.

[18] T. Widiyaningtyas, I. A. Elbaith Zaeni, and R. Al Farisi, "Sentiment Analysis Of Hotel Review Using N-Gram And Naive Bayes Methods," *Proc. 2019 4th Int. Conf. Informatics Comput. ICIC 2019*, no. October, 2019, doi: 10.1109/ICIC47613.2019.8985946.

[19] L. Breiman, "ST4_Method_Random_Forest," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1017/CBO9781107415324.004.

[20] B. Purnama, *Pengantar Machine Learning Konsep dan Praktikum dengan Contoh Latihan Berbasis R dan Python*. Informatika Bandung, 2019.

[21] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches," *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 42, no. 4, pp. 463–484, 2012, doi: 10.1109/TSMCC.2011.2161285.

[22] H. H. Parmar, "Sentiment Mining of Movie Reviews using Random Forest with Tuned Hyperparameters," *Conf. Pap.*, no. July, 2014.