Available online at www.ijournalse.org

Emerging Science Journal

(ISSN: 2610-9182)

Vol. 7, No. 1, February, 2023

Modified EDA and Backtranslation Augmentation in Deep Learning Models for Indonesian Aspect-Based Sentiment Analysis

Natasya [1*], Abba Suganda Girsang [1]

[1] *Computer Science Department, BINUS Graduate Program – Master of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia.*

Abstract

In the process of developing a business, aspect-based sentiment analysis (ABSA) could help extract customers' opinions on different aspects of the business from online reviews. Researchers have found great prospective in deep learning approaches to solving ABSA tasks. Furthermore, studies have also explored the implementation of text augmentation, such as Easy Data Augmentation (EDA), to improve the deep learning models' performance using only simple operations. However, when implementing EDA to ABSA, there will be high chances that the augmented sentences could lose important aspects or sentiment-related words (target words) critical for training. Corresponding to that, another study has made adjustments to EDA for English aspect-based sentiment data provided with the target words tag. However, the solution still needs additional modifications in the case of non-tagged data. Hence, in this work, we will focus on modifying EDA that integrates POS tagging and word similarity to not only understand the context of the words but also extract the target words directly from non-tagged sentences. Additionally, the modified EDA is combined with the backtranslation method, as the latter has also shown quite a significant contribution to the model's performance in several research studies. The proposed method is then evaluated on a small Indonesian ABSA dataset using baseline deep learning models. Results show that the augmentation method could increase the model's performance on a limited dataset problem. In general, the best performance for aspect classification is achieved by implementing the proposed method, which increases the macro-accuracy and F1, respectively, on Long Short-Term Memory (LSTM) and Bidirectional LSTM models compared to the original EDA. The proposed method also obtained the best performance for sentiment classification using a convolutional neural network, increasing the overall accuracy by 2.2% and F1 by 3.2%.

Keywords:

Easy Data Augmentation;
Backtranslation;
Long Short-Term Memory;
Bidirectional LSTM;
Convolutional Neural Network.

Article History:

Received:	17	December	2021
Revised:	21	August	2022
Accepted:	04	October	2022
Available online:	07	November	2022

1- Introduction

Customer reviews posted on online platforms bring abundant benefits for business owners and play a significant role in developing existing businesses. Customer reviews may consist of the customers' emotions, suggestions, and opinions, as well as discussions on specific aspects of products or services [1]. Through analyzing online customer reviews, business owners are able to understand how customers feel about certain products or services. Hence, by utilizing customer reviews, business owners could obtain insights to improve their businesses. This objective could be achieved through aspect-based sentiment analysis (ABSA). ABSA combines sentiment prediction with aspect extraction. Aspect extraction refers to the classification of customer review categories, which could include price, service, product, and others. Further in ABSA, the sentiment will be classified based on the relevant aspects [2]. That way, business owners could understand customers' feelings regarding specific aspects of their business. In order to support better analysis performance, deep learning models have become a promising solution.

* **CONTACT**: natasya004@binus.ac.id

DOI: http://dx.doi.org/10.28991/ESJ-2023-07-01-018

Following the evolution of artificial neural networks, deep learning has shown remarkable performance in natural language processing, especially text classification tasks [3]. Many researchers have also implemented deep learning models for ABSA. There are several approaches taken to building an ABSA deep learning model. Some researchers have proposed hybrid methods, such as the Aspect-level Recurrent Convolutional Neural Network (AARCNN) model, which combines bidirectional Long Short-Term Memory (Bi-LSTM) with Convolutional Neural Network (CNN) [4]. While another approach separates ABSA into two tasks, aspect classification and sentiment classification, Ilmania et al. [5] proposed the implementation of Gated-Recurrent Unit (GRU) and fully connected layer for aspect classification, then proceeded to classify sentiment using Bidirectional Gated Recurrent Unit (Bi-GRU) and CNN. Although good results could be produced by relying on deep learning models, the training dataset is also an important factor to consider [6]. When the dataset is not large enough, the models are not able to gain as much information to help the learning process [7]. Therefore, additional training data would be helpful for the model, as it could provide a greater variety of examples to learn. One approach to obtaining additional training data is to collect it manually, however it could be inefficient [8]. Thus, an automated process would be a better option, which is implementing text augmentation.

Text augmentation refers to the process of replicating sentences in the dataset into new sentences that carry the same meaning but with different choices of words or even structure. The methods for text augmentation range from generating new sentences using a pre-trained model, backtranslation, and methods such as n-grams and Latent Dirichlet Allocation (LDA). Wei and Zou [6] have introduced Easy Data Augmentation (EDA) as a simpler approach to text augmentation, involving less complicated operations. They evaluated the results on different classification tasks, which are not ABSA and obtained an increase in the model's performance. An extension to this research was done by proposing EDA operation adjustments to fit the Hybrid Approach for Aspect-Based Sentiment Analysis (HAABSA) model, and the results showed that the adjusted EDA improved the model's performance better than other compared methods [9]. In Liesting et al. [9], the dataset used already includes target words that represent the aspect class, and those words are not augmented to preserve the truth of the label. Additionally, due to EDA's adaptability to language diversity, another research has implemented EDA on Vietnamese datasets for machine learning classifiers [8]. However, the performance of EDA in the Vietnamese dataset is not as outstanding, which could be highly caused by EDA's random operations. The research also further shows the different effects of EDA on a different language. This prompts us to investigate EDA's impact on another language, which is Indonesian.

In this work, we intend to implement EDA on a small Indonesian aspect-based sentiment dataset. Considering the nature of the aspect-based sentiment dataset, some constraints must be fulfilled to retain the sentences' meaning and labels when implementing EDA to minimize the random characteristics of its operations [9]. Apart from that, the dataset used in this work does not have target words tagged in the training data. Hence, we also modified the EDA operations to follow specific rules and, additionally, to help extract aspect along with sentiment-related terms as the target words directly using Part-of-Speech (POS) tagging and word similarity. Based on several other data augmentation studies [8,10], backtranslation method has also constantly proven to be significant for improving a model's performance, thus, we also propose to combine the modified EDA and backtranslation at once. The augmentation is implemented to boost the performance of baseline models, which are Long Short-Term Memory (LSTM) and Bi-LSTM models for aspect classification, with CNN model for sentiment classification. This work also further investigates the effects of the modification and backtranslation integration compared to the original EDA.

2- Literature Review

The ABSA task has been done extensively in many research with various classifier implementations. The task itself may consist of several subtasks. A research has implemented both Support Vector Machine (SVM) and Recurrent Neural Network (RNN) separately to achieve three subtasks of aspect-based sentiment analysis using the SemEval-ABSA16 containing Arabic hotel reviews [11]. The subtasks include identifying aspect category and aspect sentiment polarity, as well as extracting opinion target expression. These basically summarize the objectives of ABSA, which involves detecting specific aspects in the sentence, determining words that represent a certain sentiment, and finally pair the polarity of the sentiment relevant to the aspects. In the research, SVM achieves a better accuracy and F1-score compared to RNN. However, the authors suggested exploring fastText and LSTM models for further development of the research.

LSTM models have also become the baseline to be compared with other proposed deep learning models for ABSA. An improvement to LSTM model variants is established by Zhu et al. [4] that proposed a hybrid deep learning model which incorporates Bi-LSTM with CNN altogether upon accomplishing the SemEval 2016 task 5 involving an English and Chinese dataset. The hybrid of the two model constructs the AARCNN and the Bi-LSTM model in it showed its strength on understanding texts from two directions instead of only one. Although CNN is primarily known for its performance on image classification tasks, in the research, CNN contributes to sentiment classification enhancement through identifying words that need more attention. Another way of implementing CNN on English ABSA is proposed by Ray and Chakrabarti [12], where they attempted to acquire a more detailed sentiment scale by additionally integrating a rule-based approach to deep learning. Meanwhile, Ilmania et al. [5] trained a CNN model to be compared with Bi-GRU for sentiment classification in the Indonesian ABSA. Although the Bi-GRU model performs better than CNN, the CNN result is still very competitive.

Besides deep learning model options, relevancy along with the amount of the training dataset also determine the performance of ABSA. Researchers have brought up about how deep learning model performances on text classification tasks could be increased using text augmentation. Abulaish & Sah [13] used LDA to group most relevant keywords for each sentiment class. Then, when one of the keywords from a specific class is present in the text's trigram, the trigram will be augmented to the text. The augmented data is then evaluated using CNN as the classifier. Another augmentation method is proposed by generating new sentences through fine-tuning pre-trained models such as GPT-2, BERT, and BART [10]. Contrarily, a simpler augmentation method is proposed by Wei & Zou [6], known as EDA. EDA involves random synonym replacement, random insert, random swap, and random deletion, which are notably simple compared to other augmentation methods. The operations will be executed on random words in each sentence, where afterwards, CNN and LSTM-RNN are used for evaluating the augmented training data.

Liesting et al. [9] then further studied the implementation of EDA on English ABSA dataset using the HAABSA model with some adjustments on the EDA operations related to word context and targeted words in the sentence. The adjusted EDA successfully improved the model's performance for ABSA. However, another research also implemented the original EDA on Vietnamese sentiment analysis with machine learning but did not obtain the best performance from it [8]. This could be an indication that EDA might not be fit for certain language datasets. Apart from that, the backtranslation method succeeds in obtaining competitive results when compared to the EDA method [8] and pre-trained transformer-based augmentation [10]. Meanwhile, Rizos et al. [7] have proposed a substitution-based, word position, and neural generative augmentation to manage imbalanced short-text data. In their substitution-based augmentation, they also considered to replace words with the same context, however by implementing POS tagging and word similarity.

Based on previous research, LSTM, Bi-LSTM and CNN have been considered as baseline models and show promising results on ABSA tasks. However, the performance of those models still could be improved especially in the context of limited resource dataset. Supporting this, many researchers have attempted to enhance the performance of deep learning models using text augmentation. Among the text augmentation methods, EDA is able to increase classification performance on texts with its simplicity. Despite that, EDA still needs to be modified before handling ABSA type of dataset with no target word tag. Augmenting dataset of ABSA itself relies a lot on the contexts of the word and would need to capture important target words as well. Aside from that, the impact of implementing EDA on tasks with a different language should also be further investigated. Additionally, backtranslation could also present quite significant boost for classification results using deep learning models as several research have proven its superiority. Therefore, this research will focus on modifying the existing EDA approach and combine it with the backtranslation method to improve the performance of the baseline models on a small Indonesian ABSA dataset [14].

3- Proposed Method

This work follows the strategy of Ilmania et al. [5], where the ABSA task is divided into two separate tasks which are aspect classification and sentiment classification. In general, the whole process of the proposed method starts from preprocessing the training data before augmentation, text augmentation, and continued with another preprocessing for the augmented texts before inputted to both aspect and sentiment classification model. This process is summarized in Figure 1.

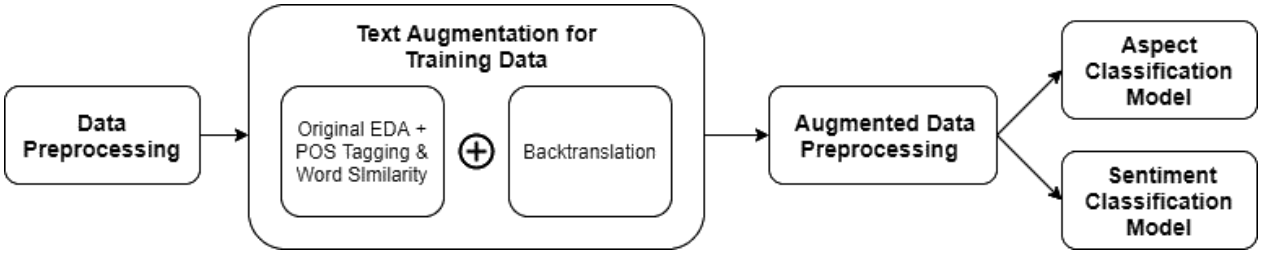

Figure 1. Overview of proposed method

3-1- Data Pre-processing

In this stage, simple preprocessing techniques are applied to the training data. These include eliminating stop words and punctuation marks or special characters [15], as well as case folding which involves changing upper case letters into lower case letters. The list of stop words is mainly taken from the Sastrawi library. Adjustments are made to the stop words list, including the removal of several existing stop words from the library that is considered not relevant for the experiment and also addition of some new stop words.

3-2- EDA with POS Tagging, Word Similarity, and Backtranslation

After pre-processing the dataset, the training data will be augmented. Originally, the EDA method [6] is comprised of four operations. The random synonym replacement is one of EDA's operations which substitutes a random word in

the sentence with its random synonym. Random insert also involves randomly picking a synonym of a random word in the sentence, however, the synonym will be inserted to the original sentence and the point of the insertion will be random as well. Synonyms are obtained using WordNet. Meanwhile, the random swap is a process of switching the position of two random words and random deletion, as the last operation, is a process of removing random words in the sentence. Each operation will be done multiple times depending on the targeted number of words to be replaced/inserted or the number of times the operation should be repeated (specifically for random swap). For random deletion, the word will be removed if the random number generated in the operation is less than the probability p being set. However, in this work, the EDA method will be modified as in Figures 2 and 3, where the yellow actions indicate the original operations, while the green actions indicate modified operations which is also adapted to Indonesian language-related resources.

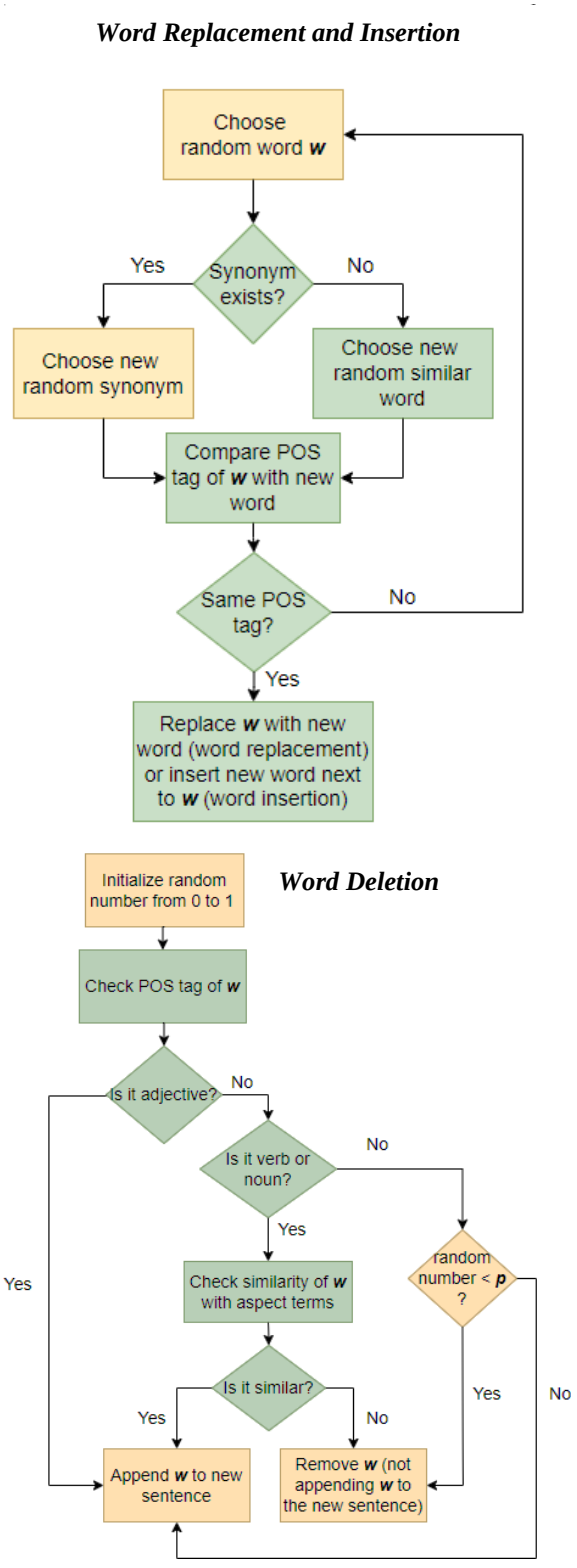

Figure 2. Modified EDA word replacement, insertion, and deletion flowchart

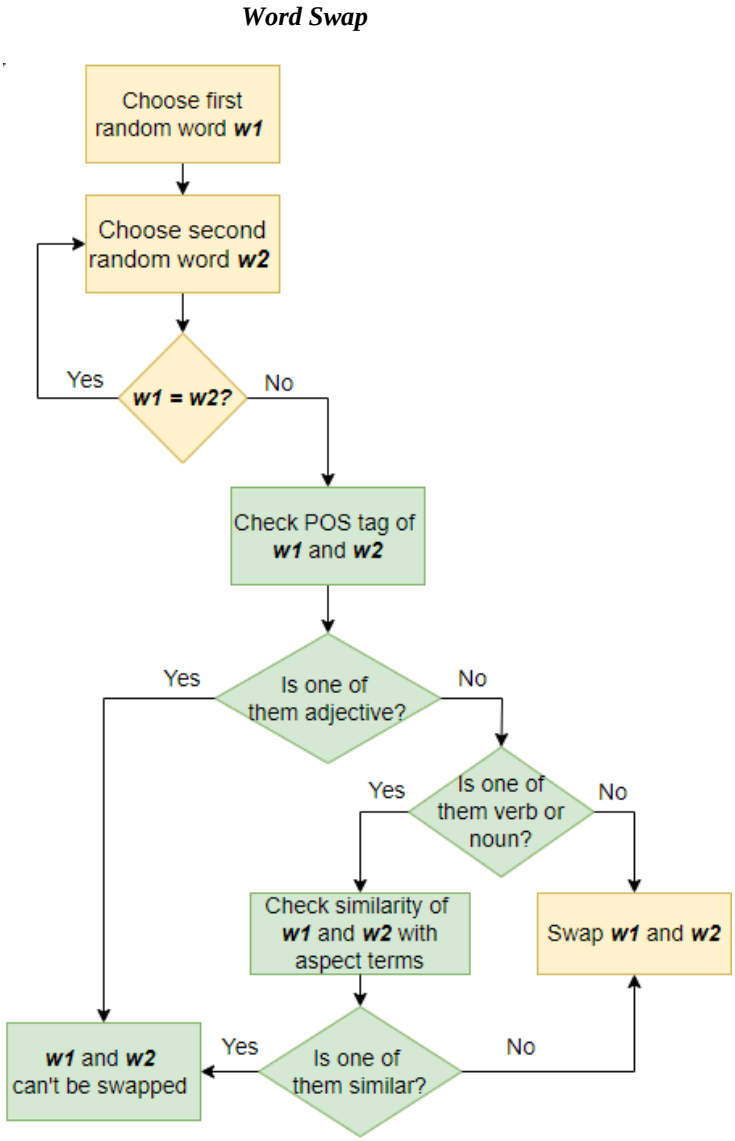

Figure 3. Modified EDA word swap flowchart

The original EDA's [6] random insert and synonym replacement operations could result in addition or replacement of new words that have a totally different context from the original word in a sentence [9]. An example could be seen from number 2 in Table 1, where "*tinggi* (high)*"* which is an adjective, replaced with a noun "*kaliber* (caliber)" that has a different connotation. On another case, a specific word might be a homonym, which means that the same word might have different meanings depending on the context. One of the example is the word "*genting*". It could be an adjective that refers to a critical situation or a noun that describes a house roof. Hence, to help select words with appropriate contexts, POS tagging is implemented [16] on the EDA operations in this work. First of all, the POS tagging will be done on the original sentence using the CRF tagger from the NLTK library. Then, a random word *w* from the original sentence is selected for the operation. When selecting synonym of *w* using the modified random synonym replacement or modified random insert operation, the synonym should have the same POS tag with *w* to make sure the synonym that is used to replace *w* or inserted in the sentence refers to the right situation. The insert position will not be random as in the original EDA, instead, the synonym will be inserted right next to *w*. Arranging the insert position will minimize the noise present in the augmented sentences. In the case of Indonesian dataset, we implemented the Indonesian WordNet to find synonyms, but this does not always guarantee the availability of every word synonyms. Synonyms of some words might not be present in the WordNet, therefore, another alternative is given. When there are no synonyms found for *w*, the word *w* is going to be replaced with other words that are similar to it and should also have the same POS tag as *w*. For instance, synonyms of the word "menu" could not be found in the WordNet. Therefore, with the modification, words similar to "menu", for instance, the word "*hidangan* (dishes)" could be used as replacement or insertion. Similar words are obtained using fastText [17]. The POS tagging and word similarity approach is mainly inspired by Rizos et al. [7]. However, in this work, the role of word similarity for modified random synonym replacement and random insert operation is just additional, not as significant as in Rizos et al. [7]. Instead, both modified operations still mainly rely on WordNet for choosing synonym.

Table 1. Comparison of original and modified EDA implementation

No.	Operations	Sentence
1	Original Sentence (after Preprocessing)	*kualitas makanan sangat baik sebanding harga tinggi* (food quality is very good worth the high price)
2	Original Synonym Replacement	*kualitas makanan sangat baik sebanding harga* ***kaliber*** (food quality is very good worth the price **caliber**)
3	Modified Synonym Replacement	*kualitas makanan sangat baik sebanding harga* ***superior*** (food quality is very good worth the **superior** price)
4	Original Random Insert	*kualitas* ***sembuh*** *makanan sangat baik sebanding harga tinggi* (food **healed** quality is very good worth the high price)
5	Modified Random Insert	*kualitas makanan sangat baik* ***bagus*** *sebanding harga tinggi* (food quality is very good **nice** worth the high price)
6	Original Random Swap	*kualitas makanan sangat* ***tinggi*** *sebanding harga* ***baik*** (food quality is very **high** worth the **good** price)
7	Modified Random Swap	*kualitas makanan sangat baik sebanding harga tinggi* (food quality is very good worth the high price)
8	Original Random Deletion	*kualitas makanan sangat (-* ***baik****) sebanding (-* ***harga****) tinggi* (food quality is very (- **good**) worth the high (- **price**))
9	Modified Random Deletion	*kualitas makanan (-* ***sangat)*** *baik sebanding harga tinggi* (food quality is (- **very**) good worth the high price)

Other than that, our method also extended the use of word similarity from Rizos et al. [7] to identify whether the selected word *w* is associated with a specific aspect or sentiment label. Different from Rizos et al. [7], the purpose of this addition is to determine whether the word *w* is a target word or not. This implementation is considered as a modification to the original EDA as the random characteristics of the original EDA could not assure that the truth of the label is maintained or kept for an ABSA dataset [9]. The new sentence produced by original EDA might lose some information regarding the original label or even alter the meaning of the overall sentence when the terms defining an aspect or sentiment (target words) in the original sentence is switched or even removed. For instance, in Table 1 example number 6, the original EDA swaps the word "*baik* (good)" with "*tinggi* (high)". Therefore, with the original EDA random swap operation, the sentiment for food and price aspects are switched. This example mainly alters the original sentiment of price aspect, opposing the original negative label, which is supposed to be expensive price but becomes good price due to the swap. Furthermore, this example clearly changes the ground truth. In order to tackle the problem, random word selection in modified EDA must fulfill a certain POS tagging and word similarity rules.

The selected random word *w* could not be removed or swapped when it is tagged as adjectives. Adjectives are very representational for sentiment related terms, hence it should not be touched. Thus, "*baik* (good)" and "*tinggi* (high)" should not be removed as in Table 1 example 9, because it is an important adjective that defines the sentiment of the sentence. Instead it only removes the word "*sangat* (very)", which would not significantly alter the original sentiment. This is different from the original random deletion operation, where the word "*baik* (good)" could be removed despite being one of the important sentiment terms. Sentiment terms are also not swapped around the sentence in Table 1 example 7 using modified random swap operation, making no changes to the original sentence. However, later the duplicate augmented sentences will be removed from the training data. Additionally, *w* that has been POS tagged as noun or verb, should not be similar with aspect terms such as "*makanan*", "service", "*tempat*", and "*harga*" which are food, service, place, and price respectively. If the similarity exceeds 0.5, then *w* could not be removed or swapped with another word in the sentence and a new *w* should be selected. This is done because nouns and verbs in the sentence have high probability of being related with the aspect terms and would be crucial if removed or swapped. In Table 1 example 9, the words "*kualitas makanan* (food quality)" and "*harga* (price)" are not removed (removed words are marked with a dash) as they are highly associated with food and price aspect respectively. Contrarily, the aspect terms have a high probability to be removed using the original EDA, as shown in example 8, which will make the price aspect unrecognized.

Additionally, a fifth operation is included to the original EDA which is the backtranslation method [18]. Rather than words, backtranslation generates a new sentence by inputting the whole sentence. Based on preliminary experiments, Chinese language have shown more diversity and alterations of the original sentence in the backtranslated results. Hence, in this work, each sentence will be backtranslated using Chinese language. Since the sentences in the dataset are Indonesian, the backtranslation operation will translate them to Chinese first. Then from the Chinese sentence, it will be translated back to Indonesian. In the data pre-processing stage, stop words in the input sentences are eliminated to ensure that the backtranslated sentences are not exactly the same as the original (input) sentences. Stop word elimination helps to create more variations in terms of the sentence structure and word choices of the backtranslated results. From Table 2, it could be seen that the backtranslated sentence has provided new relevant words producing a different sentence compared to the preprocessed original sentence.

Table 2. Example of backtranslation implementation

No.	Operations	Sentence
1	Original Sentence (after Preprocessing)	*Pelayanannya luar biasa dapat meja pas menikmati view makan siang* (service is incredible got table suitable for enjoying lunch time view)
2	Backtranslation	***Layanan ini*** *luar biasa,* ***meja dapat diadaptasi dengan pemandangan*** *makan siang* (**this service** is incredible, **table can be adapted with** lunch time **scenery**)

3-3- Augmented Data Pre-processing

After augmenting the training data, especially using EDA, the stop words could still remain in the sentences. Therefore, this pre-processing stage is dedicated for eliminating stop words left in the augmented texts. Other than that, duplicate sentences could also be accidentally generated during augmentation, therefore duplicate sentences should also be removed. Following the stop words, duplicate sentences removal, and other necessary cleaning processes, the augmented sentences are tokenized and sentences are broken into a collection of words, which is part of another common pre-processing technique [19]. Subsequently, the words are converted into numerical representations or word indexes. Equal length for all sentences should also be maintained in the training data, therefore, post-padding is added in each sentence. The padded sentences will be the final input for the multi-label classification models. However, there is a slight difference in the structure of the final input for sentiment classification model, which will be further discussed in section 3.5.

3-4- Aspect Multi-Label Classification Model

As one customer review may consist of either one or more aspects, a multi-label classification approach is taken to predict the aspects related to the sentence. The first model to be compared for this aspect classification task would be LSTM. The LSTM model is first introduced in Hochreiter and Schmidhuber [20]. In our work, the LSTM model will receive the previously pre-processed augmented sentences as the input. Each word in the input will be further represented with a 100-dimensional word embedding obtained using the same fastText [17] previously implemented for word similarity. The implementation of fastText word embedding itself helps with the representing out-of-vocabulary words [21]. The word vectors from the embedding would then be processed into the first LSTM layer with 64 units. The output from first LSTM layer will then be passed to the second layer with the same number of units as the first LSTM layer. Dense layer consisting of four neurons will receive input from the second LSTM layer and outputs values ranging from 0 to 1 for each aspect's prediction. If an aspect prediction is closer to 1, the aspect is strongly present in the sentence. In this context, the model will predict four aspects in a sentence involving food, service, price, and place. The LSTM architecture and its model summary is presented in Figure 4 and Table 3 respectively.

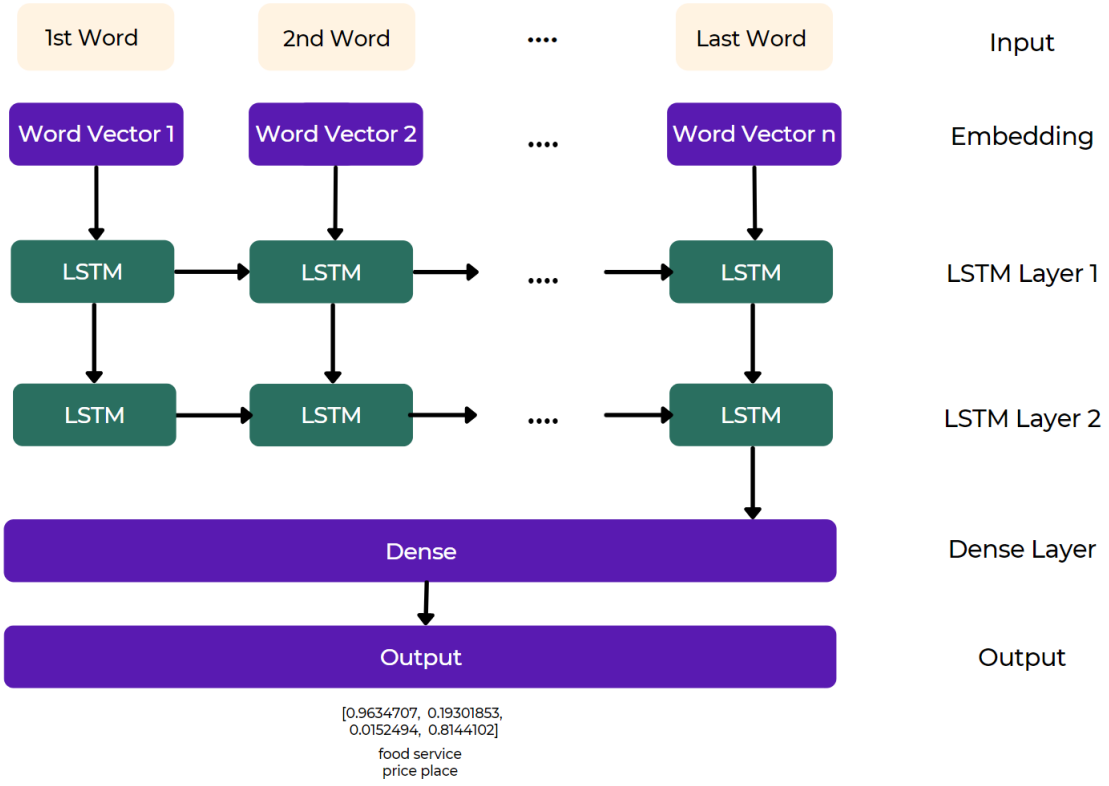

Figure 4. LSTM architecture for aspect classification

Table 3. Summary of the LSTM model

Layers	Output Shape
Embedding	(None, None, 100)
LSTM	(None, None, 64)
LSTM	(None, 64)
Dense	(None, 4)

Other than the LSTM model, we also perform the multi-label classification task using a Bi-LSTM model [22]. Essentially, the conversion process of words into vector representation in the embedding layer for Bi-LSTM model is the same as LSTM. However, instead of feeding word vectors into one forward LSTM layer, a Bi-LSTM layer feeds word vectors on two different directions of LSTM layers. Figure 5 shows that in the Bi-LSTM layer, word vectors are not only being fed to the forward LSTM, but also to a backward LSTM. As mentioned in Table 4, the Bi-LSTM layer consists of 128 units. Therefore, there are 64 units in the forward LSTM and another 64 units in the backward LSTM. Output of both forward and backward LSTM will then be the input for dense layer and produce the same form of prediction result as the LSTM model.

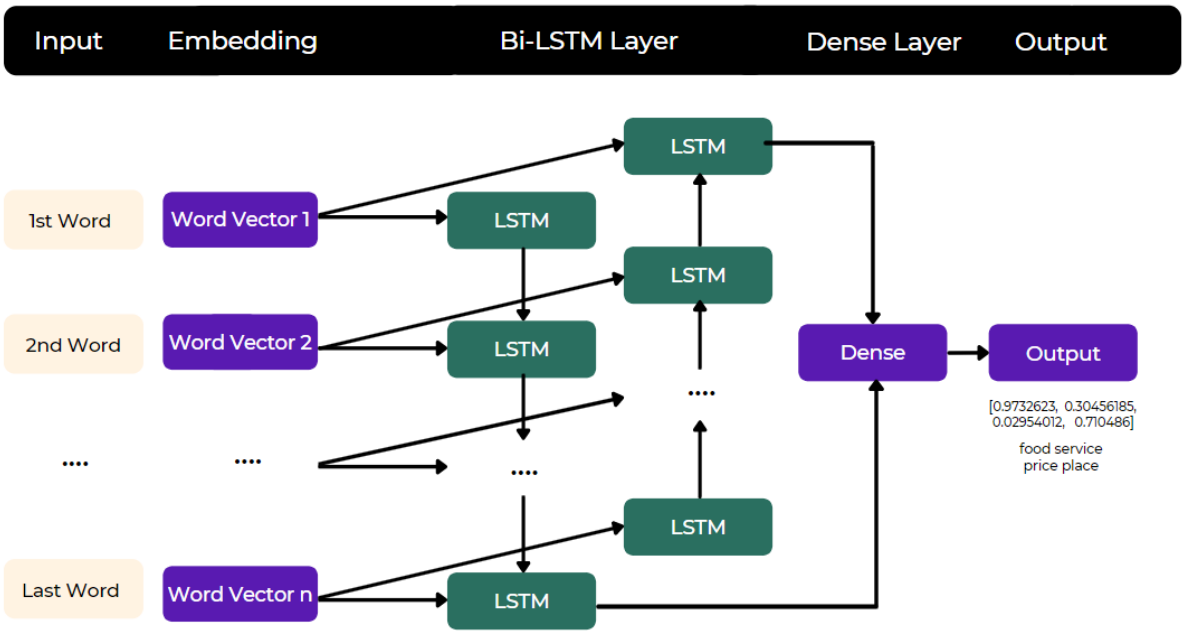

Figure 5. Bi-LSTM architecture for aspect classification

Table 4. Summary of the Bi-LSTM model

Layers	Output Shape
Embedding	(None, None, 100)
Bi-LSTM	(None, 128)
Dense	(None, 4)

3-5- Sentiment Classification Model

The input for sentiment classification model is quite similar to the multi-label classification model. However, for sentiment classification, the aspect label is added at the beginning of the sentence, similar to work by Ilmania et al [5]. Sentence with more than one different aspect label is replicated and the respective aspect label is inserted into the sentence. Thus, there would be the same sentences with different aspect labels at the beginning of each sentence. If Sentence A has two aspects, for example food and price, there would be two new sentences. The sentences are "*makanan* (food)" + Sentence A and "*harga* (price)" + Sentence A. This creates an indication for the sentiment to be evaluated based on the different aspects. The sentiment label prediction of the sentence based on the aspect label that has been inserted in the sentences is either positive or negative. After utilizing the same word embedding used in the multi-label classification, the input that is comprised of word vectors will enter the one-dimensional CNN model. The CNN model in this work is mainly referenced from previous research by Ilmania et al. [5], Kim [23], as well as Zhang and Wallace [24]. The sentiment classification model architecture consists of 1D convolution layer, global max pooling layer, and a dense layer. We also vary the filter size into 3, 4, and 5, which could be seen from Figure 6.

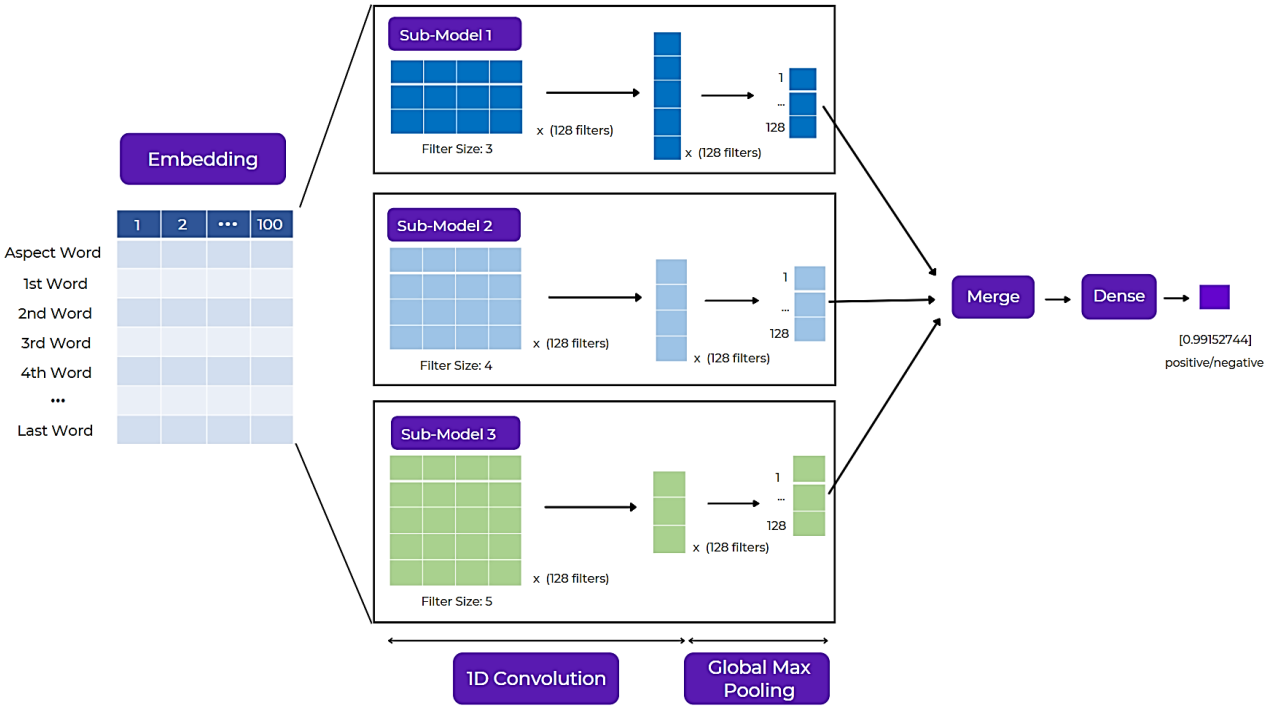

Figure 6. CNN architecture for sentiment classification

By implementing different sizes of filters for the 1D convolution layer, the model is divided into specific sub-models. In each sub-model, the text input undergoes the convolution process using 128 filters with a specific size. From Table 5, it could be seen that the convolution layer in each sub-model outputs to feature maps with different output shapes according to the respective filter size. Then for the next step, only the maximum value from each filter is taken in the global max pooling layer. The output of the global max pooling layer from each sub-models are then merged and made into an input for the dense layer of the final model as described in Table 6. Eventually, the model will output a single value that ranges from 0 to 1. If the value is closer to 1, the sentiment label is positive. Meanwhile, when the output value is closer to 0, the sentiment label is negative.

Table 5. Summary of the CNN sub-model

Layers	Output Shape
Embedding	(None, max_words, 100)
Conv1D	**Filter size 3** = (None, max_words - 2, 128) **Filter size 4** = (None, max_words - 3, 128) **Filter size 5** = (None, max_words - 4, 128)
Global Max Pooling	(None, 128)

Table 6. Summary of the CNN final model

Layers	Output Shape
Merge	(None, 128)
Dense	(None, 1)

4- Experiments

4-1- Dataset

The dataset used in this work consists of Trip Advisor restaurant reviews in Indonesian language from Ekawati & Khodra's [14] research, which is originally collected by Gojali & Khodra [25]. In Ekawati and Khodra's dataset, there is an addition to the original data for low-frequency class labels and the validity of the labelling is manually checked after going through semi-supervised learning. The dataset contains four aspect labels, which are food, service, price, and place. Each aspect label would have three sentiment labels, which are positive, negative, and neutral. However, in this work, neutral labels are ignored and the dataset is manually rechecked for the second time. There are a total of 1793 sentences used from the dataset.

From the distribution of the class labels shown in Table 7, we could observe that there is still a very low amount of data for the negative sentiment class. Additionally, the service and price aspect are represented with fewer data compared to other aspects. However, in general, the dataset is still very limited, which further supports the focus of this work. For the research purpose, we use 60% of the overall dataset for training and 20% for each validation and testing.

Table 7. Class distribution for training, validation, and testing

Aspect Class	Training		Validation		Testing	
	Positive	Negative	Positive	Negative	Positive	Negative
Food	376	73	126	24	126	23
Service	150	55	50	18	50	20
Price	153	103	52	33	52	34
Place	383	71	123	25	125	22
Total	1062	302	351	100	353	99

4-2- Experimental Design

Primarily, there are six training data settings explored for experiment. The first setting is the original training data without augmentation process. Next is the augmented training data using the original EDA method. Then the experiment would proceed using the proposed modified EDA training data and backtranslation separately. Following that, each EDA augmented data will be combined with backtranslation. Hence, for the next data setting, the original EDA augmented training data will be combined with the backtranslated data. Lastly, the training data will be augmented using a combination of the modified EDA with backtranslation method. For both original and modified EDA augmentation, we focused on adding new sentences for the classes that has fewer data or minority classes. Each sentence in the training data that belongs to the majority class is augmented into a maximum of only two new sentences. Meanwhile, each minority class sentence will be augmented up to eight new sentences. Separately, the backtranslation process will contribute to an additional sentence for every aspect class. We trained all models for a maximum of 50 epochs with batch size equals to 50. Early stopping with patience of 5 epochs is implemented for all models to prevent overfitting. The LSTM layers have 0.2 dropout, whereas in the CNN model, different dropout value is implemented, 0.1 in the sub-models and 0.5 in the final model. All of the models are trained according to the six data settings mentioned using Adam optimizer. Each epoch's losses are then evaluated using the binary cross-entropy function. Afterward, we proceed to the validation and testing process.

Performance of the models on testing data in each sentiment class is then measured using accuracy and F1-score metrics. Accuracy is a metric that shows how much a model could predict the correct labels, while F1-score is calculated based on the precision and recall values. For aspect classification model performance, we also use the macro and micro-averaged calculation for both metrics. Macro and micro-averaged metrics are commonly applied for multi-label classification tasks. When basic accuracy metrics of each class labels are averaged, it will output the macro-averaged accuracy. Differently, micro-averaged accuracy calculates accuracy metric using the total number of true positives, false positives, true negatives, and false negatives from all class labels. Meanwhile, similar to macro-averaged accuracy, macro-averaged F1-score is the average of each label's F1 value. Whereas micro-averaged F1-score involves the micro values of precision and recall. For sentiment classification model performance specifically, we focus on its overall accuracy along with both macro-average and overall F1-score to measure the classification of sentiment based on all aspects in general.

5- Results and Analysis

5-1- Performance of Aspect Classification Model

Based on the experiment results, by training the LSTM aspect classification model with augmented data, the model is mostly able to perform better on the testing data compared to no augmentation in terms of both accuracy and F1-score. This is presented in Table 8, where the accuracy and F1-score values are organized in columns, while macro and micro-average of both metrics are placed in the last row of the table. The performances are then compared over different augmentation methods. As we observe deeper on the augmented method results, the original EDA still achieves slightly higher macro-average accuracy and F1-score compared to the modified EDA. This could be due to more restrictions and conditions to be fulfilled before doing random operations in modified EDA. Although the restrictions help to generate more relevant words and keep the truth of the label, it could also limit the variety of words generated. Hence, the model might not learn as much new words. Despite that, the modified EDA still performs better in classifying the service aspect and same result for price aspect classification. This shows that the modified EDA could still maintain its performance for price aspect and able to generate more appropriate training data for service aspect. Hence, still providing competitive results.

Table 8. Accuracy and F1-score of LSTM aspect classification model on testing data

Aspect Class	Without Augmentation		Original EDA		Modified EDA		Backtranslation		Original EDA with Backtranslation		Modified EDA with Backtranslation	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Food	90.8	89.2	92.5	90.8	91.4	89.8	92.5	91.3	**93.0**	**92.0**	92.5	91.0
Service	91.9	76.0	94.7	84.8	95.0	85.7	95.0	85.2	95.0	85.9	**95.8**	**88.0**
Price	97.2	94.1	98.9	97.6	98.9	97.6	98.3	96.4	98.9	97.6	**99.2**	**98.2**
Place	91.4	89.7	92.8	91.2	92.5	91.0	**93.6**	**92.3**	92.5	90.7	92.8	91.4
Macro-Average	92.8	87.3	94.7	91.1	94.4	91.0	94.8	91.3	94.8	91.6	**95.1**	**92.2**
Micro-Average	92.8	88.5	94.7	91.4	94.4	91.1	94.8	91.8	94.8	91.8	**95.1**	**92.1**

Apart from the original and modified EDA, we also experimented on augmenting the training data using only backtranslation method. The testing result shows that the backtranslation performs better than the EDA methods by 0.2% to 0.7% in terms of the macro and micro-averaged F1-score. Primarily, backtranslation heavily relies on a larger resource compared to EDA that mainly utilizes WordNet for additional word vocabularies. Therefore, the backtranslation method could excel in predicting three out of the four aspects compared to the EDA methods independently, excluding the price aspect. Moreover, by using backtranslation, it could achieve the highest performance for classifying place aspect. However, in contrast to the EDA methods, backtranslation could not provide a better data for price aspect, which leads to a decrease in performance.

Following that, we combined each of the EDA methods with backtranslation. The combination of original EDA and backtranslation help increase performance in food and service aspect classification compared to using each method independently. Meanwhile, for price aspect classification, original EDA and backtranslation hybrid only achieves higher performance than the backtranslation method itself and equal to using the original EDA alone. Therefore, although the original EDA and backtranslation hybrid method does not give improvement to original EDA method, it could cover the weakness of the backtranslation method. However, when original EDA and backtranslation is implemented together and compared to the individual methods, it achieves lower performance for place aspect classification. This could happen due to the nature of original EDA, where the method itself could not guarantee in retaining the sentence label as much as the modified EDA. Similar to the original EDA, the generation of new sentence using backtranslation is also quite difficult to control, hence, it could alter the original label of the sentence as well, specifically for place aspect in this case. In addition to that, removing stop word before doing backtranslation is needed to produce a varied/different sentence, but this process actually also take part in less accuracy of the newly generated sentence. Hence, further affecting the model's understanding of the whole augmented training data.

However, overall, the combination of modified EDA and backtranslation, successfully gain performance increase compared to using modified EDA only, at minimum. For half of the aspects classification performance, modified EDA and backtranslation hybrid method could perform better than both methods individually. By combining backtranslation with modified EDA, it could be seen that mostly the backtranslation method is able to counterbalance the modified EDA's weakness and vice versa. In contrary to hybrid of original EDA and backtranslation, adding backtranslation to modified EDA does not result in any lower performance compared to the individual methods. The difference between original and modified EDA could be the reasoning behind this. As the modified EDA has more ability to retain the labels of the original sentence, it would be more suitable to be paired with backtranslation. Hence, as a result, the modified EDA and backtranslation combination or proposed method, has achieved the best performance compared to the rest of training data settings for LSTM model. When compared to the original EDA itself, the proposed method achieves 0.4% and 1.1% increase of macro-averaged accuracy and F1-score respectively.

Besides the LSTM model, we also compared the same performance metrics of aspect classification on testing data using the Bi-LSTM model, which is summarized in Table 9. From Table 8 and Table 9 we could observe that the Bi-LSTM model performance is also similar to the LSTM model. In general, the implementation of augmentation on training data also mainly supports the Bi-LSTM model to predict the aspects better. This is also supported in Figure 7, where LSTM and Bi-LSTM model both show lower training and validation losses when using augmented data compared to non-augmented data. The performance using original EDA and modified EDA on Bi-LSTM is also quite competitive in terms of accuracy, which is also similar when implemented on the LSTM model. However, in this case, the original EDA obtains quite an increase in terms of its macro and micro-averaged F1-score, because it achieves the best performance for place aspect prediction. Meanwhile, the macro and micro-averaged F1-score of modified EDA does not show any improvement compared to using the LSTM model. Yet, essentially, the Bi-LSTM model performs slightly better compared to the LSTM model. The Bi-LSTM model presents higher performance for the majority of the training data settings. This continually proves the strength of Bi-LSTM compared to LSTM [4].

Table 9. Accuracy and F1-score of Bi-LSTM aspect classification model on testing data

Aspect Class	Without Augmentation		Original EDA		Modified EDA		Backtranslation		Original EDA with Backtranslation		Modified EDA with Backtranslation	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Food	89.7	87.7	91.9	90.4	92.5	91.1	91.1	89.3	92.2	90.7	**92.8**	**91.2**
Service	95.0	86.2	94.7	85.9	95.0	86.2	**96.4**	**90.4**	95.5	87.3	95.8	88.2
Price	96.9	93.5	98.6	97.0	98.3	96.4	**98.9**	97.6	**98.9**	**97.7**	98.6	97.1
Place	90.3	88.0	**94.2**	**92.8**	92.2	90.1	91.4	89.5	93.3	91.7	93.3	91.9
Macro-Average	93.0	88.8	94.8	91.5	94.5	91.0	94.4	91.7	95.0	91.8	**95.1**	**92.1**
Micro-Average	93.0	88.7	94.8	91.7	94.5	91.1	94.4	91.1	95.0	91.9	**95.1**	**92.1**

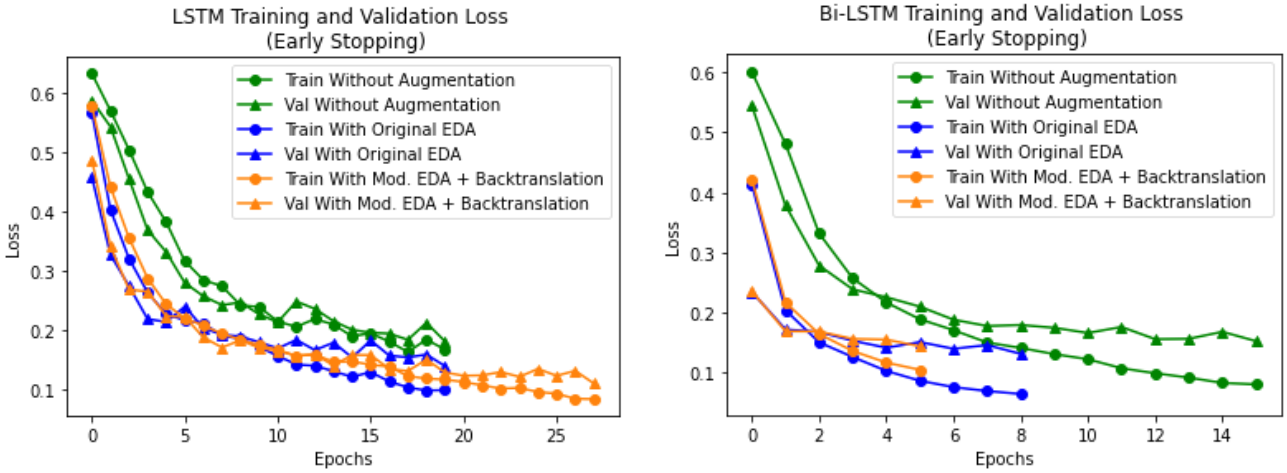

Figure 7. Training and validation loss comparison of LSTM and Bi-LSTM model

Despite performing higher than LSTM in most data settings, the implementation of backtranslation on Bi-LSTM is not as outstanding. The backtranslated data only helps improves the macro-averaged F1-score of the aspect classification, but not in terms of accuracy and micro-averaged F1-score. Regardless of that, the performance using the hybrid augmentation methods on Bi-LSTM are not critically affected by this factor due to the powerful performance achieved by original or modified EDA individually. It could be seen that although the modified EDA is combined with backtranslation and suffers a decrease compared to its performance in the LSTM model, it is not very significant and considered very competitive. Meanwhile, the hybrid of original EDA and backtranslation even managed to achieve slightly higher performance. However, although it seems that the EDA methods are more high-performing on Bi-LSTM in general, the backtranslation method actually helps improve classification performance on service and price aspects. Therefore, the methods could complement each other when used together. This leads to an increase of original EDA macro-averaged F1-score by 0.3% and backtranslation by 0.1% when combining original EDA and backtranslation. While the modified EDA and backtranslation hybrid is able to improve the modified EDA performance by 1.1% and backtranslation by 0.4%. Hence, the effects of combining two augmentation method on Bi-LSTM model is also similar to when using the LSTM model.

Essentially, implementing two augmentation methods simultaneously could increase the performance of individual methods. Integrating two methods actually helps the model to learn more patterns of the specific aspect. Therefore, allowing improvement on classifying certain aspects. However, a similar problem occurred when pairing original EDA and backtranslation as in the LSTM model, where its performance improvement is not as significant compared to combining modified EDA and backtranslation. As have been mentioned previously, this is possibly due to the nature of original EDA and backtranslation method. Thus, the proposed method involving the combination of modified EDA and backtranslation is more ideal for Bi-LSTM model, achieving the best performance with 0.3% and 0.6% increase from the original EDA alone, in terms of its macro-averaged accuracy and F1 respectively. However, since implementation of the proposed method on Bi-LSTM model suffers a narrow loss from its implementation on LSTM, the performance of LSTM model trained on our proposed method achieves the best score for the aspect classification task.

5-2- Performance of Sentiment Classification Model

The CNN model experiences several stagnant accuracy on predicting sentiment based on a specific aspect when trained using the augmented data. However, in Table 10, it is shown that the proposed method obtains the same accuracy as using no augmentation data on service aspect-sentiment prediction. The EDA methods individually along with combination of original EDA and backtranslation method also do not improve classification accuracy for service aspect-sentiment, instead presenting lower accuracy compared to training the model with non-augmented data. Besides that, the original EDA method separately also does not increase the food aspect-sentiment classification accuracy. Similar case experienced by testing the CNN model trained using backtranslated data and the proposed method on place aspect-sentiment. This denotes that the corresponding augmentation methods is not able to generate new relevant keywords for the respective aspect-sentiment and misguide the model's prediction, thus achieving a lower to equal performance accuracy compared with the model trained without augmentation. However, the accuracy of augmented methods mostly achieved higher accuracy on other aspect-sentiment classes.

Table 10. Accuracy of CNN sentiment classification model on testing data

Sentiment Based On	Accuracy					
	Without Augmentation	Original EDA	Modified EDA	Backtranslation	Original EDA with Backtranslation	Modified EDA with Backtranslation
Food	87.9	86.6	88.6	**90.6**	**90.6**	**90.6**
Service	91.4	88.6	88.6	**94.3**	88.6	91.4
Price	76.7	81.4	83.7	81.4	82.6	**86.0**
Place	88.4	**89.1**	**89.1**	87.8	88.4	87.8
Overall	86.5	86.7	87.8	88.5	88.1	**88.9**

Different from aspect classification task, the modified EDA method individually present a higher overall accuracy compared to the original EDA in accomplishing the sentiment classification task. Although modified EDA alone suffer in accuracy of service aspect-based sentiment prediction, the method excels in predicting the rest of existing aspects, hence outperforming both non-augmented and original EDA method. Compared to the original EDA method, the modified EDA alone improves accuracy by 2% for food and 2.3% for price aspect-based sentiment. Hence, implementing modified EDA for sentiment classification task using CNN does not only improve the performance of prediction for large classes, but also minority class. This clearly shows that the modifications made in EDA has successfully generated more relevant vocabularies related to sentiment for the respective aspects. Hence, it enables the model to detect sentiment based on their aspects better. At the same time, generally the backtranslation method also outperforms original EDA and non-augmented method, even also modified EDA. However, considering the backtranslation performance accuracy specifically on each aspect, it still poses some shortcomings. For instance, it does not improve the price aspect-sentiment prediction from the original EDA and even experienced a decrease in accuracy for place aspect-sentiment prediction compared to no augmentation and both EDA methods.

Regardless of that, the outstanding performance of backtranslation is mostly gained from food and service aspect-sentiment prediction, where it achieves the best accuracy for those aspects out of the compared methods. As the modified EDA and backtranslation each show the strength and weakness in specific aspect-based sentiment prediction, they are able to complement each other. Hence, the hybrid of modified EDA and backtranslation could successfully improve three out of four aspect-based sentiment prediction accuracy compared to the original EDA. Precisely, 4% on food aspect, 2.8% on service aspect, and 4.6% on price aspect. Overall, the proposed method managed to bring 2.2% improvement on aspect-based sentiment prediction accuracy from the original EDA method. Other than the proposed method, the addition of backtranslated data to original EDA also increases accuracy compared to using original EDA independently, however the result is still below the proposed method and even the backtranslation method alone.

For predictions on the positive sentiment class using augmentation method, in some cases, their F1-score also degrade compared to the model trained without augmentation. For example, in Table 11, the original EDA performed lower on predicting food-positive and service-positive label. In case of service-positive label prediction, modified EDA and the hybrid augmentation methods also follow. Another performance decrease is also experienced for place-positive label prediction using backtranslation and the hybrid augmentation methods. Besides the positive sentiment class, prediction of service-negative label using the EDA methods and combination of original EDA with backtranslation could not be on par with the performance by non-augmented method. However, those occurrences are not major. In general, the CNN model trained with augmented data has achieved a higher overall F1-score for sentiment classification based on aspects.

Table 11. F1-score of CNN sentiment classification model on testing data

Aspect-Sentiment Class	F1-Score					
	Without Augmentation	Original EDA	Modified EDA	Backtranslation	Original EDA with Backtranslation	Modified EDA with Backtranslation
Food-Positive	93.0	91.8	93.2	**94.4**	**94.4**	**94.4**
Food-Negative	57.1	63.0	63.8	70.8	69.6	**72.0**
Macro-Average (Food)	75.1	77.4	78.5	82.6	82.0	**83.2**
Service-Positive	94.1	92.2	92.2	**95.9**	92.3	94.0
Service-Negative	84.2	78.9	78.9	**90.5**	77.8	85.0
Macro-Average (Service)	89.2	85.6	85.6	**93.2**	85.0	89.5
Price-Positive	81.8	84.6	87.7	84.6	86.7	**88.5**
Price-Negative	67.7	76.5	75.9	76.5	74.6	**82.4**
Macro-Average (Price)	74.8	80.5	81.8	80.5	80.7	**85.4**
Place-Positive	93.4	93.5	**93.7**	92.9	93.3	92.7
Place-Negative	51.4	**65.2**	61.9	57.1	58.5	62.5
Macro-Average (Place)	72.4	**79.4**	77.8	75.0	75.9	77.6
Overall	78.6	81.1	81.3	83.3	81.4	**84.3**

As the sentiment classification model's performance in terms of F1-score is compared between the augmentation methods, the modified EDA is able to improve prediction of positive sentiment labels of food, price, and place aspect, as well as food-negative sentiment label, better than the original EDA. One of the example could be taken from the sentence "*harga* (price aspect label) *harga standar restoran sejenis*" which means "price (aspect label) price is standard to similar restaurants". The prediction should be a positive sentiment for the price aspect. The CNN model that has been trained with modified EDA successfully predicts the sentiment label with the value of 0.881, closer to the positive sentiment label. Meanwhile, the original EDA based CNN model outputs the value of 0.498, which leans more towards the negative sentiment label. By training the CNN model using modified EDA data, the model could grasp the precise meaning related to the sentiment in a sentence by considering the context. For instance, in the example, the word "standard" could actually refer to both positive and negative sentiment, depends on what aspect it is based on. In terms of price, "standard" would mean that the price is not expensive or just normal, which is positive. However, in terms of food, "standard" could refer to no specialty in the food being served, which is closer to a negative sentiment. With the characteristics of modified EDA that is equipped with more restrictions to produce more context relevant sentences, the model is able to learn and differentiate the two contexts, compared to the original EDA. Aside from that, the backtranslation method excels in food and service aspect prediction, achieving an average of 2.1% overall increase from the EDA methods.

Furthermore, when backtranslation is added to the modified EDA sentences, the proposed method performs better with higher macro-averaged F1-score on majority of aspects compared to the original EDA. In Figure 8, the proposed method also achieves slightly lower validation loss compared to the original EDA in its last epoch. Combining modified EDA and backtranslation altogether could predict sentiments more accurately when the sentence has more than one sentiment labels. For example, the sentence "*kalau makan restonya burangrang makanan bervariasi* western asian indonesia *tampilan menarik namun harga lebih mahal daripada foodcourtnya*" meaning "if (you) eat in burangrang restaurant the food is variative western asian indonesia presentation is interesting however price is more expensive than the foodcourt". The sentiments on this sentence are supposed to be evaluated based on two aspects, namely food and price. The sentiment is positive for food aspect and negative for price aspect. When tested using the original EDA only, the model could only predict the food aspect-sentiment correctly, but results in false prediction for price aspect-sentiment. Meanwhile, the proposed method could successfully predict both aspect-sentiments correctly. During testing, the proposed method is able to gain improvement on food and service-based sentiment classification with the superiority that backtranslation has on those respective aspects compared to the EDA methods. Although the prediction of place-sentiment label using backtranslation is not as significant, the modified EDA also individually gives support and boosts the F1-score on price and place-based sentiment classification, enabling the proposed method to achieve the best performance on both food and price-sentiment prediction over other evaluated methods. Hence, the proposed method also gains the best F1-score performance for sentiment classification model.

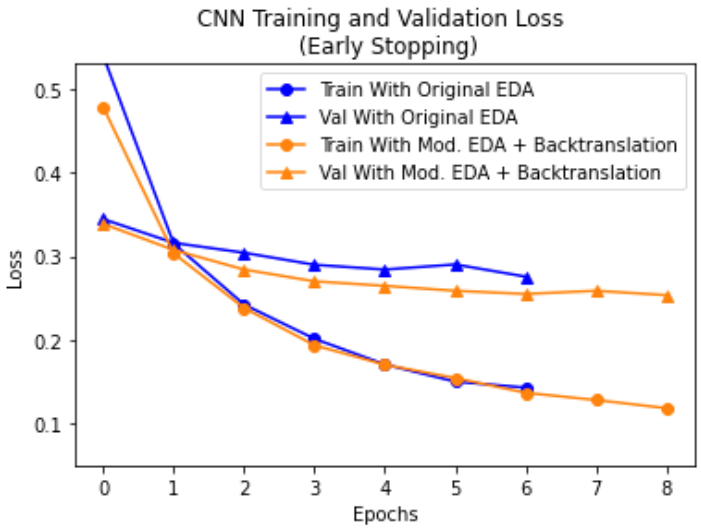

Figure 8. Loss comparison between original EDA and proposed method

Above all, the combination of the modified EDA with backtranslation has improved the macro-averaged F1-score of the original EDA by 5.8% on food aspect-sentiment, 3.9% on service aspect-sentiment, and 4.9% on price aspect-sentiment. Hence, the overall increase in F1-score that the hybrid method achieved compared to the original EDA reached 3.2%. Once again, the two methods complement each other and do not only improve the performance of original EDA but also modified EDA and backtranslation separately. However, this was not the case when combining original EDA with backtranslation, where it could only improve the performance of original EDA alone and was still inferior to the backtranslation method itself. Thus, combining augmentation methods does not always improve the methods separately, specifically in classifying aspect-based sentiment labels.

6- Conclusion

In this research, we have compared aspect and sentiment classification models using different training data settings with various augmentation methods. The augmentation methods include the original EDA, modified EDA, backtranslation, and combinations of each of the EDA methods with backtranslation. These methods are evaluated on Ekawati & Khodra's dataset [14] using baseline deep learning models. Models with augmented training data have mostly shown improvements in terms of their accuracy and F1-score compared to those not using any augmentation. For an aspect classification task, adding backtranslation to either the original or modified EDA could help improve the performance of each individual method separately. Both hybrid-augmented methods are also able to outperform the original EDA. This also applies for the sentiment classification task, except that the combination of the original EDA and backtranslation could only improve the performance of the original EDA alone but not for the backtranslation separately. Additionally, for both classification tasks, backtranslation is best paired with the modified EDA considering each method's characteristics compatibility. Ultimately, the proposed method achieves the best performance for aspect classification using both the LSTM and Bi-LSTM models. Out of the aspect classification models, the best performance is obtained using LSTM. Following that, implementing the proposed method trained using CNN for sentiment classification based on different aspects also gained the best overall accuracy and F1-score in general.

Although the combination of modified EDA with backtranslation method has contributed to the improvement of text classification accuracy and F1-score performance for the models evaluated in this work, the EDA method itself only presented a word-level augmentation. Other than that, modifications to EDA and the backtranslation method require quite plenty of resources. Hence, improvements could be made to propose a sentence-level augmentation on the Indonesian dataset with a more effective, efficient, and reliable augmentation method.

7- Declarations

7-1- Author Contributions

Conceptualization, A.S.G.; methodology, N.; formal analysis, A.S.G.; writing—original draft preparation, N.; writing—review and editing, A.S.G. All authors have read and agreed to the published version of the manuscript.

7-2- Data Availability Statement

The data presented in this study are available on request from the corresponding author.

7-3- Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

7-4- Institutional Review Board Statement

Not applicable.

7-5- Informed Consent Statement

Not applicable.

7-6- Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

8- References

[1] Zhao, Y., Xu, X., & Wang, M. (2019). Predicting overall customer satisfaction: Big data evidence from hotel online textual reviews. International Journal of Hospitality Management, 76, 111–121. doi:10.1016/j.ijhm.2018.03.017.

[2] Tao, J., & Fang, X. (2020). Toward multi-label sentiment analysis: a transfer learning-based approach. Journal of Big Data, 7(1), 1-26. doi:10.1186/s40537-019-0278-0.

[3] Zulqarnain, M., Ghazali, R., Hassim, Y. M. M., & Rehan, M. (2020). A comparative review on deep learning models for text classification. Indonesian Journal of Electrical Engineering and Computer Science, 19(1), 325–335. doi:10.11591/ijeecs.v19.i1.pp325-335.

[4] Zhu, Y., Gao, X., Zhang, W., Liu, S., & Zhang, Y. (2018). A bi-directional LSTM-CNN model with attention for Aspect-level text classification. Future Internet, 10(12), 1–11. doi:10.3390/fi10120116.

[5] Ilmania, A., Abdurrahman, Cahyawijaya, S., & Purwarianti, A. (2018). Aspect Detection and Sentiment Classification Using Deep Neural Network for Indonesian Aspect-Based Sentiment Analysis. 2018 International Conference on Asian Language Processing (IALP), Bandung, Indonesia. doi:10.1109/ialp.2018.8629181.

[6] Wei, J., & Zou, K. (2019). EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). doi:10.18653/v1/d19-1670.

[7] Rizos, G., Hemker, K., & Schuller, B. (2019). Augment to Prevent. Proceedings of the 28th ACM International Conference on Information and Knowledge Management, New York, United States. doi:10.1145/3357384.3358040.

[8] Duong, H. T., & Nguyen-Thi, T. A. (2021). A review: preprocessing techniques and data augmentation for sentiment analysis. Computational Social Networks, 8(1), 1–16. doi:10.1186/s40649-020-00080-x.

[9] Liesting, T., Frasincar, F., & Truşcă, M. M. (2021). Data augmentation in a hybrid approach for aspect-based sentiment analysis. Proceedings of the 36th Annual ACM Symposium on Applied Computing, New York, United States, 828–835. doi:10.1145/3412841.3441958.

[10] Kumar, V., Choudhary, A., & Cho, E. (2020). Data augmentation using pre-trained transformer models. arXiv preprint arXiv:2003.02245. doi:10.48550/arXiv.2003.02245.

[11] Al-Smadi, M., Qawasmeh, O., Al-Ayyoub, M., Jararweh, Y., & Gupta, B. (2018). Deep Recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels' reviews. Journal of Computational Science, 27, 386–393. doi:10.1016/j.jocs.2017.11.006.

[12] Ray, P., & Chakrabarti, A. (2022). A Mixed approach of Deep Learning method and Rule-Based method to improve Aspect Level Sentiment Analysis. Applied Computing and Informatics, 18(1–2), 163–178. doi:10.1016/j.aci.2019.02.002.

[13] Abulaish, M., & Sah, A. K. (2019). A Text Data Augmentation Approach for Improving the Performance of CNN. 2019 11th International Conference on Communication Systems Networks (COMSNETS). doi:10.1109/comsnets.2019.8711054.

[14] Ekawati, D., & Khodra, M. L. (2017). Aspect-based sentiment analysis for Indonesian restaurant reviews. 2017 International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA). doi:10.1109/icaicta.2017.8090963.

[15] Sudheer, K., & Valarmathi, B. (2018). Real time sentiment analysis of e-commerce websites using machine learning algorithms. International Journal of Mechanical Engineering and Technology, 9(2), 180–193.

[16] Abka, A. F. (2016). Evaluating the use of word embeddings for part-of-speech tagging in Bahasa Indonesia. 2016 International Conference on Computer, Control, Informatics and Its Applications (IC3INA). doi:10.1109/ic3ina.2016.7863051.

[17] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics, 5, 135–146. doi:10.1162/tacl_a_00051.

[18] Sennrich, R., Haddow, B., & Birch, A. (2016). Improving Neural Machine Translation Models with Monolingual Data. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 86–96. doi:10.18653/v1/p16-1009.

[19] Bayhaqy, A., Sfenrianto, S., Nainggolan, K., & Kaburuan, E. R. (2018). Sentiment Analysis about E-Commerce from Tweets Using Decision Tree, K-Nearest Neighbor, and Naïve Bayes. 2018 International Conference on Orange Technologies (ICOT). doi:10.1109/icot.2018.8705796.

[20] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780. doi:10.1162/neco.1997.9.8.1735.

[21] Wang, B., Wang, A., Chen, F., Wang, Y., & Kuo, C. C. J. (2019). Evaluating word embedding models: Methods and experimental results. APSIPA Transactions on Signal and Information Processing, 8, 1–14. doi:10.1017/ATSIP.2019.12.

[22] Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11), 2673–2681. doi:10.1109/78.650093.

[23] Yoon, K. (2014). Convolutional Neural Networks for Sentence Classification [OL]. arXiv Preprint. doi:10.48550/arXiv.1408.5882

[24] Zhang, Y., & Wallace, B. (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint, arXiv:1510.03820. doi:10.48550/arXiv.1510.03820

[25] Gojali, S., & Khodra, M. L. (2016). Aspect based sentiment analysis for review rating prediction. 2016 International Conference on Advanced Informatics: Concepts, Theory and Application (ICAICTA). doi:10.1109/icaicta.2016.7803110.