# Multi-Label Aspect-Sentiment Classification on Indonesian Cosmetic Product Reviews with IndoBERT Model

Ng Chin Mei[1], Sabrina Tiun[2], Gita Sastria[3]
Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia, Malaysia[1, 2]
Jurusan Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Riau, Indonesia[3]

***Abstract*—For an existing cosmetic company to expand, it is crucial to understand customers' opinions regarding cosmetic products through product reviews. Aspect-based sentiment classification (ABSC), which consists of text representation and classification stages, is typically employed to automatically extract the interested insights from review. Existing studies of ABSC primarily used single-label classification, which fails to capture relationships between multiple aspects in a review. Additionally, the use of contextual embeddings like IndoBERT for representing Indonesian-language cosmetic product reviews has been underexplored. This study addresses these issues by developing a multi-label classification model that leverages IndoBERT, including IndoBERT[b], IndoBERT[k], and IndoBERTweet, to better represent context and capture relationships across multiple aspects in a review. The model is trained and evaluated using a dataset of Indonesian-language cosmetic product reviews from Female Daily. The multi-label models can be constructed using IndoBERT directly as end-to-end model or employing IndoBERT solely as word embedding model. The latter model, also known as conventional multi-label model, needs to be coupled with problem transformation approach and classifier for classification. Single label classification model with Word2Vec serves as baseline to assess the improvement of multi-label model's performance on Female Daily cosmetic product reviews dataset. The empirical results revealed that the multi-label approach was more effective in identifying sentiments for pre-defined aspects in reviews. Among the models, end-to-end IndoBERT[b] achieved the highest accuracy (86.98%), while conventional multi-label models combining IndoBERT[b], Label Powerset (LP), and Support Vector Machine (SVM) performed best with 69.64%. This study is significant as it provides a more generalized understanding of the BERT embedding within the context of multi-labels classification and explores the effect of contextual embedding in the cosmetic domain.**

***Keywords*—Aspect-based sentiment analysis; IndoBERT; multi-label classification; IndoBERTweet; problem transformation***

## I. Introduction

Recently, it can be observed that the worldwide cosmetic industry has generally experienced tremendous growth [1]. Specifically, in Indonesia, the cosmetic industry market size is anticipated to grow from USD 1.17 billion from 2020 to roughly double the amount, which is USD 2.38 billion within eight years period [2].

In order for cosmetic companies to capitalize on these future prospects, it is imperative to understand the customer's needs and opinion, and one of the methods to achieve this is through customer's product reviews, as these heavily influence purchasing decisions [3] [4]. Sentiment analysis is a suitable technique to extract the insights from reviews due to the nature of review itself, which is opinionated with a sentiment polarity, either positive, negative, or neutral. In sentiment analysis, classification of sentiment can be performed at three levels of extraction in terms of granularity, namely document level, sentence level and aspect level. Analyzing reviews at aspect level using aspect-based sentiment analysis (ABSA) is crucial since it identifies sentiments tied to specific product aspects, offering deeper insights than document or sentence-level analysis.

Aspect-based sentiment classification (ABSC) is one of the tasks in ABSA that involves solely sentiment classification. Typically, ABSC is implemented as a pipeline consisting of two stages, namely text representation and classification. The first stage involves the transformation of the text data into its numerical representation. Traditional text representation methods such as Term Frequency - Inverse Document Frequency (TF-IDF) and static word embedding model (e.g., Word2Vec) are relatively inaccurate in representing text as they fail to capture contextual meanings between words [5]. The emergence of contextual word embedding model addresses this limitation. One state-of-the-art contextual model is BERT (Bidirectional Encoder Representations from Transformers). The effectiveness of BERT is demonstrated by the findings of several studies [6] [7] [8] [38].

Subsequently, text representation enables the next stage which is classification. Classification is where sentiment for each aspect of an individual review text can be determined. A notable limitation in the classification stage of ABSC in prior research is the reliance on a single-label classification approach [28] [29] [30] [31] [32]. This method determines the sentiment for each aspect independently, failing to capture correlations between sentiments across different aspects of a review [9]. In real-world scenarios, multiple aspects in a customer review may have related sentiments, and ignoring these dependencies can reduce the performance of the classification model. Although [9] proposed a multi-label approach to account for these correlations, their work did not directly compare it with single-label models, and their dataset was from a different domain. Furthermore, there is a gap in exploring multi-label classification for Indonesian-language datasets, which have

been under-explored compared to more dominant languages like English.

This study focuses on investigating ABSC of cosmetic products reviews written in the Indonesian language. Several issues have been observed in this context. Firstly, there is a scarcity of prior ABSC research in Indonesian cosmetic products reviews. Secondly, none of the prior works have explored the effect of contextual embedding models such as BERT on this domain. Thirdly, existing studies primarily employed single label classification, but this classification approach presents limitations by disregarding possible relationship between multiple aspects within single review [9]. Thus, this study aimed to address these issues by developing a reliable multi-label ABSC model, at the same time, exploring the performance of contextual word embedding in representing words from cosmetic domain. As the focus is on the Indonesian language, IndoBERT, which is a BERT designed for Indonesian language, was employed as the contextual embedding model in this study.

In this study, this multi-label ABSC model was built with two alternative methods, one using IndoBERT as an end-to-end model, and the other built with a combination of text representation method, multi-label problem transformation approach, and machine learning classifier. The former method performs multi-label classification by directly processing the text and generating sentiment predictions within a single model, while the latter method analyses the word vectors transformed by IndoBERT, then combines the multi-label approach with a classifier to categorize the aspect-sentiment labels for each cosmetic product review. The second method was referred to conventional multi-label model. This study is organized as follows: Section II presents the background information and related works. Section III demonstrates the methodology. Section IV presents the results. Discussion is given in Section V. Finally, the paper is concluded in Section V.

## II. Literature Review

### A. IndoBERT

*1) Variants of IndoBERT*: There are three variants of IndoBERT, which are IndoBERT[k], IndoBERT[b], and IndoBERTweet. The key differences among IndoBERT variants lie in the training datasets used. IndoBERT[k], introduced by [10] was trained on the INDOLEM dataset, which consists solely of formal text, limiting its effectiveness with colloquial Indonesian. To address this, [11] introduced IndoBERT[b], trained on the mixed formal and informal dataset INDONLU. Additionally, IndoBERTweet was developed specifically for informal social media language [12].

*2) Word embedding model*: Previous comparative studies consistently highlight the superiority of IndoBERT variants when applied to Indonesian product review datasets across various domains. For example, [13] demonstrated that IndoBERT[b] outperformed Word2Vec when paired with a Convolutional Neural Network (CNN) classifier in categorizing restaurant customer reviews across four dimensions: Price, Food, Place, and Service. This finding is corroborated by study [14], who showed that IndoBERT[k] performed better than both Word2Vec and FastText in representing the reviews related to COVID vaccines. Moreover, the IndoBERT[b] showed more effectiveness than Word2Vec with all seven classifiers such as SVM, Naïve Bayes (NB), and Random Forest (RF) when dealing with hotel reviews in study [15], further supporting the superiority of IndoBERT as word embedding models in transforming the text.

*3) End-to-end model*: IndoBERT also plays a good role as end-to-end model in ABSC as shown in study [16]. In the study of [17], sentiment classification on online reviews was conducted using IndoBERT as end-to-end model, with the aim to investigate the satisfaction of customer towards ride-hailing company Gojek from seven aspects. A relatively high accuracy of 96% was achieved, suggesting the superiority of IndoBERT. Similar promising results were obtained in the study of [18]. Other studies such as [19] and [20] showed the effectiveness of IndoBERT implementations as an end-to-end model when comparing its performance to other deep learning language models and traditional machine learning method.

### B. Multi-Label Classification

Existing research categorizes multi-label approaches into three categories: Problem transformations, algorithm adaptations, and pre-trained language model. Problem transformation methods, such as Binary Relevance (BR), Classifier Chain (CC), and Label Powerset (LP), convert multi-label problems into binary or multi-class problems for classification. BR treats each label as a separate binary problem, while CC predicts labels sequentially, and LP transforms labels into a multi-class problem. RAkEL D, an ensemble of LP, addresses LP's limitations by training on label subsets. Algorithm adaptation modifies existing algorithms, like ML-kNN and ML-DT, to handle multi-label tasks directly.

In most previous works on multi-label classification, emphasis was placed mainly on text categorization [21] [22] [23] [24]. The literature review reveals that there was only a limited of studies on ABSC that specifically address multi-labels classification. Related existing studies primarily assessed the multi-label aspect-sentiment model efficiency through multi-label metrics: Accuracy and Hamming Loss.

The study in [25] explored drug effectiveness using problem transformation methods (BR, CC, LP) combined with various classifiers, finding that SVM performed best, followed by DT and NB. Among the transformation methods, LP outperformed CC and BR. Despite this study evaluating all three popular problem transformation methods, there was a lacking exploration of the algorithms that specifically adapted for multi-label problems.

The research in [9] compared three categories of multi-labels approaches using customer reviews from restaurants, wine, and movies, demonstrating pre-trained language models such as BERT outperformed other categories, followed by problem transformation and algorithm adaptation. For problem transformation, the author obtained the same results as [25]: LP consistently outperformed CC and BR.

The study in [26] further confirmed the findings of [9], demonstrating BERT's superiority in sentiment classification

and showing that LP outperformed BR and CC within problem transformation approaches.

In the Indonesian context, BERT variants, IndoBERT has shown excellent performance in multi-label tasks, outperforming other models, as evidenced by studies like [27] and [15].

### C. Aspect-based Sentiment Classification (ABSC) on Indonesian Cosmetic Product Reviews

Previous ABSC studies of cosmetic product reviews in Indonesian language, primarily using datasets from Female Daily, focused on determining the overall sentiment of user reviews. These studies explored various aspects, with most focusing on four predefined aspects: packaging, quality, scent, and price [28] [29] [30] [31].

Traditional text representation approaches like TF-IDF and NB yielded moderate performance, with studies like [28] achieving an average F1-score of 62.81% across these four aspects. Their results indicated room for improvement, potentially due to the limitations of TF-IDF and NB, which rely on word frequency for representation and independence assumptions for classification.

Subsequent research improved performance by employing Word2Vec static word embeddings and different classifiers, such as SVM, on similar datasets with the same aspects. [29] achieved a 68.25% F1-score using Word2Vec for text representation while maintaining NB as the classifier. The improvement is likely due to Word2Vec enhanced the contextual richness of sentiment representations compared to TF-IDF. The study in [30] demonstrated further improvements by using SVM with TF-IDF as the text representation method.

Other studies, like those by [32], explored additional aspects but found limitations in performance using approaches like Bag of Words (BOW), achieving only a 53.04% F1-score. The study in [31] employed a hybrid method of TF-IDF and semantic similarity, achieving a high accuracy of 90.33%, likely due to fewer predicted aspects.

Notably, none of these studies applied contextual embeddings or addressed multi-label classification in the cosmetic domain, highlighting areas for future improvement in ABSA methods.

## III. Methodology

### A. Research Methods Overview

Generally, there are two strategies that can be used to build a multi-label model using IndoBERT, as shown in Fig. 1. In the first strategy, IndoBERT was used as an end-to-end model to directly perform multi-label classification. In second strategy, IndoBERT was used solely as a word embedding model to transform text into dense word vectors, which were then passed to machine learning classifiers. The model from second strategy, known as the conventional multi-label model, involved using word embeddings for text representation, multi-label problem transformation methods, and classifiers.

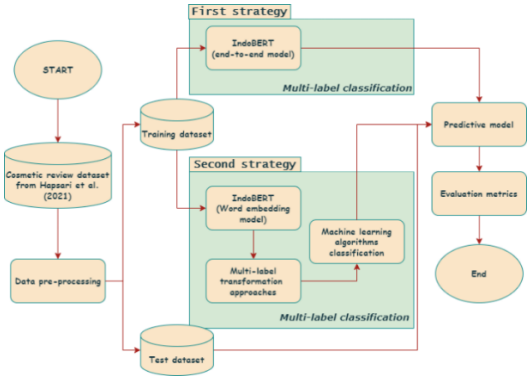


Fig. 1. General workflow for building multi-label models using IndoBERT.

In this study, three interconnected experiments were conducted: -

*1) Experiment I*: This experiment was focused on evaluating the performance of multi-label classification in ABSC context, with single-label model from [29] serving as the baseline. To rule out any external factors which might affect the results, the experiment established the baseline by replicating the literature experiment [29].

*2) Experiment II*: Experiment II concentrated on assessing the different variants of IndoBERT as word embeddings, alongside the promising multi-label approaches identified from experiment I. This experiment exploited Word2Vec as the baseline word embedding. The IndoBERT variants used were IndoBERT[b], IndoBERT[k], and IndoBERTweet.

*3) Experiment III*: Building upon the findings of Experiment I and II, this experiment further explored the performance of IndoBERT in multi-label classification by incorporating the multi-label capabilities into model, using IndoBERT directly as end-to-end model, serving as both text representation and classifier. Conventional multi-label models with classifiers such as Gaussian NB, SVM, Linear SGD, and RF were also employed for comparison.

### B. Dataset

The study investigated a secondary dataset that was exclusively sourced from journal article published by [29]. The dataset consists of a total of 3960 customer reviews on cosmetic products collected from Female Daily website and written in Indonesian language. Each review was pre-annotated with sentiments across four different aspects, with the sentiment distribution summarized in Table I.

TABLE I. Sentiment Distribution Across Four Aspects in Cosmetic Product Reviews

| Aspect | Sentiment Count | | |
|---|---|---|---|
| | ***Positive*** | ***Negative*** | ***Neutral*** |
| Product | 659 | 688 | 2612 |
| Packaging | 447 | 189 | 3323 |
| Price | 1056 | 716 | 2187 |
| Scent | 669 | 218 | 3072 |

### C. Exploratory Data Analysis (EDA)

Before data modeling, a Chi-square test was conducted during the EDA stage to assess dependencies between the targeted aspects. This test provided insights into the relationships between aspect-sentiments, helping determine if single-label or multi-label classification would be more suitable. In this study, the null hypothesis claimed that there is no significant association between the targeted aspects in the investigated dataset. A p-value threshold of 0.05 was used; any value below this indicated significant associations between aspects, suggesting the need for a multi-label classification approach in the dataset.

### D. Preprocessing

For multi-label classification, the aspect columns were transformed into aspect-sentiment labels for further analysis. Each aspect was further divided into three labels, following the patterns of 'aspect_pos,' 'aspect_neg,' and 'aspect_other', with the number 0 or 1 indicating the presence of each aspect-sentiment label.

### E. Word Embedding (Text Representation) using IndoBERT

Before being processed by IndoBERT, the review text must first pass through the IndoBERT tokenizer to obtain special input format. In this stage, the input sentence was concurrently tokenized and added with special tokens of [CLS] and [SEP] tokens at the beginning and the end of tokenized sequence respectively. For the input length, the maximum sentence length was limited to 128 tokens in this study. After that, these modified sequences were fed into token, segment, and position embeddings sequentially to generate the initial input embedding, which in turn fed into encoder layers within IndoBERT for further processing.

In IndoBERT, the initial embedding of each token was passed through multiple sub-layers consisting of multi-head self-attention and Feed-forward Network to incorporate the contextual information. The final output contextualized embedding was extracted through a mean pooling strategy. The overview of the word embedding process for IndoBERT is illustrated in Fig. 2.

### F. Multi-Label Classification

*1) Conventional multi-label classification*: To build conventional multi-label models, the output contextualized embeddings and target labels were transformed using problem transformation methods. Binary Relevance (BR), Classifier Chain (CC), and Label Powerset (LP) were used to convert the data into binary or multi-class problems. Classifiers such as Gaussian Naïve Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), and Linear Stochastic Gradient Descent (SGD) were then applied for classification.

*2) IndoBERT (As end-to-end model)*: To perform multi-label classification directly using IndoBERT, an additional classifier layer was added at the uppermost level of the model. Final output embedding from mean pooling layer was fed into the classifier layer directly for multi-label classification as shown in Fig. 3. In this strudy, the hyperparameters values of classifier layer mirrored the literature findings of [33], which set to "learning rate = 2e-5, batch size = 8, and epoch =5", optimizing with Adam optimizer. Because IndoBERT was designed to perform multi-label classification, sigmoid function was used rather than SoftMax in classification layer. The formula of sigmoid function is shown in (1). The output probability from sigmoid activation function for each aspect-label is a real number within the range of 0 to 1. Given that the study applied the default threshold value of 0.5, any predicted probability of the label greater than 0.5 was referred to present and less than 0.5 was considered as absent.

$$\sigma(x)= 1/(1+e^{(-x)}) \quad (1)$$

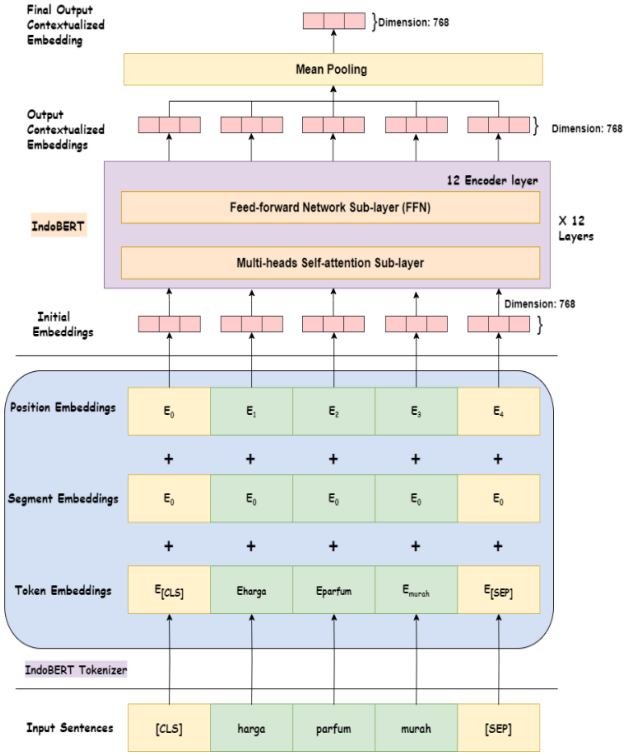


Fig. 2. Overview of word embedding process for input sentence by IndoBERT.

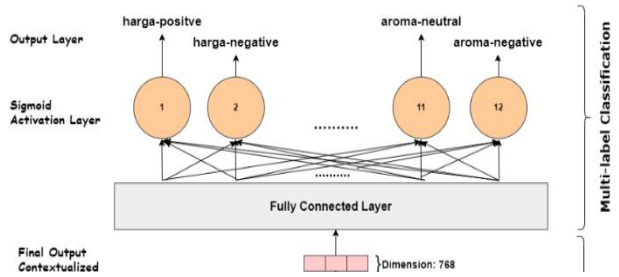


Fig. 3. Multi-label classification process within IndoBERT.

### G. Evaluation Metrics

In this study, several evaluation metrics were employed to evaluate the sentiment classification model performance.

*1) Accuracy per label*: This metric assesses the number of correctly predicted labels over the total number of instances,

computing using (2), where TP and TN refer to True Positive and True Negative, respectively. Given that it evaluates the model's correctness on individual labels (column) independently, accuracy is commonly used in single label classification. In this study, this metric was utilized for comparing the performance of multi-label with single-label models in Experiment I.

$$\text{Accuracy per label} = (TP + TN) / Total \quad (2)$$

*2) Accuracy*: This accuracy metric calculates the probability of the correctly classified labels by considering the overlap between the true data, $Y_i$ and predictive data, $Z_i$ for each instance, as computed by the (3). In this study, this metric was used as one of the main references in assessing the multi-label sentiment classification model performance in Experiment II and III.

$$Accuracy = 1/N * \Sigma | Yi \cap Zi | / | Yi \cup Zi | \quad (3)$$

*3) Hamming loss*: Hamming Loss is the prominent evaluation metric of multi-label classification that measures the proportion of wrongly classified labels over all instances. In contrast with typical metrics, the lower the value hamming loss represents the higher performance of model. The formula for computing hamming loss is shown in (4). It was used to evaluate the multi-label model performance in this study.

$$\text{Hamming Loss} = 1/N * \Sigma | Yi \, \Delta \, Zi | \quad (4)$$

*4) Micro-F1 score*: Micro F1-score is a harmonic mean of precision and recall calculated based on classes, shown in (5), where P and R represent precision and recall respectively. Precision measures the proportion of the chosen items that are correct over the actual instances that are predicted as chosen while recall is a metric uses to gauge the percentage of correct items that are selected. The formula of precision and recall is shown in (6) and (7).

$$F1-score = 2PR/(P+R) \quad (5)$$

$$Precision = 1/N * \Sigma | Yi \cap Zi | / | Zi | \quad (6)$$

$$Recall = 1/N * \Sigma | Yi \cap Zi | / | Yi | \quad (7)$$

## IV. Results

### A. (EDA) Results

Table II presents the Chi-square test results between aspects, showing that most p-values were below 0.05, indicating variable dependency. From the results, it can be observed that nearly all the resulting p-values were less than threshold value 0.05.

TABLE II. Chi-Square Test Results Between Different Aspects of Cosmetic Product

| Variable 1 | Variable 2 | p value | Null hypothesis |
|---|---|---|---|
| harga | pengeamsan | 0.0147 | Rejected |
| | produk | 6.5113e-24 | Rejected |
| | aroma | 3.4954e-12 | Rejected |
| pengemasan | produk | 4.3069e-13 | Rejected |
| | aroma | 0.0689 | Accepted |
| produk | aroma | 1.2633e-25 | Rejected |

### B. Experimental Results I

In this experiment, the baseline single label classification model was replicated based on the methodology from [29]. The method primarily employed Word2Vec to vectorize the text reviews and utilized Gaussian NB to classify the sentiment for each aspect independently. Given that there were four aspects, the model iterated four times to complete the prediction for all aspects. Table III presents the empirical outcomes of baseline from our experiment and [29].

TABLE III. Comparison of Accuracy Per Label of Baseline Single Label Model and Our Experiment

| Source | Average Accuracy per Label (%) | Accuracy per Label (%) | | | |
|---|---|---|---|---|---|
| | | ***Price*** | ***Packaging*** | ***Product*** | ***Aroma*** |
| [29] | 68.17 | 70.96 | 68.79 | 56.36 | 76.57 |
| Our experiment | 62.47 | 60.33 | 74.39 | 48.18 | 66.99 |

From Table III, it shows that there was approximately 6% reduction in average accuracy when comparing the replicated model with [29]. The differences can be attributed to factors such as differences in the parameters setting of Gaussian NB as [29] did not include any details about their parameters while this study proceeded with default settings. To mitigate the influences from external factors, an accuracy of 62.47% is used as a reference value for comparing the performance of multi-label models.

Fig. 4 shows that multi-label models perform differently in determining sentiment for Indonesian cosmetic product reviews. Models using BR and CC problem transformation methods had performance comparable to the baseline. The BR model achieved the same accuracy of 62.47% as the baseline, likely due to their similar classification approach. However, the other two multi-models with LP and RAkEL D, exhibited notable enhancement in accuracy over the baseline single label model, achieving 70.18% and 70.16%, respectively.

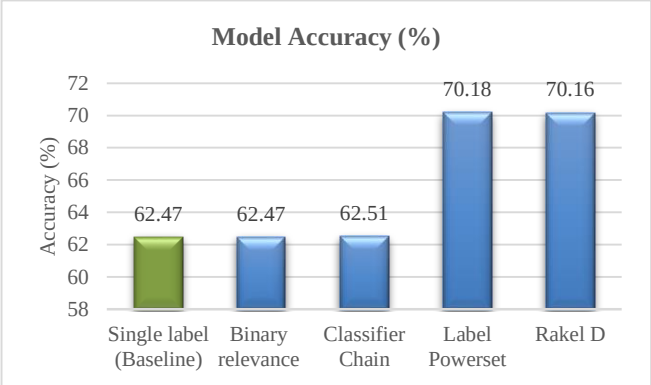


Fig. 4. Average of accuracy per label for single label and multi-label models.

### C. Experimental Results II

In Experiment II, the study evaluates the performance of IndoBERT as word embeddings in representing the words from cosmetic domain, using Word2Vec as the baseline. Building upon the findings of Experiment I, LP and RAkEL D, each was used to transform the multi-label problem in the model while Gaussian NB classifier was utilized to classify the aspect-sentiment labels to each review, along with Word2Vec. The IndoBERT word embedding model was employed in parallel, aiming to determine its effectiveness compared to Word2Vec. The results are summarized in Table IV.

TABLE IV. PERFORMANCE OF MULTI-LABEL MODELS USING DIFFERENT EMBEDDING MODELS IN EACH MULTI-LABEL APPROACH

| Transformation Approach | IndoBERT version | Accuracy (%) | Hamming Loss (%) | Micro F1-score (%) |
|---|---|---|---|---|
| Label Powerset | Word2Vec | 54.75 | 22.27 | 66.66 |
| | IndoBERT[b] | 60.39 | 19.15 | 71.28 |
| | IndoBERT[k] | 47.70 | 26.18 | 60.73 |
| | IndoBEETweet | 61.10 | 18.72 | 71.91 |
| RAkEL D | Word2Vec | 53.60 | 24.96 | 66.12 |
| | IndoBERT[b] | 57.79 | 20.81 | 69.97 |
| | IndoBERT[k] | 41.86 | 31.87 | 56.21 |
| | IndoBERTweet | 57.55 | 21.09 | 69.76 |

Overall, both IndoBERT[b] and IndoBERTweet consistently outperformed the baseline Word2Vec embedding model, regardless of the problem transformation approaches used in classification model. For models with LP approach, there was an approximate 5% to 6% improvement in accuracy and micro F1-score when comparing the multi-label models with Word2Vec to IndoBERT[b] and IndoBERTweet. On the other hand, the models employing RAkEL D exhibited only a 2% to 3% enhancement in accuracy and micro F1-score. For hamming loss in the models employing both LP and Rakel D, an approximately range of 3% to 4% reduction was observed when comparing the baseline with each of the IndoBERT[b] and IndoBERTweet, indicating a decrease in misclassification occurrences in models employing both methods.

### D. Experimental Results III

This experiment evaluated IndoBERT's performance in direct multi-label classification. Building on results from Experiments I and II, conventional multi-label models using IndoBERT[b] and IndoBERTweet were developed with various classifiers. The results are shown in Tables V and VI, while Table VII presents IndoBERT's empirical performance.

TABLE V. PERFORMANCE OF CONVENTIONAL MULTI-LABEL MODEL USING INDOBERT[B] AS TEXT REPRESENTATION METHOD

| Model | Accuracy (%) | Hamming Loss (%) | Micro F1-score (%) |
|---|---|---|---|
| Label Powerset + NB | 60.45 | 19.12 | 71.33 |
| Label Powerset + SVM | 69.64 | 14.20 | 78.69 |
| Label Powerset + RF | 63.3 | 17.45 | 73.82 |
| Label Powerset + SGD | 65.48 | 16.36 | 75.47 |
| RakEL D + NB | 56.81 | 21.43 | 68.90 |
| RakEL D + SVM | 68.31 | 14.26 | 78.59 |
| RakEL D + RF | 62.93 | 16.60 | 74.18 |
| RakEL D + SGD | 65.31 | 15.99 | 75.94 |

TABLE VI. PERFORMANCE OF CONVENTIONAL MULTI-LABEL MODEL USING INDOBERTWEET AS TEXT REPRESENTATION METHOD

| Model | Accuracy (%) | Hamming Loss (%) | Micro F1-score (%) |
|---|---|---|---|
| Label Powerset + NB | 61.15 | 18.70 | 71.95 |
| Label Powerset + SVM | 68.14 | 14.92 | 77.61 |
| Label Powerset + RF | 62.54 | 17.82 | 73.26 |
| Label Powerset + SGD | 63.38 | 17.54 | 73.69 |
| RakEL D + NB | 56.71 | 21.49 | 69.13 |
| RakEL D + SVM | 66.88 | 14.86 | 77.57 |
| RakEL D + RF | 61.94 | 17.42 | 73.26 |
| RakEL D + SGD | 64.71 | 16.35 | 75.75 |

TABLE VII. PERFORMANCE OF MULTI-LABEL MODELS WITH INDOBERT AS END-TO-END MODEL

| Model | Accuracy (%) | Hamming Loss (%) | Micro F1-score (%) |
|---|---|---|---|
| IndoBERT[b] | 86.98 | 5.45 | 91.70 |
| IndoBERTweet | 86.21 | 5.85 | 91.12 |

The results show that the end-to-end model significantly outperformed the conventional multi-label models. It can be observed that there was a significant enhancement in end-to-end model performance when using IndoBERT[b] and IndoBERTweet, with accuracy increasing by approximately 18% to 30% and the micro F1-score by 13% to 22%, compared to both the highest and lowest performing conventional models based on the same IndoBERT embedding. In terms of hamming loss, using IndoBERT[b] and IndoBERTweet directly for multi-label classification reduced significantly (8% to 16%) in classifying wrongly the aspect-sentiment label for cosmetic product review.

## V. DISCUSSION

The results of p-values from EDA showed that most of the null hypotheses were rejected, suggesting that the variables are dependent on each other. These dependencies indicate the presence of an aspect's sentiment might affect the prediction of sentiment of another aspect. Given this dependency, it implies a need to explore the Indonesian cosmetic product review dataset with multi-label model.

There are three categories of multi-label classification methods: Problem transformation, algorithm adaptations, and pre-trained language model. For the problem transformation, the BR model transforms the multi-label task into 12 single-label problems, performing independent classification for each label, much like the baseline. Unlike BR, multi-label model with CC

problem transformation method considered label correlation into account. However, it still demonstrated comparable accuracy with 62.51% to baseline. One of the possible reasons might be the influences by the arbitrary arrangement of labels, which could lead to the poor performance in CC model [34] [35]. In contrast with BR and CC, the other two multi-models with LP and RAkEL D, exhibited notable enhancement in accuracy over the baseline single label model, achieving 70.18% and 70.16%, respectively. There was approximately 8% improvement of accuracy in multi-label models with LP and Rakel D methods over baseline. This suggests LP and RAkEL D are more suitable for multi-aspect sentiment classification. The improvement is probably due to both multi-label models considering label dependencies, with LP capturing co-occurrence relationships by converting the problem into combinations of labels, and RAkEL D enhancing performance by training Gaussian NB on distinct label subsets, improving label correlation handling.

In terms of text representation, as expected, IndoBERT outperformed Word2Vec as a word embedding model due to its architectural design. Word2Vec, using the CBOW architecture with a fixed window produces static embeddings, which lack contextual information. In contrast, IndoBERT generates contextualized embeddings by learning from masked tokens, capturing semantic relationships and nuances. This contextual richness allows IndoBERT to provide more refined input vectors, enabling the Gaussian NB classifier to more accurately classify aspect-sentiment labels. IndoBERT[k] performed worse than the baseline Word2Vec, indicating that contextualized embeddings don't always outperform static embeddings, as seen in study [36]. The varying performance among IndoBERT variants may stem from differences in pre-training datasets. IndoBERT[k], trained primarily on formal text, struggles with the informal language in product reviews. In contrast, IndoBERT[b] and IndoBERTweet, pre-trained on informal data like social media, are better equipped to handle the mixed linguistic style found in the reviews.

Given that multi-label classification can be performed directly using pre-trained language models as end-to-end model, this paper proposes using IndoBERT and IndoBERTweet for classification, as these two pre-trained language models have demonstrated high accuracy in terms of text representation. The results show that the proposed end-to-end model significantly outperformed the conventional multi-label models. These findings align with outcomes from [9] and [15], suggesting that IndoBERT generalizes multi-label classification tasks better than conventional models. This is likely due to the architectural differences between IndoBERT and traditional classifiers. IndoBERT, using neural networks, is more complex, flexible, and better equipped to handle intricate patterns in the dataset compared to simpler linear, tree-based, or probabilistic classifiers like SGD, SVM, RF, and Gaussian NB [37].

For conventional multi-label models, it can be observed that a similar trend of classifier performance was exhibited across multi-label transformation approach and word embedding. Notably, SVM consistently demonstrated superiority in correctly multi-classifying labels, followed by linear SGD, RF and Gaussian NB. These results aligned with classification results for two out of three datasets in the study of [9], which found that SVM outperformed linear SGD, followed by RF, regardless of whether LP or Rakel D was used.

## VI. Conclusion

This study developed a reliable multi-label ABSC model while exploring the performance of contextual word embedding in representing words from the Indonesian cosmetic domain. Three experimental experiments evaluated different multi-label classification approaches. The results showed that IndoBERT[b] and IndoBERTweet provided more refined text representation, improving performance approximately by 2% to 6% compared to Word2Vec. The findings also demonstrated that multi-label models using IndoBERT as an end-to-end model outperformed conventional methods. IndoBERT[b] achieved the best accuracy of 86.98%, showing a 17.34% to 30.27% improvement over the baseline, confirming its superiority for multi-label classification in this domain. Although this study demonstrated notable improvements, certain limitations remain. The study investigated only one type of contextual embedding model, IndoBERT. To further enhance the multi-label model, future work could explore other contextual embeddings, such as DistilBERT and RoBERTa. Additionally, the hyperparameter settings for the end-to-end IndoBERT model were restricted to "learning rate = 2e-5, batch size = 8, and epoch = 5." Future research could experiment with different hyperparameter combinations, as there is no one-size-fits-all setting for optimizing the model.

## Acknowledgment

The authors would like to acknowledge the Ministry of Higher Education (MoHE) for supporting this research through the Fundamental Research Grant Scheme (FRGS), under grant number FRGS/1/2020/ICT02/UKM/02/1.

## References

[1] [A. Berg and S. Hudson, “The beauty market in 2023: A special state of Fashion report,” May 2023. Accessed: May 14, 2024. [Online]. Available: https://www.mckinsey.com/industries/retail/our-insights/the-beauty-market-in-2023-a-special-state-of-fashion-report#/

[2] Satista, “Cosmetics - Indonesia.” Accessed: May 23, 2024. [Online]. Available: https://www.statista.com/outlook/cmo/beauty-personal-care/cosmetics/indonesia

[3] K. Saleh, “The importance of online customer reviews,” Invesp. Accessed: Apr. 24, 2024. [Online]. Available: https://www.invespcro.com/blog/the-importance-of-online-customer-reviews-infographic/

[4] Boxwell, “Why customer reviews are important for business,” Boxwell. Accessed: Apr. 24, 2024. [Online]. Available: https://boxwell.co/why-customer-reviews-are-important-for-business/#:~:text=86%25%20of%20people%20will%20hesitate,increase%20in%20a%20business’s%20revenue.&text=Customer%20reviews%2C%20whether%20it’s%20a,inspire%20confidence%20in%20your%20brand.

[5] D. S. Asudani, N. K. Nagwani, and P. Singh, “Impact of word embedding models on text analytics in deep learning environment: a review,” Artif Intell Rev, vol. 56, no. 9, pp. 10345–10425, Sep. 2023, doi: 10.1007/s10462-023-10419-1.

[6] K. Ethayarajh, “How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings,” in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 55–65. doi: 10.18653/v1/D19-1006.

[7] N. Lakshmidevi, S. Keshari Swain, and M. Vamsikrishna, "A hybrid enhancing aspect-based sentiment analysis with BERT for aspect extraction and diverse ml classifiers," in 2023 International Conference on Network, Multimedia and Information Technology, NMITCON 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/NMITCON58196.2023.10275957.

[8] Y. Peng, S. Yan, and Z. Lu, "Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets," in Proceedings of the 18th BioNLP Workshop and Shared Task, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 58–65. doi: 10.18653/v1/W19-5006.

[9] J. Tao and X. Fang, "Toward multi-label sentiment analysis: a transfer learning based approach," J Big Data, vol. 7, no. 1, Dec. 2020, doi: 10.1186/s40537-019-0278-0.

[10] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP," Online, 2020. [Online]. Available: https://huggingface.co/

[11] B. Wilie et al., "IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding," 2020. [Online]. Available: https://github.com/annisanurulazhar/absa-playground

[12] F. Koto, J. H. Lau, and T. Baldwin, "IndoBERTweet: A pretrained language model for Indonesian Twitter with effective domain-specific vocabulary initialization," in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 10660–10668. [Online]. Available: https://huggingface.co/huseinzol05/

[13] P. R. Amalia and E. Winarko, "Aspect-based sentiment analysis on Indonesian restaurant review using a combination of convolutional neural network and contextualized word embedding," IJCCS (Indonesian Journal of Computing and Cybernetics Systems), vol. 15, no. 3, pp. 285–294, Jul. 2021, doi: 10.22146/ijccs.67306.

[14] C. B. P. Putra, D. Purwitasari, and A. B. Raharjo, "Stance Detection on Tweets with Multi-task Aspect-based Sentiment: A Case Study of COVID-19 Vaccination," International Journal of Intelligent Engineering and Systems, vol. 15, no. 5, pp. 515–526, Oct. 2022, doi: 10.22266/ijies2022.1031.45.

[15] N. K. Nissa and E. Yulianti, "Multi-label text classification of Indonesian customer reviews using bidirectional encoder representations from transformers language model," International Journal of Electrical and Computer Engineering, vol. 13, no. 5, pp. 5641–5652, Oct. 2023, doi: 10.11591/ijece.v13i5.pp5641-5652.

[16] A. Jazuli, Widowati, and R. Kusumaningrum, "Aspect-based sentiment analysis on student reviews using the Indo-Bert base model," E3S Web of Conferences, vol. 448, p. 02004, Nov. 2023, doi: 10.1051/e3sconf/202344802004.

[17] N. Mahfudiyah and A. Alamsyah, "Understanding user perception of ride-hailing services sentiment analysis and topic modelling using IndoBERT and BERTopic," in 2023 International Conference on Digital Business and Technology Management, ICONDBTM 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICONDBTM59210.2023.10327320.

[18] H. Imaduddin, F. Yusfida A'la, and Y. S. Nugroho, "Sentiment analysis in Indonesian healthcare applications using IndoBERT approach," 2023. [Online]. Available: www.ijacsa.thesai.org

[19] E. I. Setiawan, L. Kristianto, A. T. Hermawan, J. Santoso, K. Fujisawa, and M. H. Purnomo, "Social media emotion analysis in Indonesian using fine-tuning BERT model," in 3rd 2021 East Indonesia Conference on Computer and Information Technology, EIConCIT 2021, Institute of Electrical and Electronics Engineers Inc., Apr. 2021, pp. 334–337. doi: 10.1109/EIConCIT50028.2021.9431885.

[20] L. F. Simanjuntak, R. Mahendra, and E. Yulianti, "we know you are living in Bali: Location prediction of Twitter users using BERT language model," Big Data and Cognitive Computing, vol. 6, no. 3, Sep. 2022, doi: 10.3390/bdcc6030077.

[21] N. Endut, W. M. A. F. W. Hamzah, I. Ismail, M. Kamir Yusof, Y. Abu Baker, and H. Yusoff, "A systematic literature review on multi-label classification based on machine learning algorithms," TEM Journal, vol. 11, no. 2, pp. 658–666, May 2022, doi: 10.18421/TEM112-20.

[22] H. Setiawan, C. Fatichah, and A. Saikhu, "Multilabel classification of student feedback data using BERT and machine learning methods," in 2023 14th International Conference on Information and Communication Technology and System, ICTS 2023, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 147–152. doi: 10.1109/ICTS58770.2023.10330849.

[23] R. K. Shah, S. Kumar, and Shashank, "Multilabel news category classification using machine learning," in Proceedings of the 8th International Conference on Communication and Electronics Systems, ICCES 2023, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 1245–1250. doi: 10.1109/ICCES57224.2023.10192826.

[24] N. K. Singh and S. Chand, "Machine learning-based multilabel toxic comment classification," in 3rd IEEE 2022 International Conference on Computing, Communication, and Intelligent Systems, ICCCIS 2022, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 435–439. doi: 10.1109/ICCCIS56430.2022.10037626.

[25] J. Ashok Kumar, S. Abirami, and T. E. Trueman, "Multilabel aspect-based sentiment classification for abilify drug user review," in Proceedings of the 11th International Conference on Advanced Computing, ICoAC 2019, Institute of Electrical and Electronics Engineers Inc., Dec. 2019, pp. 376–380. doi: 10.1109/ICoAC48765.2019.246871.

[26] Z. Jin, X. Lai, and J. Cao, "Multi-label sentiment analysis base on BERT with modified TF-IDF," in 2020 IEEE International Symposium on Product Compliance Engineering-Asia (ISPCE-CN), IEEE, Nov. 2020, pp. 1–6. doi: 10.1109/ISPCE-CN51288.2020.9321861.

[27] R. Rivaldo, A. Amalia, and D. Gunawan, "Multilabeling Indonesian toxic comments classification using the Bidirectional Encoder Representations of Transformers model," in 2021 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA), IEEE, Nov. 2021, pp. 22–26. doi: 10.1109/DATABIA53375.2021.9650126.

[28] C. H. Yutika, A. Adiwijaya, and S. Al Faraby, "Analisis sentimen berbasis aspek pada review Female Daily menggunakan TF-IDF dan Naïve Bayes [Aspect-based sentiment analysis on Female Daily reviews using TF-IDF and Naïve Bayes]," Jurnal Media Informatika Budidarma, vol. 5, no. 2, p. 422, Apr. 2021, doi: 10.30865/mib.v5i2.2845.

[29] C. C. P. Hapsari, W. Astuti, and M. D. Purbolaksono, "Naive Bayes Classifier and Word2Vec for sentiment analysis on Bahasa Indonesia cosmetic product reviews," in 2021 International Conference on Data Science and Its Applications, ICoDSA 2021, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 22–27. doi: 10.1109/ICoDSA53588.2021.9617544.

[30] N. P. Arthamevia, Adiwijaya, and M. D. Purbolaksono, "Aspect-based sentiment analysis in beauty product reviews using TF-IDF and SVM algorithm," in 2021 9th International Conference on Information and Communication Technology, ICoICT 2021, Institute of Electrical and Electronics Engineers Inc., Aug. 2021, pp. 197–201. doi: 10.1109/ICoICT52021.2021.9527489.

[31] I. Salsabila and Y. Sibaroni, "Multi aspect sentiment of beauty product reviews using SVM and semantic similarity," Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), vol. 5, no. 3, pp. 520–526, Jun. 2021, doi: 10.29207/resti.v5i3.3078.

[32] S. Liviani Mahfiz and A. Romadhony, "Aspect-based opinion mining on beauty product reviews," 2020, [Online]. Available: https://github.com/syitilv/Opinion-Mining

[33] M. R. Mahardika, I. P. J. Wijaya, A. R. Prayoga, H. Lucky, and I. A. Iswanto, "Exploring the performance of BERT models for multi-label hate speech detection on Indonesian Twitter," in 2023 4th International Conference on Artificial Intelligence and Data Sciences: Discovering Technological Advancement in Artificial Intelligence and Data Science, AiDAS 2023 - Proceedings, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 256–261. doi: 10.1109/AiDAS60501.2023.10284596.

[34] A. Kleczewski, "Multilabel classification using a classifier chain," Scikit learn. Accessed: May 08, 2024. [Online]. Available: https://scikit-learn.org/stable/auto_examples/multioutput/plot_classifier_chain_yeast.html#multilabel-classification-using-a-classifier-chain

[35] Serafín. Moral-García, J. G. Castellano, C. J. Mantas, and J. Abellán, "A new label ordering method in Classifier Chains based on imprecise

probabilities," Neurocomputing, vol. 487, pp. 34–45, May 2022, doi: 10.1016/j.neucom.2022.02.048.

[36] P. C.-I. Pang, "Performance evaluation of text embeddings with online consumer reviews in retail sectors," in 2022 IEEE/ACIS 22nd International Conference on Computer and Information Science (ICIS), IEEE, Jun. 2022, pp. 170–175. doi: 10.1109/ICIS54925.2022.9882478.

[37] F. Mota, "Gradient Boosted Machines vs. Transformers (the BERT Model) with KNIME," Medium. Accessed: May 11, 2024. [Online].

[38] S. Alshattnawi, A. Shatnawi, A. M. R. AlSobeh, and A. A. Magableh, "Beyond word-based model embeddings: Contextualized representations for enhanced social media spam detection," Applied Sciences, vol. 14, no. 6, p. 2254, Mar. 2024, doi: 10.3390/app14062254.