Aspect Detection and Sentiment Classification using Deep Neural Network for Indonesian Aspect-Based Sentiment Analysis

Arfinda Ilmania
Prosa Solusi Cerdas
Bandung, Indonesia
arfinda.ilmania@prosa.ai

Abdurrahman
Prosa Solusi Cerdas
Bandung, Indonesia
abdurrahman.prasetyo@prosa.ai

Samuel Cahyawijaya
Prosa Solusi Cerdas
Bandung, Indonesia
samuel.cahyawijaya@prosa.ai

Ayu Purwarianti
Institut Teknologi Bandung
Bandung, Indonesia
ayu@stei.itb.ac.id

***Abstract*—Sentiment analysis can categorize an overall opinion from a sentence or a document. However, there are sentences with more than one opinion in a single sentence statement. This problem is solved by aspect-based sentiment analysis. We conduct experiments on this problem using Indonesian dataset with 2-step process: aspect detection and sentiment classification. On aspect detection, we compare two deep neural network models with different input vector and topology: word embedding vector which is processed using gated recurrent unit (GRU), and bag-of-words vector which is processed using fully-connected layer. On sentiment classification, we also compare two approaches of deep neural network. The first approach uses word embedding, sentiment lexicon and POS tags as the input vector, with bi-GRU based as the topology. The second one uses aspect matrix to rescale the word embedding vector as the input vector and convolutional neural network (CNN) based as the topology. Our work is compared to a baseline framework which uses different model for each aspect. The dataset has approximately 9800 reviews collected from various categories on popular online marketplaces in Indonesia. Our models generalize well over all aspects and achieve state-of-the-art performance on 4 out of 7 aspects compared to the baseline framework.**

Keywords- aspect-based; sentiment analysis; aspect detection; sentiment classification; Indonesian; deep neural network

I. INTRODUCTION

Sentiment analysis is becoming one of the important topics in natural language processing (NLP). It identifies and extracts subjective information then classifies them into three following polarities: positive, negative, and neutral. However, regular sentiment analysis can only identify one polarity for each sentence. Whereas it is possible that one sentence can have different sentiment polarities at once. For such reason, aspect-based sentiment analysis is needed.

Aspect-based sentiment analysis is a task to find the sentiment polarity of each predefined aspects which expressed in a document. The sentiment polarity of a sentence is dependent on both content and aspect. For example, in the sentence *"The food is delicious, but the service is too slow.",* the expected sentiment polarity of 'food' and 'service' aspects are positive and negative respectively. Data can be extracted from any sources like websites, newspapers, or social media. However, extracting the sentiment polarity from these sources is not easy. Many contests are held to solve this kind of problem such as, task 4 in SemEval-2014 (SE-ABSA14) [1], task 12 in SemEval-2015 (SE-ABSA15) [2], and task 5 in SemEval-2016 (SE-ABSA16) [3]. Many approaches are used, from traditional model into more complex one like deep neural network.

Recently, deep learning-based approach has become a powerful technique to address this problem, such as Long Short-Term Memory [4], GRU [5], and CNN. However, there have been a few studies conducted to solve this problem in Indonesian language. Some of them use Support Vector Machine (SVM) and Naïve Bayes Classifier [6, 7]. Ekawati and Khodra [8] uses Conditional Random Field (CRF). Based on the research we have known so far, deep learning-based approach has been used only once to solve Indonesian aspect-based sentiment analysis. Different to Cahyadi and Khodra [9], instead of using three components, we chose to divide the process into two modules with a general model of aspect sentiment classification. We employ different deep learning models in sentiment classification, RNN and CNN.

This paper is organized in the following manner. Section 2 discusses related works. Section 3 discusses the proposed method. Section 4 shows the experimental and evaluation results. Finally, the conclusion and future works are in section 5.

II. RELATED WORKS

In this section, we present some previous works which are related on aspect-based sentiment analysis. In 2014, 2015, and 2016, the Semantic Evaluation provided dataset with annotation along some baselines for aspect-based sentiment analysis [1,2,3]. The baseline of aspect detection and sentiment classification use SVM. It also had term extraction process which occurs at some domains. The model could benefit from the exact location of aspect's term because polarity tends to be determined by near adjective words. Before extracting aspect, sentence could be filtered by subjectivity classifier [10]. Sentence having sentiment must be subjective. Performing subjectivity classifier would decrease objective sentence so aspect detection model could focus more to label other than none aspect, this aspect is present in some models.

978-1-7281-1175-9/18/$31.00 ©2018 IEEE

Target-dependent Long Short-Term memory (TD-LSTM) and Target-Connected Long Short-Term Memory (TC-LSTM) is made to keep attention only on the aspect terms [11]. Those models can only consider the terms but not aspect information which is important for aspect-based classification. Attention-based Aspect Embedding with Long-Short Term Memory (ATAE-LSTM) uses aspect embedding for each aspect which later will be concatenated with corresponding hidden vector of those aspects [12]. However, if there are a lot of aspects there will be more embeddings needed which make it become computationally expensive.

Bitmask Bidirectional Long Short-Term Memory (BB-LSTM) uses bitmask layer to keep attention only on the aspects [13]. Their model only resolves sentiment polarity classification problem and can obtain state-of-the-art result in SemEval-ABSA. Another approach by Jebbara and Cimiano [14] uses two models to extract aspect terms and determine their polarity. They use bidirectional GRU and a fully connected layer for each model, but they put many features for the input. They also use semantic resources like SenticNet and WordNet. Their approach obtained the best result on task 2 in ESCW 2016 Challenge for Semantic Sentiment Analysis and became the most innovative approach.

Another approach on accomplishing NLP task using deep learning is CNN. The problems involve polarity classification on various reviews, subjectivity classification, question type classification which are done by Kim [15]. The main reference we used to build CNN-based deep learning baseline to solve aspect-based sentiment analysis is Wang and Liu [16]. The aspect classification not only produce predicted aspects but produce probability of every word occurred in every aspect as well. This probability will be used as input in sentiment classification model. Xue and Li [17] use Gated-CNN to do the sentiment classification. The aspect feature controls the propagation of sentiment with aspect embedding of the given aspect category.

While in Indonesian, Fachrina and Widyantoro [7] develop aspect-sentiment classification in opinion mining using the combination of rule-based and machine learning. The algorithm they use for machine learning are SVM and naïve bayes classifier. Gojali and Khodra [6] use supervised learning for subjectivity classification. They use naïve bayes classifier and SVM to classify sentences and CRF for information extraction. Ekawati and Khodra [8] build a system consists of three steps: aspect detection, aspect categorization, and sentiment classification. Cahyadi and Khodra [9] build a similar system and make a new approach by using deep learning on aspect detection and sentiment classification. The number of models used on sentiment classification is as many as the number of predefined aspects.

III. Aspect-based Sentiment Analysis

Basically, our system consists of two models: aspect detection and sentiment classification. In both models, we compared two approaches each. In the aspect detection model, we compared the one using GRU based aspect classification with word embedding vector as the input layer, and bag-of-word vector input as the input layer with aim to get additional output of word weight. The first approach is common technique for text classification with deep learning approach, while the second approach is similar with Wang. The input for both approaches is taken from the pre-processing module which consists of word normalization, tokenization, and punctuation-symbol deletion. The result of aspect detection model along with the original word vector is passed onto the aspect sentiment classification model to get the sentiment class for each aspect. The flow is shown in Figure 1.

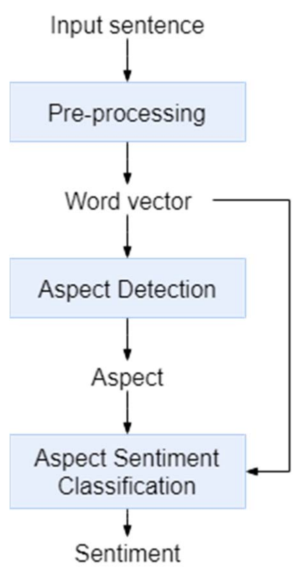

Figure 1. Process Flow of the Aspect based Sentiment Analysis

A. Aspect Detection Model

Such as mentioned in previous paragraph, there are two compared approaches in the aspect detection model. The first approach on this model adopts RNN for handling aspect detection. The input is a word embedding vector which is calculated by word2vec approach [18].

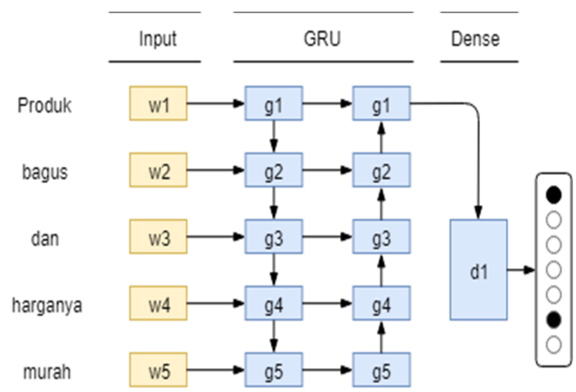

Figure 2. First Approach of Aspect Detection Model

There are two layers of GRU, where the first GRU returns a sequence of input for the second GRU. We connect the last output state of the second GRU to a fully connected layer with sigmoid activation function. The output of the fully connected layer is a logistic distribution of each aspect. The architecture is given in Figure 2.

The input for this approach is a set of words with their corresponding 500-dimensional word vectors w_i. The output of this approach is logistic distribution over 7 aspects. A sentence can contain an aspect if the corresponding output value of that aspect exceeds the threshold. We train the model with binary cross-entropy loss between expected aspects distribution and predicted aspect distribution of each sentence.

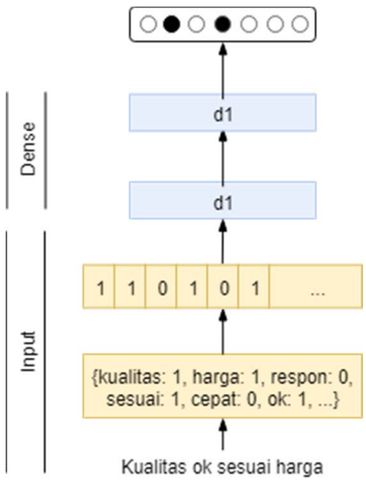

Figure 3. Second Approach of Aspect Detection Model

Second approach of our aspect detection model, as shown in Figure 3, receives bag-of-word as word vector input. The type of bag-of-word is unigram and the size is not limited to a number but all of words occur at training data is included to dictionary. The model has 2 fully connected layers as hidden layers. The output of this model is probability distribution over number of labels. Each label is a vector of size n, the number of predetermined aspects, $O = \{O_1, O_2, \ldots, O_n\}$. For $k \in [1,n]$ $O_k = 1/y$ if related aspect is present where y is total number of aspects present, otherwise $O_k = 0$. The model would compute loss using softmax cross-entropy. For generating aspect matrix, which denote the probability of each word occurred in each aspect, we build similar model with no hidden layer so weights on output layer has dimension of $W \times A$ where W is the number of distinct words and A is the number of aspects. The weights are extracted and used as aspect matrix.

B. Sentiment Classification Model

Every sentence has label over all predefined aspect. If a sentence is predicted not having a certain aspect, its label in that aspect is none. Otherwise, the sentence and the aspect are pipelined to sentiment classification. We compare two approaches in the sentiment classification. The first approach use Bi-GRU to classify the sentiment of each aspect resulted by the aspect detection model. It is similar to aspect-specific sentiment extraction model from Jebbara and Cimiano [14] but we use different features. The other differences are the use of predefined aspects and we do not pay attention to the corresponding terms of aspects. The input layer consists of a set of word vectors w_i which is concatenated with their corresponding sentiment embedding s_i and POS tag p_i for each word. The sentiment embedding is taken from sentiment lexicon (similar approach with Do [13]) which consists of words with positive and negative sentiment value. The sentiment lexicon is translated by Wahid and Azhari [19] from Hu and Liu [20]. Each word is represented by one-hot encoded vector s_i according to its sentiment polarity. When a word has no sentiment polarity (positive or negative) it will be represented as 'others'. For the POS tag, we employed POS tagger provided in prosa.ai which uses INACL standard POS tag list with 26 POS tags [21].

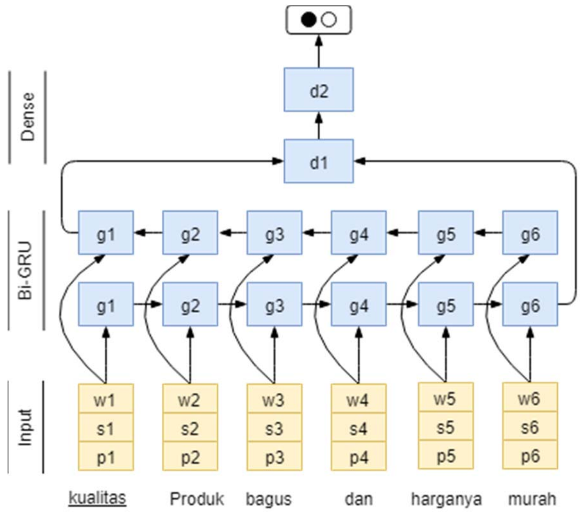

Figure 4. First Approach of Sentiment Classification Model

Before converting word into vector, we decided to add the corresponding aspects to the beginning of the sentences. We managed it so that sentences can only have one aspect. As a result of doing this, the sentence which contain more than one aspect need to be processed as many as the number of the aspects. For example, in the sentence: *"Kualitas oke, cepat, ramah, terima kasih banyak"* ("Good quality, fast, friendly, thank you so much") have two aspects: 'quality' and 'communication'. The input become, *"<u>kualitas</u> kualitas oke cepat, ramah, terima kasih banyak"* ("<u>quality</u> good quality, fast, friendly, thank you so much") and, "*<u>komunikasi</u> kualitas oke, cepat, ramah, terima kasih banyak*" ("<u>communication</u> good quality, fast, friendly, thank you so much").

The result of Bi-GRU layer is inputted into a fully connected layer with a softmax activation function and outputs polarity distribution over two polarity labels: 'positive' and 'negative'. Here, we do not handle 'neutral' label due to no sentence with 'neutral' label on the datasets. The index with the highest estimated probability becomes the predicted polarity label for the given aspect. The architecture is shown in Figure 4. The network is trained to minimize the categorical cross-entropy between expected and predicted sentiment polarity label.

The second approach of sentiment classification model, as given in Figure 5, receives word vectors which is padded to the maximum length of sentence in training data. The word vector then rescaled according to aspect given. The rescaling works on word level with aspect matrix that could be produced by second baseline of aspect model. The function of aspect matrix is selecting which word should be considered more by the model. It could be achieved because the presence of max pooling layer which select the highest value on feature map. Bigger magnitude on a word vector means it is more likely to be selected by max pooling layer. The aspect matrix is scaled a few times and yield the best result when it is scaled with minimum of zero and maximum of one.

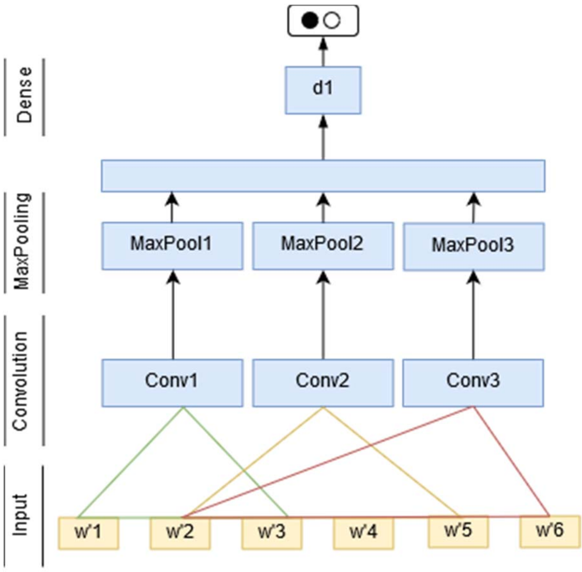

Figure 5. Second Approach of Sentiment Classification Model

We adopt temporal convolutional layers (one dimensional convolution). The windows would be applied to all dimensions of word vectors with stride of one. Then we would apply max pooling layer on each feature map. The output layer is softmax distribution over positive, and negative labels.

IV. EXPERIMENTS

A. Datasets

TABLE I. Statistics of the datasets

Aspect	Count	Pos	Neg
Accuracy	2967	2481	486
Quality	1466	1236	229
Service	966	777	189
Communication	1232	1087	145
Packaging	1849	1351	498
Delivery Time	937	818	119
Price	243	232	11

In order to see the performance of overall system, we performed an evaluation on the provided data for aspect-based sentiment analysis. We conducted experiment on the same dataset as Fachrina and Widyantoro [7] used. We divided the data into two parts, 70% for training data and the rest for testing data.

There are 7 aspects on the dataset: accuracy, quality, service, communication, packaging, delivery time, dan price. The average of aspect occurrences in a sentence is 0.98. Each sentiment aspect can be classified as positive or negative. The statistics of the datasets are shown in Table I.

B. Experiments

We used different settings for different models. For aspect detection model, the first approach used Adagrad optimizer with learning rate 0.01. We applied weight constraint in dense layer and dropout in recurrent layer. The second approach used Adam optimizer with learning rate 0.00025 and dropout rate 0.5 is applied. We did early stopping to both approaches to prevent overfitting.

Before performing the whole experiment, we compare two baselines of aspect model. One should be chosen so two baselines of sentiment classification model could also be compared fairly. Evaluation on aspect detection result is done by comparing average f1-score value which is shown at Table II. The fully-connected baseline is inferior to GRU baseline because GRU baseline has advantages on some words which only occur at test dataset by using word embedding.

TABLE II. F1-score for each baseline of aspect model

Model	Precision	Recall	F-score
GRU	**0.8936**	**0.8775**	**0.8855**
Fully-Connected	0.8367	0.8693	0.8527

For sentiment classification model, the first approach has some features as stated before. We did some experiments by using combination of features as shown in Table III. The highest result is obtained with the combination of word embeddings (WE), sentiment lexicons (SL), and POS tags (POS). It can perform well for predicting both positive and negative labels.

TABLE III. Average weighted F1-score with different features

Feature	Pos	Neg	Avg
WE	0.8843	0.6475	0.8547
WE+SL	0.8833	0.6543	0.8545
WE+POS	0.8850	0.6505	0.8555
WE+SL+POS	**0.8852**	**0.6575**	**0.8563**

The second approach use setup from Kim [15]. Convolution layer has multiple filter window size of 3, 4, 5 with 100 feature maps each. Dropout rate 0.5 is applied after each filter layer and after those filters result are concatenated. Optimal number of epochs is chosen by 5-fold cross validation. The best one is decided by highest mean.

For knowing the impact of using aspect matrix generated by aspect model, we would use word embedding to generate similar matrix using cosine similarity function. We use 500-dimensional word embedding which is pre-trained on 27 million sentences of Indonesian news data. For each word, similarity function would be computed relative to aspect's name. Evaluation done to all aspects in test dataset result which is the output of the better aspect model. Table IV shows that using general word embedding does not benefit as much as using the weights on aspect model because aspect model could distinct words related to aspect better.

TABLE IV. Average weighted F1-s_core with different choice of aspect matrix.

Aspect Matrix	Pos	Neg	Avg
Aspect Model	**0.8803**	**0.6366**	**0.8535**
General Word Embedding	0.8540	0.4502	0.8194

TABLE V. Evaluation comparison of different methods

Aspect	Model	Pos	Neg	Avg
Accuracy	Fachrina	0.9142	0.5333	0.8601
	GRU-based	0.9260	0.6474	0.8965
	CNN-based	**0.9303**	**0.6716**	**0.9044**
Price	Fachrina	**0.9114**	**0.4000**	**0.8927**
	GRU-based	0.8571	**0.4000**	0.8520
	CNN-based	0.8547	0.2857	0.8435
Communication	Fachrina	**0.9617**	0.6279	0.9209
	GRU-based	0.9523	**0.6567**	**0.9326**
	CNN-based	0.9438`	0.5965	0.9285
Quality	Fachrina	0.8261	0.5980	0.7895
	GRU-based	0.8701	0.6484	0.8406
	CNN-based	**0.8744**	**0.6617**	**0.8512**
Service	Fachrina	**0.8539**	**0.6042**	**0.8065**
	GRU-based	0.7634	0.5135	0.7348
	CNN-based	0.7586	0.4928	0.7305
Packaging	Fachrina	**0.9533**	**0.6071**	**0.9170**
	GRU-based	0.8468	0.5633	0.8133
	CNN-based	0.8299	0.5507	0.7989
Delivery time	Fachrina	0.8446	0.7483	0.8175
	GRU-based	**0.8954**	**0.7589**	**0.8696**
	CNN-based	0.8680	0.6838	0.8417

We choose the best model for each evaluation shown in Table III and Table IV. Those models are compared to Fachrina and achieved competitive results as given in Table V. We include aspect column because Fachrina and Widyantoro [7] approach has one model for each aspect. 'Packaging' and 'Price' aspects use rule-based approach, while the rest of the aspects use SVM approach.

V. EVALUATION ANALYSIS

We perform manual error analysis of our models to get better understanding of the misclassified data. In sentence *"Barang sudah diterima dan sudah uji quality semua berfungsi dengan baik, hanya beda sama yang digambar"* ("The product has already been received and the quality has been tested, all functionality is good, just different from its picture"), the label of 'accuracy' aspect and 'quality' aspect are negative and positive respectively. The GRU sentiment classification model predicts it correctly but CNN-based gets it wrong on 'accuracy' aspect. The result on 'accuracy' and 'quality' in CNN-based is 0.509 and 0.83 respectively which shows that 'accuracy' value is a small point away to be predicted as negative. In this case, it indicates the model prediction is not completely wrong.

In other sentence *"Barang oke, kecepatan pengiriman dan input resi ditingkatkan yang akan datang"* ("The product is okay, delivery time and process of confirming the purchase should be improved next time") the CNN-based model fails to classify the right sentiment by predicting positive for all sentiment aspects. The 'service' and 'delivery time' aspects should be negative in this case. This case is caused by the word 'okay' which has moderately high probability in all aspects. The phrase, which signal the sentiment is negative, is also quite long (having more than 5 words) so CNN-based has difficulty on processing it.

There are some sentences which none of our models get these sentences right. For instance, *"Pelayanan ya sangat buruk ga ada respon sama sekali baik pertanyaan ataupun keluhan"* ("Bad service, no response for both questions or complaints"), *"sip dipertahankan yang akan datang kecepatan pengirimannya ya jangan sampai ada yang ketinggalan ketinggalan lagi paketnya ..."* ("Yep, delivery time should be keep this way in the future, do not allow any more package left behind..."). The first example is actually quite easy for us to understand, the label should be positive in 'communication' but the model gets it wrong because word 'response' is commonly associated with fast or slow in our training data. A phrase like 'there is no response' is very rare. The second sentence is a little bit tricky, the reviewer gave a good opinion but present evidence of bad act (there is "package left behind", which should be considered as a negative sentiment).

VI. CONCLUSION

In this paper, we propose two different approaches for solving aspect-based sentiment analysis. The first one uses state-of-the-art of text classification using deep neural network for both modules of aspect-based sentiment analysis: aspect detection and sentiment classification. The second one employs aspect matrix to rescale the word vector of input sentence which aspect matrix is resulted by using dense layer for bag of words input layer. Both approaches obtain competitive result compared to previous research on Indonesian aspect based sentiment analysis using SVM and rule based methods [7].

From our experiments, we derive the following conclusions. Aspect detection with GRU layers performs better than fully-connected layer. Additional output from aspect detection is a matrix which denote probability of each word given an aspect. Using aspect model to generate aspect matrix gives a better performance compared to a general word embedding approach. The aspect matrix could distinct same sentence with different target aspect well according to some words which we have been analyzed. Although CNN-based sentiment classification could benefit from aspect matrix, GRU-based is still better by using derived features such as sentiment lexicons and POS tags.

Based on our evaluation, our model achieves state-of-the-art performance on 4 out of 7 aspects. Compared to Fachrina and Widyantoro [7], which use different model for each aspect, our model could generalize well over all aspects within the dataset.

ACKNOWLEDGMENT

This work was supported by Prosa.ai. We would also like to thank Fachrina for providing the annotated dataset regarding to this work.

REFERENCES

[1] Pontiki M. et al. SemEval-2014 Task 4: Aspect Based Sentiment Analysis, Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014) pp. 27 - 35, Dublin, Ireland, August 23-24, 2014.

[2] Pontiki M. et al. SemEval-2015 Task 12: Aspect Based Sentiment Analysis, Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015) pp. 719 - 724, Denver, Colorado, June 4-5, 2015.

[3] Pontiki M. et al. Semeval-2016 task 5: Aspect based sentiment analysis, Proceedings of SemEval 2016, San Diego, California, 2016.

[4] Hochreiter, S. and Schmidhuber, J. Long Short-Term Memory. Neural Computation, 9(8):1735–1780, 1997.

[5] Cho, K et al. "Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation". In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1724{1734. Association for Computational Linguistics, Doha, Qatar (Oct 2014)

[6] Gojali S. and Khodra M. L., "Aspect based sentiment analysis for review rating prediction," in Advanced Informatics: Concepts, Theory and Application (ICAICTA), 2016.

[7] Fachrina, Z. and Widyantoro, D. H., "Aspect-Sentiment Classification in Opinion Mining using the Combination of Rule-Based and Machine Learning". *International Conference on Data and Software Engineering (ICoDSE)*, Palembang, 2017, pp. 1-6.

[8] Ekawati D. and Khodra M. L., "Aspect-Based Sentiment Analysis for Indonesian Restaurant Reviews," in Advanced Informatics: Concepts, Theory and Application (ICAICTA), 2017.

[9] Cahyadi A. and Khodra M. L., "Aspect-Based Sentiment Analysis Using Convolutional Neural Network and Bidirectional Long Short-Term Memory" in Advanced Informatics: Concepts, Theory and Application (ICAICTA), 2018.

[10] Liu, B. "Sentiment analysis and subjectivity". In Handbook of Natural Language Processing, Second Edition, N. Indurkhya and F.J. Damerau, Editors. 2010.

[11] Tang, D., Qin, B., Feng, X., and Liu, T. "Target-Dependent Sentiment Classification with Long-Short Term Memory". CoRR, abs/1512.01100, 2015.

[12] Wang, Y., Huang, M., Zhu, X., and Zhao, L. "Attention-Based Lstm for Aspect-Level Sentiment Classification". In EMNLP, 606–615, 2016.

[13] Do, B. "Aspect-Based Sentiment Analysis Using Bitmask Bidirectional Long Short-Term Memory Networks". Florida Artificial Intelligence Research Society Conference, North America, May. 2018.

[14] Jebbara, S., and Cimiano, P. "Aspect-Based Sentiment Analysis Using a Two-Step Neural Network Architecture". In: Sack et al. [26], pp. 153–167

[15] Kim Y. "Convolutional Neural Networks for Sentence Classification". Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) pp. 17461751, Doha, Qatar, October 25-29, 2014.

[16] Wang B. and Liu M., "Deep Learning for Aspect-Based Sentiment Analysis", Stanford University report, 2015.

[17] Xue W. and Li T. "Aspect Based Sentiment Analysis with Gated Convolutional Networks", May 18, 2018.

[18] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. Accepted to NIPS 2013.

[19] Wahid, D. H., and Azhari, S. N. "Peringkasan Sentimen Ekstraktif di Twitter Menggunakan Hybrid TF-IDF dan Cosine Similarity". IJCCS (Indonesian Journal of Computing and Cybernetics Systems), 10(2), 207-218, 2016.

[20] Liu, Bing, Hu, Minqing, and Cheng, Junsheng. "Opinion Observer: Analyzing and Comparing Opinions on the Web." Proceedings of the 14th International World Wide Web Conference (WWW-2005), May 10-14, Chiba, Japan, 2005.

[21] Purwarianti, A., Gunarso, Suhardijanto, T. F. "INACL POS Tagging Convention Konvensi Pelabelan Kelas Kata INACL/MALKIN," 2017.