JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY
http://e-journal.uum.edu.my/index.php/jict

How to cite this article:
Ahmad Genadi, R., & Khodra, M. L. (2022). Opinion triplet extraction for aspect-based sentiment analysis using co-extraction approach. *Journal of Information and Communication Technology, 21*(2), 255-277. https://doi.org/10.32890/jict2022.21.2.5

# Opinion Triplet Extraction for Aspect-Based Sentiment Analysis Using Co-Extraction Approach

[1]Rifo Ahmad Genadi & *[2]Masayu Leylia Khodra
School of Electrical Engineering and Informatics,
Institut Teknologi Bandung, Indonesia

23520033@std.stei.itb.ac.id
*masayu@staff.stei.itb.ac.id
*Corresponding author

Received: 2/10/2021 Revised: 25/12/2021 Accepted: 10/1/2022 Published: 7/4/2022

## ABSTRACT

In aspect-based sentiment analysis, tasks are diverse and consist of aspect term extraction, aspect categorization, opinion term extraction, sentiment polarity classification, and relation extractions of aspect and opinion terms. These tasks are generally carried out sequentially using more than one model. However, this approach is inefficient and likely to reduce the model's performance due to cumulative errors in previous processes. The co-extraction approach with Dual crOss-sharEd RNN (DOER) and span-based multitask acquired better performance than the pipelined approaches in English review data. Therefore, this research focuses on adapting the co-extraction approach where the extraction of aspect terms, opinion terms, and sentiment polarity are conducted simultaneously from review texts. The co-extraction approach was adapted by modifying the original frameworks to perform unhandled subtask to get the opinion triplet.

Furthermore, the output layer on these frameworks was modified and trained using a collection of Indonesian-language hotel reviews. The adaptation was conducted by testing the output layer topology for aspect and opinion term extraction as well as variations in the type of recurrent neural network cells and model hyperparameters used, and then analysing the results to obtain a conclusion. The two proposed frameworks were able to carry out opinion triplet extraction and achieve decent performance. The DOER framework achieves better performance than the baselines on aspect and opinion term extraction tasks.

**Keywords:** Aspect-based sentiment analysis, opinion triplet, co-extraction, Dual crOss-sharEd RNN, span-based multitask.

## INTRODUCTION

Aspect-based sentiment analysis (ABSA) analyses review texts for specific opinion targets to obtain meaningful information. It consists of at least 5 subtasks, namely aspect expression extraction, aspect categorization, sentiment expression extraction, sentiment polarity classification, and the extraction of aspect relations and sentiment expression (Pontiki et al., 2016; Chen et al., 2018). SemEval-2016 categorized 2 of the approaches, aspect term extraction, and sentiment polarity classification, as Task 5 Subtask 1 (Pontiki et al., 2016). Aspect expression extraction is used to obtain the attributes (or aspects) in an opinion. For example, in the following laptop review text, "I like the keyboard and the monitor, but the price is too expensive." The extracted terms are keyboards, monitors, and prices, in addition sentiment polarity classification was also determined (Hu & Liu, 2004). It was positive for the keyboard and monitor and negative for the price.

Generally, aspect term extraction that is a sequence labelling task, and sentiment classification that is a classification task are carried out separately, one-by-one (pipelined). The problem arises from the inefficient separation of these 2 methods because there are two model constructed, therefore, the error in the previous stage is carried over to the next process (Luo et al., 2020). Several research, such as Luo et al. (2019), and Zhao et al. (2020), proposed a framework that can be improved to solve the numerous issues associated with these sub-tasks and able to perform opinion triplet extraction.

This research focuses on modifying the Dual Cross-Shared RNN (DOER) (Luo et al., 2019) and SpanMLT architectural models (Zhao et al., 2020) to carry out aspect plus opinion terms extractions and sentiment polarity classification. Its performance was further compared to baseline models designed by Fernando et al. (2019) and Azhar et al. (2019). A collection of Indonesian hotel review texts on AiryRooms was used as case research because it contributed to creating an end-to-end ABSA framework. This simultaneously performs aspect term extraction, sentiment term extraction, and polarity classifications, enabling fine-grained information to be extracted without cumulative errors.

The remainder of the paper is organized as follows: Section 2 discusses the related works in ASBA on Indonesian language and co-extraction method for ASBA, Section 3 describe the detailed method of the experiment, section 4 discusses the result of experiment and section 5 discuss the conclusion and future works of the study.

## RELATED WORKS

Aspect-based sentiment analysis research on Indonesian language review texts on AiryRooms was carried out by Azhar et al. (2019) and Fernando et al. (2019) and is focused on completing either 1 or 2 tasks. Research by Azhar et al. (2019) was based on the multi-label aspect categorization and sentiment classification. Meanwhile, Fernando et al. (2019) analyzed the extraction of aspect and opinion terms. Azhar et al. (2019) further adapted the Extreme Gradient Boosting (XGBoost) Convolutional Neural Network (CNN) technique designed by Ren et al. (2017) for aspect categorization and sentiment classification with reference to the research carried out by Chen et al. (2017). The research was able to resolve the multi-label categorization problem with 10 selected aspect categories using the classifier chain method. The mean F1-macro, F1-micro, and hamming loss scores obtained for the aspect categorization tasks are 0.93, 0.93, and 0.02, respectively. The average F1-measure value for the sentiment classification and the combined test is 0.97 and 0.79, respectively **(**Azhar et al., 2019). Fernando et al. (2019), focusses on completing aspect and opinion term extractions by adapting models designed by Xu et al. (2018), using double embedding as word representation, along with implementing coupled multi-layer attentions architecture proposed by

Wang et al. (2017). The F1-measure obtained in the work of Fernando et al. (2019). is 0.91 for the token level and 0.91 for the entity level. Both of these research uses a pipelined approach and suffer from cumulative error.

A co-extraction or joint method may help reduce the error, for example, Luo et al. research (2020) proposed a DOER architecture that viewed aspect term extractions and polarity classification as sequence labelling task. Based on the experiments carried out, its performance is better compared to the previous state-of-the-art model, DE-CNN (Xu et al., 2018) combined with TNet (Li et al., 2018), reported in 3 datasets, namely laptop SemEval-2014 and restaurant reviews from SemEval 2014, 2015, 2016 (Pontiki et al., 2016) including English tweets. The F1-measure is 0.60 and 0.72 for laptop and restaurant review data, respectively (Luo et al., 2020). These results exceeded the state-of-the-art pipeline models DE-CNN (Xu et al., 2018) combined with TNet (Li et al., 2018), of 0.56 and 0.67 for laptop and restaurant review data, respectively. However, this framework does not handle the sentiment term extraction task and only predicts aspect-polarity pairs.

Zhao et al. (2020) proposed another approach that executed several tasks simultaneously, named the SpanMLT framework. It performed aspect and opinion term extractions and pair extractions by using span-based classification rather than sequence labelling (Zhao et al., 2020). The framework achieved state-of-the-art performance for aspect, and opinion term extractions, including pair extraction for restaurant and laptop review datasets (Pontiki et al., 2016; Fan et al., 2019; Wang et al., 2017). The results of the experiment indicated that their model significantly outperformed all compared methods. However, the proposed approach did not handle sentiment polarity classification.

Moreover, recent research on NLP also shows that using a pre-trained language model with architectural transformers (Vaswani et al., 2017) such as BERT (Devlin et al., 2019), XLM (Lample and Conneau, 2019) and IndoBERT (Wilie et al., 2020) lead to the realization of better performances in the downstream tasks. The utilization of these pre-trained language model can be useful for opinion triplet extractions.

## METHOD

In this section, two systems based on DOER (Luo et al., 2019), and SpanMLT frameworks (Zhao et al., 2019), were proposed to carry out opinion triplet extraction. Each of these is thoroughly explained in the following sections.

### Dataset

The experiment carried out on the dataset obtained from AiryRooms consists of 5,000 review texts or 78,603 tokens. These were split into 3,000 train data, 1,000 validation data, and 1,000 test data, as shown in Table 1. Each token in the review was given two labels, namely term, and polarity. The evaluation method used is F1-score, based on the exact match of triples contained in the review text, besides each sentence either contained several aspects and opinion terms, or none. An aspect is referred to one or more opinion terms, and vice versa. This needs to be considered in the pairing of these terms. In addition, several sentences contained opinion terms without explicitly mentioning their aspects, meanwhile, the average length of reviews in the dataset was 15.72 words, while only 253 sentences (~ 5% of the total) had over 40 lengths. Based on this information, it was concluded that the reviews are usually brief because they are straight to the point and convey an overall impression of the client's stay as not many customers comment on every aspect of the hotel. The dataset is annotated by 2 annotators and further examined by a reviewer. The result shows that 123.268 unannotated reviews were used to perform masked language model post-training for IndoBERT (Xu et al., 2019).

**Table 1**

*Distribution of Labels*

| Term Label | Number of Data | Polarity Label | Number of Data |
|---|---|---|---|
| **B-ASPECT** | 8,762 | **POSITIVE** | 5,607 |
| **I-ASPECT** | 2,871 | **NEGATIVE** | 6,026 |
| **B-SENTIMENT** | 12,036 | **OTHER** | 66,970 |
| **I-SENTIMENT** | 5,333 | **Total** | 78,603 |
| **OTHER** | 49,601 | | |
| **Total** | 78,603 | | |

## Improved DOER-based Framework

The modifications made to the original framework include changes in the classifier and the creation of an additional module used to perform relation extraction of aspect and opinion terms. The proposed system uses a dense softmax activation layer to execute multi-class classification, which categorizes the token as $\in$ {Aspect, Opinion, Other} in IOB notation, rather than only determining the aspect term (Luo et al., 2019). The additional module adopted for the relation extraction is a heuristic algorithm that pairs each aspect to the 'closest' opinion terms. For example, in Table 2, the aspect term '*Kamar'* (room) was paired with the opinion '*bersih sekali*' (very clean) because of its close position.

The proposed system is made up of several modules, explained in the next subsections. An example of the overall process is shown in Table 2, including the input and output of each module. The architectural design of the proposed system and co-extraction models are shown in Figure 1 and Figure 2.

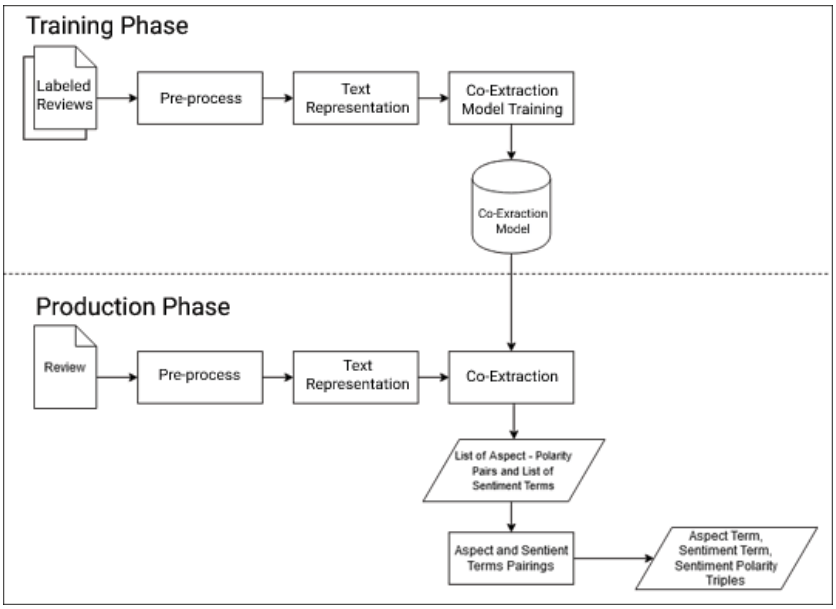
**Table 2**

*Overall DOER-based System Process*

| Module | Input | Output |
|---|---|---|
| **Pre-processing** | *"Kmr brsh sekali. Pelayanan memuaskan."*<br>(The room is very clean. The service is satisfactory) | [*'kamar', 'bersih', 'sekali', '.', 'pelayanan', 'memuaskan', '.'*]<br>(The room is very clean with satisfactory service) |
| **Text Representation** | *['kamar'* (room*), 'bersih'* (clean*), 'sekali' (very), '.', 'pelayanan'* (service*), 'memuaskan'* (satisfactory*), '.']* | Word vector<br>[-3.6405697 … 3.9813783, 0.12211756 … -0.278016, ..., 0.5733097 … 5.7085686] |
| **Model** | Word vector<br>Size: (400, max_sentence_length)<br>[-3.6405697 … 3.9813783, … -0.278016, ..., 0.5733097 … 5.7085686] | List of predicted labels (joint tag):[B-ASPECT, B-SENTIMENT, B-SENTIMENT, …]<br>[PO, O, O, …] |
| **Final Output Generation** | List of predicted labels (joint tag): [B-ASPECT, B-SENTIMENT, B-SENTIMENT, …]<br>[PO, O, O, …] | List of triplets:<br>[*('kamar'* (room*), 'bersih sekali'* (very clean*), PO), ('pelayanan'* (service*), 'memuaskan'* (satisfactory*), PO)*] |

**Figure 1**

*DOER-based System Architecture*

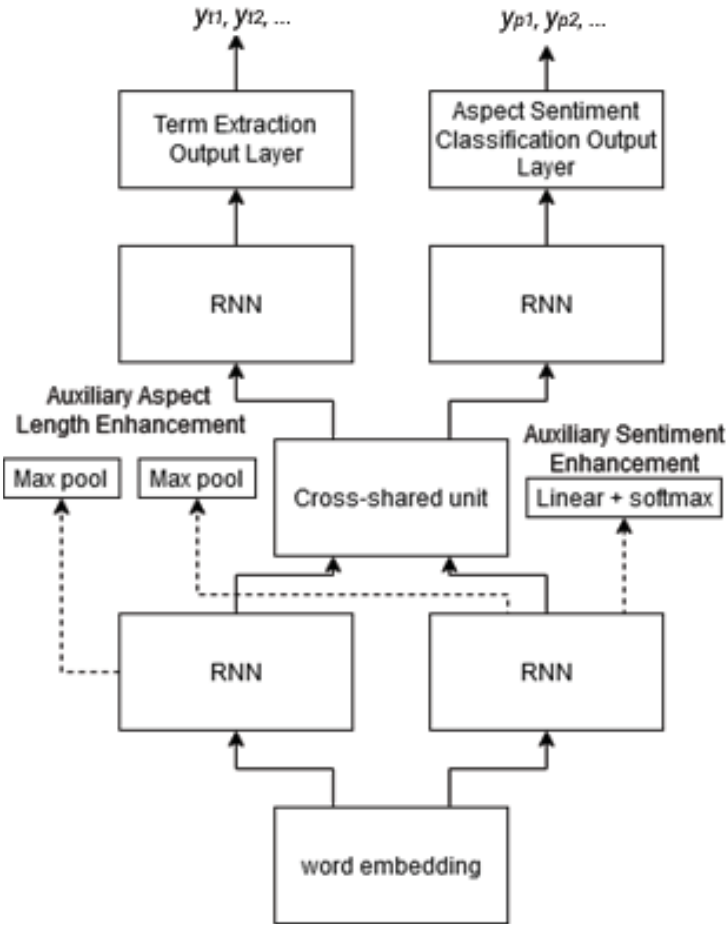



*1) Pre-processing*

The reviews were initially pre-processed with additional word normalization rules used to determine unhandled cases in the original inaNLP library (Purwarianti et al., 2016). The procedure consists of case-folding, word normalization, and tokenization. Normalization is carried out to correct typos, errors, abbreviations and translate informal words--often found in reviews-- to its formal form. For example, the text *"Hotelnya brsih, fasilitas lkp. Cuma resepsionisnya jutek banget"* is transformed into a list of tokens ["*hotelnya*" (the hotel), "*bersih*" (clean), ",", "*fasilitas*" (facilities), "lengkap" (complete), ".", "*cuma*" (but), "*resepsionisnya*" (the receptionist), "*jutek*" (stiff), "*banget*" (very), "."].

**Figure 2**

*Co-extraction Model Architecture for DOER-based system*




*2) Text Representation*

Each token on the list is transformed into its vector representation, while the double embedding (Xu et al., 2018) technique is used to generate the general-purpose and domain-specific embedding. These differ according to whether they are trained by an in-domain corpus. FastText (Bojanowski et al., 2017) is applied due to its ability to use sub-word N-gram embedding to decipher out-of-vocabulary words that often appear in the dataset. Padding is performed, thereby enabling the mini-batch learning to be executed, and all reviews are set to a specified maximum length. For example, assuming the specified length is *n*, any sentence shorter than it tends to be added to the dummy tokens after the last one, thereby making it equal. However, those that are lengthier are deducted and only take *n* initial tokens.

The text representation module is used to convert pre-processed input into a vector with a certain dimension. The commonly used ones are 300 and 100 for general-purpose and domain-specific embedding, and then both are concatenated. For example, the results obtained from the pre-processing module is a list of tokens, [“*hotelnya*” (the hotel), “*bersih*” (clean), “,”, “*fasilitas*” (facility), “*lengkap*” (complete), “.”, “*cuma*” (but), “*resepsionisnya*” (the receptionist), “*jutek*” (stiff), “*banget*” (very), “.”], converted to a vector with a (400, max_ sentence_length) dimension.

*3) Model Training*

The model is used to perform term extraction (both aspect and opinion terms) and aspect sentiment classification. These are both viewed as sequence labelling tasks, which enable every token to have 2 labels, term and polarity tags. The term tags are labelled with IOB notation and are defined with 5 labels, namely B-ASPECT (beginning of aspect term), I-ASPECT (inside of aspect term), B-SENTIMENT (beginning of opinion term), I-SENTIMENT (inside of opinion term), and O (other). On the contrary, the polarity tags are described using 3 labels, namely PO (positive), NG (Negative), and O (other). A modified DOER model is used to predict each token’s label in a sequence. The difference is based on the aspect term extraction output layer topology. It also involves 2 auxiliary tasks, namely Aspect Term Length Enhancement (AuL) and Sentiment Lexicon Enhancement (AuS) (Luo et al., 2019). The model hyperparameters are hidden units for the RNN layers, cross share k, dropout rate, and the term extraction output layer topology, which consists of 2 variations. This involves using either a single or separate layer to predict aspect and opinion terms. Meanwhile, Keras (Chollet et al., 2015) was used to design and train the model.

*4) Final Output Generation*

Labels generated by the model were further processed to obtain the final output, comprising aspect and opinion terms and sentiment polarity in the review text. The aspect term acts as the boundary for the polarity labels, and then the number of times each category appeared within it is counted, and the one that appears most is selected. However, the first label is chosen, assuming 2 or more labels have equal occurrences. Each aspect term–polarity pair then chooses the closest opinion term, for example, the desired output for “*hotelnya*

*bersih, fasilitas lengkap. Cuma resepsionisnya jutek banget"* (The hotel is clean, the facilities are complete, however, the receptionist is very stiff) is a list of triples [(*hotelnya* (the hotel), *bersih* (clean), PO], (*fasilitas* (facilities), *lengkap* (complete), PO), (resepsionisnya (the recepsionist), *jutek banget* (very stiff), NG)].

## Improved Span-based Framework

The original SpanMLT was modified in accordance with the following: 1) the relation scorer part that performs a multi-class classification was used to categorize each span pair into ∈ {Positive, Negative, Null}. The positive, negative, and null classes mean that the span pair is related positively, negatively, and the existence of no relationship respectively, 2) apply the logits from term scorer instead of using the new FFNN in relation scorer to rank and select the top k span that is paired, and 3) the architecture of FFNN used in the relation scorer is larger, in addition 2 hidden layers with sizes 512, and 256 were used. This led to several assumptions on its implementation, namely: 1) The span representation is the average sum of each token, and 2) The base encoder weight is updated during the fine-tuning part.

The proposed system constitutes several modules, each of which were explained in the next subsections. The overall procedure, including the input and output in each module as well as the model architecture, is shown in  and Figure 3.

**Table 3**

*Overall Span-Based System Process*

| Module | Input | Output |
|---|---|---|
| **Tokenization** | *"Handuk warnanya putih kehitaman"* (The towels are blackish white) | Sub-word tokenization result: ["[CLS]", "Hand", "##uk", "warnanya", "putih", "kehitaman", "[SEP]"] Token ids: [3, 4414, 156, 321, 54, 8744, 4] |
| **Encoder** | [3, 4414, 156, 321, 54, 8744, 4] | Tensor with size (8, 768) 8 is the number of tokens |

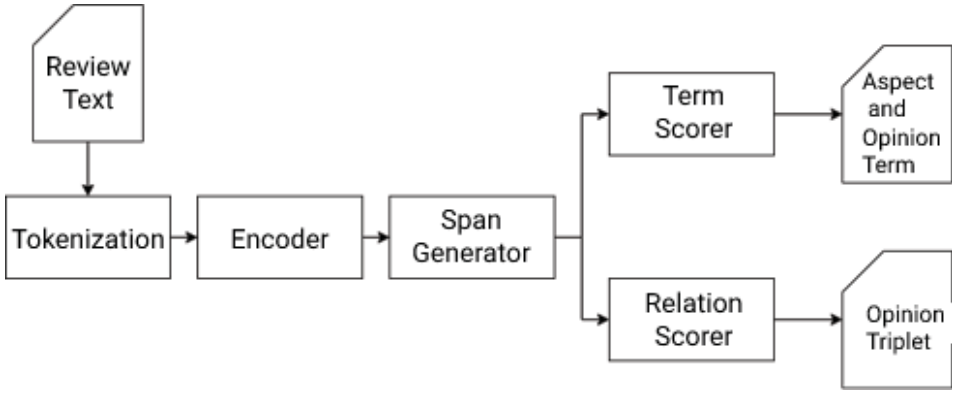
(continued)

| Module | Input | Output |
|---|---|---|
| **Span Generator** | Encoder output, tensor with size (number_of_ tokens, 768) | Tensor with size (num_ of_spans, 768) Number of generated spans affected by max_span_length and sentence length. |
| **Term scorer** | Tensor with size (num_of_spans, 768) | List of aspect and opinion terms (*"Handuk"* (Towel), Aspect) (*"warnanya"* (colored), O) (*"putih kehitaman"* (blackish white), Sentiment) |
| **Relation Scorer** | Tensor with size (num_of_spans, 768) | List of opinion triplets ("*Handuk*" (Towel), "*putih kehitaman" (blackish white)*, Positive) |

The span-based framework applied a similar pre-processing approach as the DOER-based one. Additionally, the *whitespace cleaning*, *punctuation splitting*, and *word piece tokenization* were leveraged from the IndoBERT model because of its pre-trained tokenizer. Afterward, it was processed by the base encoder, a pre-trained language model (Wilie et al., 2020), to generate a context-aware representation of the sentence. The span generator further process this to enumerate all possible spans and return span representations. Finally, it becomes the input of the 2 framework classifiers, namely term and relation scorers. The term scorer classifies each span as aspect and opinion terms, or neither of them (Zhao et al., 2020). Meanwhile, the relation scorer classifies each pair as positively and negatively related or not related.

**Figure 3**

*Span-based Framework Architecture*



## Experiment

The experimental goal is to determine the best configurations of hyperparameters for the co-extraction models. The analysis of the DOER-based framework consists of 6 scenarios, and their various goals are shown in Table 4. The model was developed using Adam optimizer, batch size 16, and categorical cross-entropy as the loss function. In addition, early stopping was adopted, with patience set to 5 and the number of epochs at 20. The default configuration for model hyper-parameters is similar to the configuration designed by Luo et al. (2020). The ideal configuration for each scenario is used to determine the best configuration.

**Table 4**

*DOER-based Framework Experiment Scenarios*

| Experiment Id | Goal |
|---|---|
| **D0** | Find the best padding method |
| **D1** | Find the best RNN Cell type |
| **D2** | Find the best output layer topology for term extraction |
| **D3** | Find the best number of hidden units, number of cross share k, and dropout rate. |
| **D4** | Compare the performance between models that execute AuL and those that do not perform AuL |
| **D5** | Compare the performance between models that execute AuS and those that do not perform AuS |

Meanwhile, the span-based model experiment consists of 4 scenarios, and their diverse goals are shown in Table 5. The default configuration follows the settings designed by Zhao et al. (2020). Its details for the controlled hyper-parameters are shown in Table 5. Additionally, the configuration used for the post and pre-trained language models are shown in Table 5.

**Table 5**

*Span-based Framework Experiment Scenarios*

| Experiment Id | Goal |
|---|---|
| **S0** | Find the best pre-trained language model |
| **S1** | Find the best top *k* span percentage to be paired |
| **S2** | Find the best λt/ λr |
| **S3** | Find the best maximum span length |

**Table 6**

*Default Configuration of Controlled Hyperparameters for Span-based Framework*

| No | Hyperparameter | *Default value* |
|---|---|---|
| **1.** | Batch size | 8 |
| **2.** | Optimizer | 1 |
| **3.** | Seed | 42 |
| **4.** | Learning rate | 2e-5 |
| **5.** | Maximum sentence length | 40 kata |
| **6.** | Term scorer's number of hidden layers | 1 |
| **7.** | Term scorer's hidden layer size | 512 |
| **8.** | Patience | 5 |
| **9.** | Max epochs | 200 |
| **10.** | Dropout rate | 0.10 |
| **11.** | Relation scorer's number of hidden layers | 2 |
| **12.** | Relation scorer's hidden layer sizes | 512, 256 |

## Evaluation

Models with the best configuration are further evaluated using the test data. Besides, model performance is compared with some previous research on AiryRooms hotel reviews for extraction of aspects and opinions. It was also compared to the CMLA + DE (Fernando et al., 2019), and BERT-base frameworks, fine-tuned to sequence labelling task and opinion triplet extraction, were analyzed. Evaluation for aspect and opinion term extraction task was executed by comparing the entity level F1-scores to the test data. The model was compared with CMLA + DE (Fernando et al., 2019) to determine the best configuration batch size was set to 32, double embedding for word

embedding. Consequently, the number of hidden units, layer-coupled attentions, tensors, and the dropout rate were set to 50, 2, 20, and 0.5 respectively. An example of the entity-level is shown in Table 7, furthermore the evaluation for opinion triplet extraction is carried out by measuring the model's accuracy. A prediction is assumed to be valid, supposing there are exact triplets in the ground truth. In addition, the number of correct predictions is divided by the number of expected triplets.

**Table 7**

*Example of Evaluation of Entity-Level Term Extraction Performance*

| Label | TP | FP | FN | Precision | Recall | F1-score |
|---|---|---|---|---|---|---|
| **ASPECT** | 1 | 1 | 3 | 0.5 | 0.33 | 0.4 |
| **SENTIMENT** | 0 | 0 | 1 | 0 | 0 | 0 |

## Results and Discussion

The results of experiment D0 are shown in Table 8, with a post-padding process at a fixed length similar to the average review text used gives a better performance compared to using post-padding to the longest sentence length in the dataset. In this case, 95 percent of the review text had less than 20 words, the longest was 114 words, and 40 were selected as the maximum sentence length. Shorter padding implies fewer data to process, which results in faster training. Meanwhile, the results of experiment D1 are shown in Table 9. Based on this, BiReGU achieved better results than other RNN cell variations. This customized RNN cell, proposed by Luo et al. (2019), is specifically tailored for aspect-based sentiment analysis tasks. BiReGU architecture effectively enables information transfer to the next layer (Luo et al., 2019).

**Table 8**

*Result of Experiment D0*

| Padding-Method | Opinion Triplet Extraction Accuracy |
|---|---|
| ***Post-padding*** **to a fixed length (40) that isn't too far from average sentence length** | 0.73 |
| ***Post-padding*** **to longest sentence length in dataset** | 0.72 |

**Table 9**

*Result of Experiment D1*

| RNN cell type | Opinion Triplet Extraction Accuracy |
|---|---|
| **BiGRU** | 0.61 |
| **BiLSTM** | 0.65 |
| **BiReGU** | 0.73 |

Based on Experiment D2, using the same dense layer for both aspect and opinion term extraction is more effective compared to using the separate ones for each of them. The model with a single output layer categorizes tokens into 5 classes ('B-ASPECT,' 'I-ASPECT,' 'B-SENTIMENT,' 'I-SENTIMENT,' 'O') achieved better accuracy, as shown in Table 10. Several values related to the number of hidden units, cross-shared-k, and dropout rate were tested in experiment D3. However, 27 combinations were tried, and the best configuration for these hyperparameters was 250, 5, and 0.5 for the number of hidden units, cross-shared-k, and dropout rate, respectively. The 3 most ideal ones are shown in Table 11.

**Table 10**

*Result of Experiment D2*

| Term Extraction Output Layer Topology | Opinion Triplet Extraction Accuracy |
|---|---|
| **Same output layer for term extraction** | 0.73 |
| **Separate output layer for term extraction** | 0.72 |

**Table 11**

*Result of Experiment D3*

| Hyperparameter | | | Opinion Triplet Extraction Accuracy |
|---|---|---|---|
| **Hidden units** | **dropout rate** | **cross share k** | |
| **250** | 0.5 | 5 | 0.73 |
| **250** | 0.5 | 1 | 0.73 |
| **300** | 0.25 | 5 | 0.72 |

Table 12 shows the results of D4, based on the experiment, the model that performed AuL was less effective compared to the other one. This result is inconsistent with Luo et al. (2019), although it needs to be noted that a different task was performed, it did not execute opinion term extraction and uses English dataset. However, this is indicates predicting the average length of the aspect approach tends not to be helpful for Indonesian dataset. Based on Experiment P5, a model that performs AuS achieves better performance compared to the other one. This is consistent with Luo et al. (2019), which was concluded that predicting the subjectivity of each word correctly aids term and polarity co-extraction. The F1-score of D5 is shown in Table 13.

**Table 12**

*Result of Experiment D4*

| Model | Opinion Triplet Extraction Accuracy |
|---|---|
| **Model with AuL** | 0.73 |
| **Model without AuL** | 0.74 |

**Table 13**

*Result of Experiment D5*

| Model | Opinion Triplet Extraction Accuracy |
|---|---|
| **Model with AuS** | 0.74 |
| **Model without AuS** | 0.71 |

Meanwhile, the span-based framework's experiments result are shown in Table 14, Table 15, Table 16, and Table 7. Experiment S0 shows that the masked language model post-training using domain-specific data helps achieve better performance when it is fine-tuned to aspect-based sentiment analysis tasks. Such process helps the framework to generate better representation suited to the used domain. IndoBERT-IDPT simply means IndoBERT that is post-trained to the domain-specific dataset (IndoBERT in-domain post-training). Based on scenarios S1, S2, and S3, it is obvious that those parameters insignificantly affected the model's performance. However, there is a need to note that having less k span candidates and a shorter max span length means less data to compute and a faster training process.

**Table 14**

*Result of Experiment S0*

| Pre-trained Language Model | Opinion Triplet Extraction Accuracy |
|---|---|
| **IndoBERT-base** | 0.46 |
| **IndoBERT-IDPT** | 0.55 |

**Table 15**

*Result of Experiment S1*

| Paired Top k span candidate's percentage | Opinion Triplet Extraction Accuracy |
|---|---|
| **0.2** | 0.53 |
| **0.3** | 0.53 |
| **0.4** | 0.53 |

**Table 16**

*Result of Experiment S2*

| Term scorer and Relation scorer weight ratio ($\lambda t / \lambda r$) | Opinion Triplet Extraction Accuracy |
|---|---|
| **0.75** | 0.52 |
| **1** | 0.53 |
| **1.25** | 0.52 |

**Table 17**

*Result of Experiment S3*

| Maximum span length | Opinion Triplet Extraction Accuracy |
|---|---|
| **2** | 0.53 |
| **4** | 0.53 |
| **8** | 0.53 |

## Evaluation and analysis

The results of the evaluated term and opinion triplet extraction are shown in Table 18 and Table 19. GD and GS denote the outcome

of the evaluated DOER and span-based models designed for this research. FR and SL represent the evaluated model's results designed by Fernando et al. (2019) and the BERT-based framework that is fine-tuned to sequence labelling tasks. These results indicate that the GD model achieved better performance compared to the baseline. However, the GS framework was ineffective due to the modifications and assumptions of the span-based model implementation. The GD model mispredicted 942 words in the test data, and one of the cases involving misclassification was when the aspect and opinion terms were excluded in the train or validation data. Meanwhile, 643 words in the test data did not appear in the train or validation data, 129 of these were part of the aspect term, while 183 were part of the opinion term, and 331 were neither. The model failed to extract 45 and 96 aspect and opinion terms, respectively. For example, the model failed to extract the opinion term *"bekas ada spot"* (dirty marks) and *"putih kehitaman"* (blackish white) from the review text *"Kamar bersih tetapi sayang linen kotor, bekas ada spot, kamar mandi bau, handuk sudah waktunya di ganti karena warnanya putih kehitaman, wifi susah (sinyal ada tetapi tidak mau terhubung)";* (The room was clean but, unfortunately, the linen had dirty marks, the towel needs to be replaced sooner because of its color which was blackish white, additionally the wifi was difficult to use (signal found but unable to connect).

Another case of misclassification in the GD model occurred when it incorrectly predicted a word that is often part of an opinion term. For example, the model failed to extract the opinion term, *"perlu diperbaiki"* (needs to be repaired) from the review text *"yang perlu diperbaiki itu wifi lantai 2, sering disconnect. Jadi suka lag Kalau Lagi main mobile legends;"* (wifi on the 2nd floor has to be replaced, because it is often disconnected. Besides, it often lags when I am playing mobile legends." However, this issue is due to inconsistency in the data labelling phase. In accordance with the opinion triplet extraction task, the DOER-based model achieved better performances compared to the span-based platform, irrespective of the fact that it only adopted a heuristic approach to pair the aspect and opinion terms. Since both proposed frameworks constrained the maximum sentence length to 40, some extractions from longer texts were missing, such cases frequently occurred in the span-based model because it uses sub-word tokenization. The difference between the two frameworks

is likely caused by the span-based model's inability to generalize properly, it is also prone to error in the data annotations. For instance, in the test set, the review *"Lumayan bagus, tempat dkat Malioboro, harga hemat"* means "It is quite good, affordable and located close to Malioboro", both *"harga"* and *"hemat"* are predicted as others. This is because there is only one illustration in the train set with the co-occurrence of '*harga'* and '*hemat'* and it was mislabeled as others. Another issue in the GS model is the huge number of enumerated spans and possible pairs between the spans.

**Table 18**

*Evaluation of Term Extraction Results for Entity-Level*

| Label | F1-score | | | |
|---|---|---|---|---|
| | GD | GS | FR | SL |
| **ASPECT** | 0.89 | 0.74 | 0.87 | 0.87 |
| **SENTIMENT** | 0.90 | 0.76 | 0.88 | 0.88 |
| **Average** | 0.90 | 0.75 | 0.87 | 0.87 |

**Table 19**

*Evaluation of Opinion Triplet Extraction Results*

| Model | Accuracy |
|---|---|
| **GD** | 0.71 |
| **GS** | 0.56 |

## CONCLUSION AND FUTURE WORKS

The ABSA co-extraction approach was used to achieve better performance on the term extraction task and able to do opinion triplet extraction. The two modified frameworks used to execute opinion triplet extraction, have exhibited decent performances. The experiment results for aspect and opinion term extraction tasks show the DOER framework's effectivity, which outperforms sequence labelling approaches designed by Fernando et al. (2019) and even the fine-tuned BERT model. Although, the proposed span-based multitask framework doesn't seemingly work like the one formulated by Zhao

et al. (2020), which was affected by modifications and assumptions made by the present implementation.

For further research, there is a need to combine some components of the two frameworks or try different base encoders with the span-based model. The use of span representation with width representation and span pair representation with distance representation which proposed by Xu et al. (2021) might give better representation and leads to better performance. A different approach might also leads to better performance in opinion triplet extraction, for example using Graph Neural Network (GNN) (Chen et al., 2021), using a generative text to text model (Zhang et al., 2021), decomposing triplet extraction into target tagging, opinion tagging and sentiment tagging (Chen et al., 2022), or uses span-sharing joint extraction (Li et al., 2022) . The importance of consistency in the data annotation process also needs to be emphasized, such as the various aspects that needs to be marked, and the specificity of the aspects.

## ACKNOWLEDGMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

## REFERENCES

Azhar, A. N., Khodra, M. L., & Sutiono, A. P. (2019). Multi-label Aspect Categorization with Convolutional Neural Networks and Extreme Gradient Boosting. *Proceedings of the International Conference on Electrical Engineering and Informatics*. https://doi.org/10.1109/ICEEI47359.2019.8988898

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. https://doi.org/10.1162/tacl_a_00051

Chen, F., Yang, Z., & Huang, Y. (2022). A multi-task learning framework for end-to-end aspect sentiment triplet extraction. *Neurocomputing*,*479*,12–21.https://doi.org/10.1016/j neucom.2022.01.021

Chen, G., Ye, D., Xing, Z., Chen, J., & Cambria, E. (2017). Ensemble application of convolutional and recurrent neural networks for multi-label text categorization. *Proceedings of the International Joint Conference on Neural Networks.* https://doi.org/10.1109/IJCNN.2017.7966144

Chen, G., Zhang, Q., & Di Chen. (2018). A Pair-Wise Method for Aspect-Based Sentiment Analysis. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-319-94307-7_2

Chen, Z., Huang, H., Liu, B., Shi, X., & Jin, H. (2021). *Semantic and Syntactic Enhanced Aspect Sentiment Triplet Extraction*. https://doi.org/10.18653/v1/2021.findings-acl.128

Chollet, F., & others. (2015). Keras. https://keras.io.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL HLT 2019. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*.

Fernando, J., Khodra, M. L., & Septiandri, A. A. (2019). Aspect and Opinion Terms Extraction Using Double Embeddings and Attention Mechanism for Indonesian Hotel Reviews. *Proceedings - 2019 International Conference on Advanced Informatics: Concepts, Theory, and Applications, ICAICTA 2019*. https://doi.org/10.1109/ICAICTA.2019.8904124

Hu, M., & Liu, B. (2004). Mining and Summarizing Customer Reviews. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 168–177. https://doi.org/10.1145/1014052.1014073

Lample, G., & Conneau, A. (2019). Cross-lingual Language Model Pretraining. Advances in Neural Information Processing Systems (NeurIPS).

Li, X., Bing, L., Lam, W., & Shi, B. (2018). Transformation networks for target-oriented sentiment classification. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*. https://doi.org/10.18653/v1/p18-1087

Li, Y., Lin, Y., Lin, Y., Chang, L., & Zhang, H. (2022). A span-sharing joint extraction framework for harvesting aspect sentiment triplets. *Knowledge-Based Systems*, *242*, 108366. https://doi.org/10.1016/j.knosys.2022.108366

Luo, H., Li, T., Liu, B., & Zhang, J. (2020). Doer: Dual cross-shared RNN for aspect term-polarity Co-extraction. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference.* https://doi.org/10.18653/v1/p19-1056

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Al-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., Hoste, V., Apidianaki, M., Tannier, X., Loukachevitch, N., Kotelnikov, E., Bel, N., Jiménez-Zafra, S. M., & Eryigit, G. (2016). SemEval-2016 task 5: Aspect based sentiment analysis. *SemEval 2016 - 10th International Workshop on Semantic Evaluation, Proceedings.* https://doi.org/10.18653/v1/s16-1002

Purwarianti, A., Andhika, A., Wicaksono, A. F., Afif, I., & Ferdian, F. (2016). InaNLP: Indonesia natural language processing toolkit, case study: Complaint tweet classification. 4th IGNITE Conference and 2016 International Conference on Advanced Informatics: Concepts, Theory and Application, ICAICTA 2016. https://doi.org/10.1109/ICAICTA.2016.7803103

Ren, X., Guo, H., Li, S., Wang, S., & Li, J. (2017). *A novel image classification method with CNN-XGBoost model.* Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). https://doi.org/10.1007/978-3-319-64185-0_28

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems.*

Wang, W., Pan, S. J., Dahlmeier, D., & Xiao, X. (2017). Coupled multi-layer attentions for co-extraction of aspect and opinion terms. *31st AAAI Conference on Artificial Intelligence, AAAI 2017.*

Wilie, B., Vincentio, K., Winata, G.I., Cahyawijaya, S., Li, X., Lim, Z., Soleman, S., Mahendra, R., Fung, P., Bahar, S., & Purwarianti, A. (2020). IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding. AACL.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. (2020). *Transformers: State-of-the-Art Natural Language Processing.* https://doi.org/10.18653/v1/2020.emnlp-demos.6

Xu, L., Chia, Y. K., & Bing, L. (2021). Learning span-level interactions for aspect sentiment triplet extraction. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*. https://doi.org/10.18653/v1/2021.acl-long.367

Xu, H., Liu, B., Shu, L., & Yu, P. S. (2018). Double embeddings and cnn-based sequence labeling for aspect extraction. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*. https://doi.org/10.18653/v1/p18-2094

Xu, H., Liu, B., Shu, L., & Yu, P. S. (2019). BERT post-training for review reading comprehension and aspect-based sentiment analysis. NAACL HLT 2019 - *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*. https://doi.org/10.18653/v1/N19-1242

Zhang, W., Li, X., Deng, Y., Bing, L., & Lam, W. (2021, August). Towards Generative Aspect-Based Sentiment Analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (pp. 504-510). https://doi.org/10.18653/v1/2021.acl-short.64

Zhao, H., Huang, L., Zhang, R., Lu, Q., & xue, hui. (2020). *SpanMlt: A Span-based Multi-Task Learning Framework for Pair-wise Aspect and Opinion Terms Extraction*. https://doi.org/10.18653/v1/2020.acl-main.296