# Aspect-based sentiment analysis for hotel reviews using an improved model of long short-term memory

Rahmat Jayanto [a,1,*], Retno Kusumaningrum [a,2], Adi Wibowo [a,3]

[a] Department of Informatics Universitas Diponegoro, Jl. Prof. Soedarto SH Tembalang Semarang 50275, Indonesia
[1] rahmatjayanto@students.undip.ac.id; [2] retno@live.undip.ac.id ; [3] bowo.adi@live.undip.ac.id
* corresponding author

ARTICLE INFO

**Article history**
Selected paper from The 2020 3rd International Symposium on Advanced Intelligent Informatics (SAIN'20), Nanjing, China (Virtually), November 25-26, 2020, http://sain.ijain.org/2020/. Peer-reviewed by SAIN'20 Scientific Committee and Editorial Team of IJAIN journal

Received October 30, 2020
Revised July 17, 2021
Accepted November 30, 2022
Available online November 30, 2022

**Keywords**
Aspect based sentiment analysis
Deep neural network
Long short-term memory
Hotel review
Indonesian language

ABSTRACT

Advances in information technology have given rise to online hotel reservation options. The user review feature is an important factor during the online booking of hotels. Generally, most online hotel booking service providers provide review and rating features for assessing hotels. However, not all service providers provide rating features or recap reviews for every aspect of the hotel services offered. Therefore, we propose a method to summarise reviews based on multiple aspects, including food, room, service, and location. This method uses long short-term memory (LSTM), together with hidden layers and automation of the optimal number of hidden neurons. The F1-measure value of 75.28% for the best model was based on the fact that (i) the size of the first hidden layer is 1,200 neurons with the tanh activation function, and (ii) the size of the second hidden layer is 600 neurons with the ReLU activation function. The proposed model outperforms the baseline model (also known as standard LSTM) by 10.16%. It is anticipated that the model developed through this study can be accessed by users of online hotel booking services to acquire a review recap on more specific aspects of services offered by hotels.



## 1. Introduction

Hotel quality is one aspect considered by users when booking a hotel. Information technology developments have made it easier for users to ascertain the quality of a hotel. Almost every online reservation service comes with a review section. Users refer to the review section to appraise the hotel quality. While some online hotel reservation services provide hotel-rating features, others do not. However, although the quality of a hotel may change from time to time, the review data is left unaltered. Implementing sentiment analysis is one strategy that can be used to transform the collected review data into more useful information as a summary of the quality of a hotel.

The implementation of sentiment analysis consists of three granularity levels from the most general level to the most detailed level, namely the document, sentence, and aspect levels. The downside of document-level sentiment analysis is that a guest review is forced into the dominant positive or negative sentiment, even when it has elements of both polarities. In addition, at the document level, it is also not known which aspects or features generally have positive or negative reviews. Sentence-based sentiment analysis overcomes the weakness of single-polarity detection for a review, but it still does not solve the problem of identifying which aspects or features have positive and negative reviews. However, these weaknesses can be overcome using aspect-based sentiment analysis (ABSA). The implementation of ABSA can extract the reviewed aspects in a document, and each of these aspects can be associated with its sentiment polarity, whether positive, negative, or neutral.

ABSA is a feeling-based approach that considers the entity type and the aspect [1]. ABSA involves aspect term extraction, category detection, and category polarity [2]. Several studies on the application of ABSA in Indonesian-language reviews, both open-domain documents and domain-dependent documents, have been developed. One study that focused on aspect extraction and opinion terms used the coupled multilayer attentions (CMLA) mechanism and double embeddings [3], whereas another used the Bidirectional Encoder Representations from Transformers (BERT) transfer learning mechanism [4]. The first study adapted CMLA, which was proposed by [5] and double-word embeddings using fastText proposed by [6]. Subsequently, the testing process involved combinations of these methods with several RNN architectures such as gated recurrent units (GRU), bi-directional GRU (Bi-GRU), long short-term memory (LSTM) and bi-directional LSTM (Bi-LSTM). The best combination was observed with Bi-LSTM. The model using Bi-LSTM achieved an F1-measure of 0.918 and 0.9 for the term aspect and opinion term extraction, respectively [3]. The second study implemented BERT-based transfer learning proposed by [7], and it attained F1-measures of 0.87 and 0.89 for the aspect term extraction and opinion term extraction, respectively [4].

In addition, various ABSA studies for the Indonesian language have been developed related to reviews of clothing distro [8], restaurants [1], [9]–[11], and marketplaces [12], [13]. The study of ABSA for clothing distro reviews was conducted by implementing the Naive Bayes classifier and bag-of-words as the feature extraction method. This effort attained 89.86% and 97.24% for recall and precision, respectively [8]. Subsequently, the ABSA study for restaurant reviews proposed by [9] examined ABSA regarding aspect term and opinion term extraction. This study used the conditional random field (CRF) classifier [14] to predict the aspect term and opinion term. The proposed model achieved an F1-measure of 0.794. The third ABSA study concerned aspect term extraction, which was proposed by [1]. They focused on identifying the best combination of feature extraction methods for ABSA specifically for aspect term extraction. The combination of continuous bag-of-words (CBOW) and global vectors (GloVe) was found to be the best for feature extraction. This combination achieved an F1-measure of 0.642 for sentiment polarity. The study by [10] investigated aspects of term detection and orientation. They used combinations of lexicon and semantic orientation to identify the aspect terms of the review. This model attained F1-measures of 0.8840 and 0.7576 for term extraction and aspect orientation aspects, respectively. A study about the combination of classical machine learning methods and the appropriate feature extraction method for ABSA was proposed by [12]. They achieved an F1-measure of 0.92 using the support vector machine (SVM) model. Furthermore, a study by [11] investigated ABSA with respect to aspect category classification, opinion target extraction, and sentiment polarity extraction. They used a CNN model for aspect category and sentiment polarity classification. CRF and Bi-LSTM was used for opinion target extraction. The model achieved an F1-measure of 0.87, 0.78, and 0.764 for aspect category classification, opinion target extraction, and sentiment polarity extraction, respectively. The study of ABSA with respect to aspect term detection and sentiment classification using the combination of Bi-GRU and GRU was proposed by [13]. The best model identified through this study achieved an F1-measure of 0.9326.

ABSA studies conducted outside Indonesia include those conducted by Tang [15], Al-Smadi [16], Akhtar [17], Alqaryouti [18], and Liu [19], [20]. Tang's ABSA study focused on fine-grained data using datasets from Amazon and Yelp [15]. He examined a joint aspect-based sentiment topic (JABST) model, which is a combination of topic modelling and ABSA [16], in addition to a MaxEnt-JABST model, which is a combination of max entropy and JABST. In this study, the performance of the MaxEnt-JABST model was better than the JABST as the baseline model, with an increase in accuracy level of 5%. Al-Smadi conducted ABSA research on Arabic hotel reviews, focusing on sentence-level granularity [16]. The SVM and RNN methods were investigated in this study. SVM is known for its capacity to perform binary classification, whereas RNNs manage sequential data better than CNNs. The results from this investigation show that the performance of SVM is superior to the other tested models. However, when it comes to training and testing, the performance of the RNN proved to be superior to that of SVM. Akhtar conducted ABSA research and topic modelling on the hotel reviews data from TripAdvisor [17]. ABSA was employed to identify the aspects contained in the review, and subsequently, topic

modelling was conducted based on these aspects. Alqaryouti's investigation focused on government application review data [18]. This undertaking delved into a combination of integrated lexicon and rule-based ABSA. In the context of quality, the results revealed the superiority of integrated lexicon and rule-based ABSA over other lexicon baselines and rule-based methods. Integrated lexicon also comes with the capacity to manage sentiment analysis issues, such as implicit and explicit aspects, as well as negation. Liu conducted ABSA research to find a new algorithm or method to outperform existing state-of-the-art methods. In [19], Liu focused on developing a recurrent memory neural network (ReMemNN). This method was an improvement over MemNN [21], which consists of embedding adjustment learning, multi-element attention modules, and explicit memory modules. Multi-element attention modules were an improvement over binary attention that was used in [21]. This research used cross entropy as the objective function and accuracy as the performance metre. Performance of this proposed method outperformed state-of-the-art methods in almost all datasets. Liu also conducted other ABSA research [20] to find a method to overcome the weaknesses of RNNs, CNNs, attention methods, and memory networks. This research proposed a gated alternate neural network (GANN). GANN consists of convolution, max-pooling, gate truncation RNN (GTR), and fully connected layers. Convolution and max-pooling are used to divide sentences into sentiment clues and capture local features to overcome the RNN weakness of long-term dependency. GTR can capture denoising informative aspect-dependent sentence clue representation and is used to overcome CNN weaknesses. This method also used a concept of a memory network that viewed sentences not as facts and viewed aspect as a query. This research [20] used four datasets in Chinese and three datasets in English, and the proposed method outperformed state-of-the-art methods in ABSA.

The rest of this paper is structured as follows: section 2 describes the research methodology, section 3 analyses and discusses the research results, and section 4 presents the research conclusions.

## 2. Method

The stages of this investigation are displayed in Fig. 1. The process begins with the collection of data from the Traveloka website. This data is subjected to pre-processing in preparation for the training of the word2vec model, which is used for feature extraction. The feature extraction results are then used for the embedding of data at the model training stage.

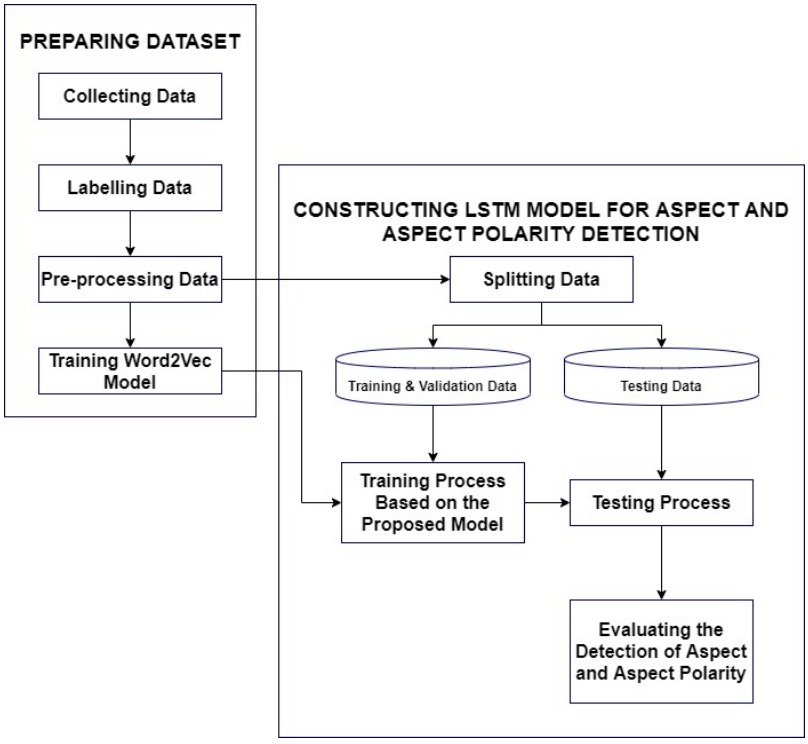


**Fig. 1.** Overview of the Research Methodology

The data are separated into training data and validation data. The purpose of the training stage is to train the models for each combination of parameters. The results from this stage are scrutinised at the model-testing stage to determine the performance of each constructed model. The model-testing phase uses data that are distinct from the data used in the training stage. This data are also obtained from the Traveloka website. The performance of each constructed model is assessed during the testing phase. A confusion matrix is used to evaluate the output of the testing phase with the micro-average F1-measure as the performance metric.

### 2.1. Collecting, Labelling, and Splitting Data

Selenium was used to acquire data by crawling reviews on the Traveloka website. Ten randomly selected hotels were the source of 2,700 data elements acquired by crawling. In addition to this data, we included 2,500 data elements used in the study conducted by [22], bringing the total dataset collected to as many as 5,200 hotel reviews as the research dataset. We split the dataset into two categories, the training-validation dataset and the testing dataset. Training-validation data consisted of 5,000 reviews with of 10,283 aspects. The data were labelled according to the aspects and sentiment polarity of these aspects identified in the review. The aspect labels used in this study were "makanan" (food), "kamar" (room), "layanan" (service), "lokasi" (location), and "lainnya" (miscellaneous), and the sentiment aspect polarities used were positive, neutral, and negative. Each review was required to contain at least one aspect.

Moreover, the testing data consisted of 200 reviews, which were also labelled as training-validation data. Table 1 exhibits the data distribution for each aspect. The one-hot encoding method was employed to label the data. Each data label holds 15 binary elements, which represent combinations of five aspects with three sentiment polarities for each aspect.

**Table 1.** Data distribution for each aspect

| Aspect | Count |
|---|---|
| Food | 1,490 |
| Room | 3,184 |
| Service | 2,169 |
| Location | 1,487 |
| Miscellaneous | 1,953 |

### 2.2. Pre-processing

Pre-processing is concerned with cleaning and preparing data for classification [23]. It involves case folding, stop-word removal, stemming, tokenisation, padding, and vectorisation. The case folding process renders all the characters in the data similar in kind, whether in lower case or upper case [24]. Stop-word removal improves machine learning performance through the elimination of conjunctions [25]. Stemming is the process of reducing words into uniform basic stems [25]. Tokenisation implies converting sentences formed from stemming into smaller parts represented by words referred to as tokens [25]. Padding involves levelling the tokenised data length by insertion of dummy tokens behind the original data. Vectorisation renders the data numeric through references to the constructed dictionary. The construction of the dictionary is based on all the words in the data.

### 2.3. Training the Word2Vec Model

This study used word2vec, a form of feature extraction method introduced by [26]. It can train a substantial quantity of data in a relatively short time [26]. The word2vec parameters used in this study were ascertained through the skip-gram model and the employment of hierarchical SoftMax as the

evaluation method [22]. The vector size of this study was 300, as the output vector of the word2vec is anticipated to increase in tandem with the significance of the dataset used [22].

### 2.4. Training and Testing Process for Aspect and Aspect Polarity Detection

In this stage, a model based on LSTM was constructed to detect aspect and aspect polarity. In this study, those two tasks were within the scope that can be handled through ABSA. In other words, this research could detect not only the polarity of aspects in a review, but also various aspects discussed in a review. In general, the implementation of ABSA follows the LSTM-based architectural standards, as illustrated in Fig. 2.

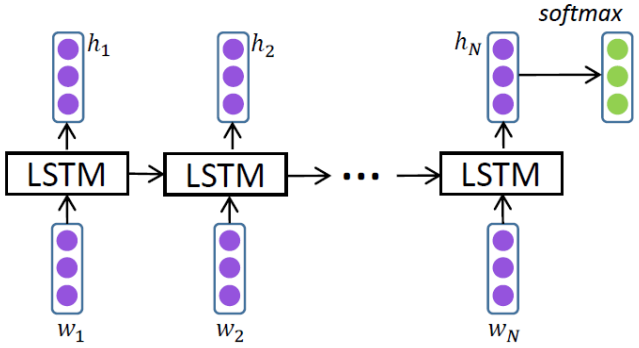


**Fig. 2.** Architecture of a Standard LSTM

LSTM was first introduced by [27] to manage the long-term dependency problem associated with RNNs. The use of LSTM eliminated this problem through the addition of gates and a memory cell. The gates used in LSTM are the forget gate, input gate, and candidate gate. Each gate has its equations. Equations (1) and (2) are the equations for the input gate and memory state, respectively.

$$\tilde{C}_t = tanh(W_c x(t) + U_c h(t-1) + b_c) \quad (1)$$

$$i_t = \sigma(W_i x(t) + U_i h(t-1) + b_i \quad (2)$$

Equation (3) is for the forget gate. Equation (4) is for the cell state, and equation (5) is for the candidate gate.

$$f_t = \sigma(W_f x(t) + U_f h(t-1) + b_f \quad (3)$$

$$C_t = i_t * \tilde{C}_t + f_t * C_{t-1} \quad (4)$$

$$o_t = \sigma(W_o x(t) + U_o h(t-1) + b_o \quad (5)$$

Each cell in LSTM has one output from the candidate gate and one hidden state. Equation (6) is used to calculate the result of the hidden state output.

$$h_t = o_t * \tanh(C_t) \quad (6)$$

This study used the LSTM model architecture, together with fully connected layers, as depicted in Fig. 3. The input from the model, generated at the pre-processing stage, is put through the embedding layer using the matrix embedding weights. Following the insertion of the input data into the embedding layer, the data are routed into the LSTM layer. The embedding matrix is constructed from the pair of words and its word2vec vector. The LSTM layer produces a two-dimensional matrix of the same size as the input embedding matrix. Before entering the fully connected layer, the LSTM output must be changed to one dimension using the flatten layer followed by two fully connected layers with size and activation functions as tested in this study. As shown in Fig. 3, this study uses standard LSTM architecture in [28] as the baseline model. The final layer of the baseline model is adjusted to correspond

to the size of the data label. During the training stage, we considered several combinations of parameters to determine the best model. Table 2 shows the combinations of parameters considered.

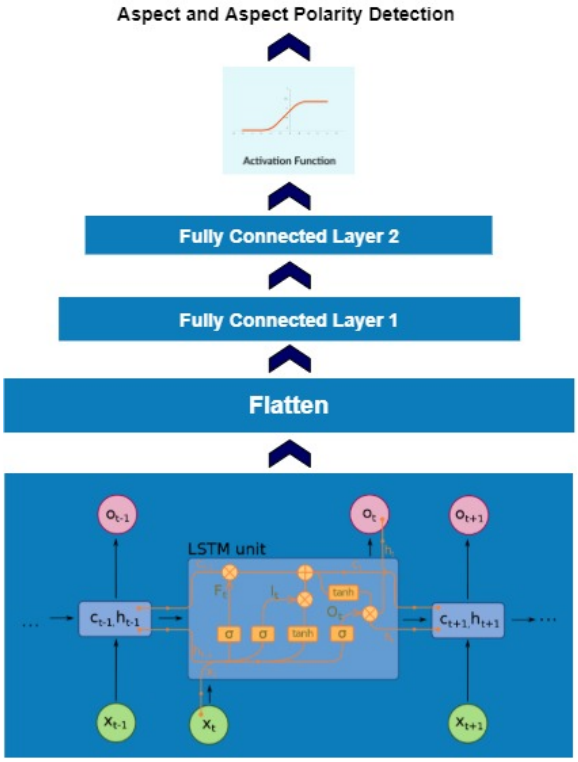


**Fig. 3.** Proposed Architecture

The testing phase involves a process designed to identify the best combination of parameters for the proposed architecture model. The data used during the testing stage are different from the data used during the training stage. The most useful model is the one with the best combination of parameter values. This study considered four research scenarios to find the best values for the parameters. The running of each scenario is based on the combination of parameters shown in Table 2. The first scenario determines the best size in the first fully connected layer, the second scenario determines the best activation function in the first fully connected layer, the third scenario determines the best size in the second fully connected layer, and the fourth scenario determines the best activation function in the second fully connected layer.

**Table 2.** Parameter combination for the proposed architecture

| Parameter | Value |
|---|---|
| Size of fully connected layer 1 | 700, 750, 800, 850, 900, 950, 1000, 1050, 1100, 1150, 1200 |
| Size of fully connected layer 2 | 50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550, 600, 650, 700 |
| Activation function | ReLU, tanh, sigmoid |

### 2.5. Evaluation

For evaluation, we opted for the micro-average F1-measure because it is highly responsive to the most common classes or labels [29]. The confusion matrix plays a supporting role during the micro-average F1-measure calculation. An example of the confusion matrix is provided in Table 3. In an ideal confusion matrix, the main diagonal entry value is more significant than the other entries. That confusion matrix produces an F1-measure value close to 1.

**Table 3.** Example of a confusion matrix in this research

| | | Prediction | | | |
|---|---|---|---|---|---|
| | | *Positive* | *Neutral* | *Negative* | *None* |
| Actual | Positive | 108 | 4 | 8 | 91 |
| | Neutral | 4 | 7 | 5 | 12 |
| | Negative | 9 | 5 | 80 | 67 |
| | None | 34 | 2 | 28 | 531 |

The F1-measure formula is expressed in Equations (7), (8), and (9).

$$Precision_{mic} = \frac{\sum_{i=1}^{N} TP_1}{\sum_{i=1}^{N}(TP_1 + FP_1)} \quad (7)$$

$$Recall_{mic} = \frac{\sum_{i=1}^{N} TP_1}{\sum_{i=1}^{N}(TP_1 + FP_1)} \quad (8)$$

$$F1_{micro\ average} \frac{2 \times Precision_{mic} \times Recall_{mic}}{Precision_{mic} + Recall_{mic}} \quad (9)$$

In these equations, TP represents true positive, FP represents false positive, and FN represents false negative. TP for each sentiment is the main diagonal. FP is located one row below, while FN is in one horizontal column. Based on Equations (7) and (8), if the value of the denominator is similar for precision and recall, it follows that the value of the micro-averages of precision and recall would also be similar. This leads to the materialisation of Equation (11), which contends that the value of the micro-average F1-measure is equal to that of precision and recall.

$$Precision_{micro\ average} = Recall_{micro\ average} \quad (10)$$

$$F1_{micro\ average} = Precision_{micro\ average} \quad (11)$$

## 3. Results and Discussion

As mentioned before, we used 200 reviews as testing data, and six scenarios were observed in this study. In the first testing scenario, the performance of the model was grouped based on the size of the fully connected layer 1. The combination of parameters used for the fully connected layer 1 can be observed in Table 2. The results of this scenario are exhibited in Fig. 4.

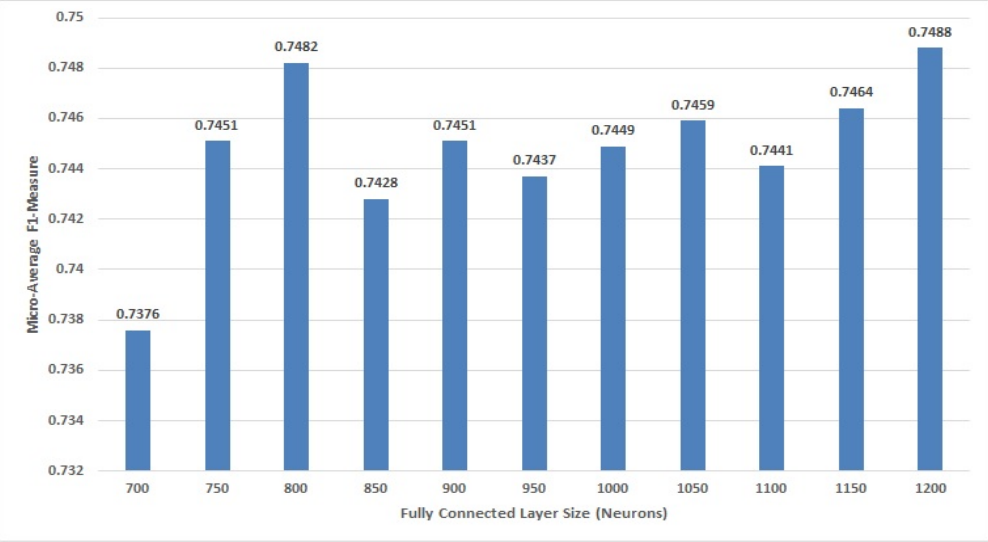


**Fig. 4.** Impact of Size of Fully Connected Layer 1 on Micro-Average F1-Measure

Fig. 4 shows no statistically significant link between the size of the fully connected layer and the performance of the model. It can be observed that there is an upward movement from 700 to 1,200, though there are some anomalies in the middle of this upward movement. These anomalies stem from the random initialisation of weights during each iteration. As illustrated in Fig. 4, the model with 1,200 neurons as the size of fully connected layer 1 is the best model. It has a micro-average F1-measure of 0.7488. The model achieved this value because a smaller number of neurons in the hidden layer led to a higher probability of providing incorrect information to the next layer [30] and vice versa.

In the second scenario, the performance of the model was grouped based on the activation function used on the fully connected layer 1. The activation functions investigated were tanh, ReLU, and sigmoid, which were applied for each combination of parameters. The performance evaluation of this scenario was based on the mean of the micro F1-measure. The results of this scenario can be observed in Fig. 5.

The results showed that the tanh activation function, with a micro-average F1-measure of 0.7462, delivered the best performance. ReLU, with a micro-average F1-measure of 0.7419, had the worst performance. The performance of the sigmoid function was close to that of tanh; both functions have an upper limit value of 1, which facilitates the convergence of the results to the desired range of 0 to 1 [31]. By using tanh, the results did not converge too quickly, and various data patterns could be recognised.

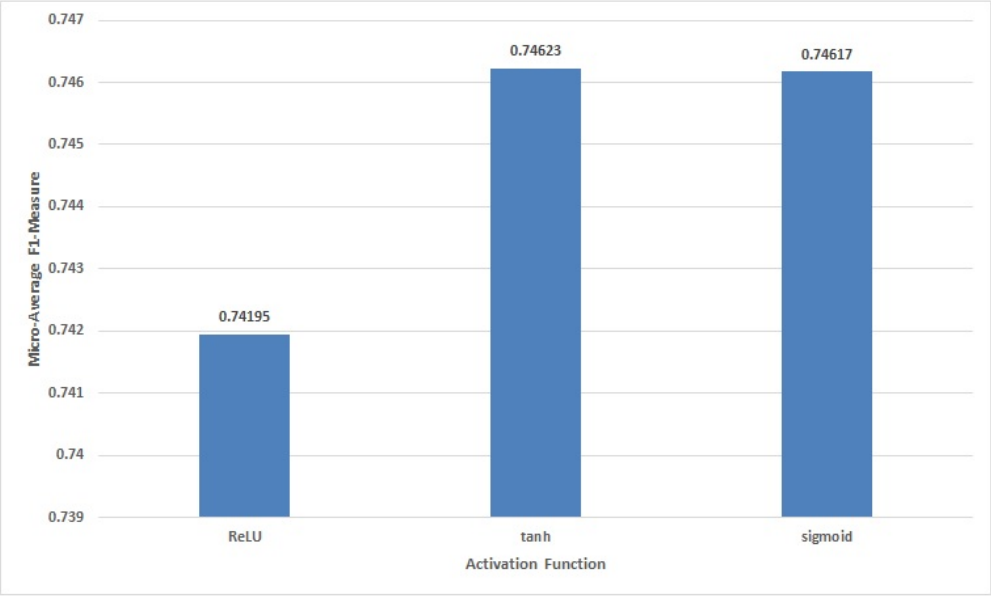


**Fig. 5.** Impact of Activation Function of Fully Connected Layer 1 on Micro-Average F1-Measure

In the third research scenario, the performance value of the model was grouped based on the size of fully connected layer 2. Table 2 portrays the size of fully connected layer 2. The results for this scenario were calculated in the same manner as in the other scenarios, i.e., using the micro-average F1-measure for each group. The performance models in this scenario can be observed in Fig. 6.

According to Fig. 6, there is an upward pattern in the model's performance until the size of 400 neurons, beyond which the performance level remained stable. This result is in contrast with the first scenario, in which no specific pattern was discerned. Several anomalies were apparent in the performance of the model. These can be connected to the random initialisation of weights for each iteration. While the upward pattern for this scenario increased until 400 neurons, the result indicated 600 neurons to be the best size for fully connected layer 2. A size of 600 neurons registers a micro-average F1-measure value of 0.7590, which is superior to the micro-average F1-measure of 0.7686 attributed to a size of 400 neurons. These results are in agreement with those realised by [30], which verified that the smaller the neuron size in the hidden layer, the more likely the failure to convey information to the next layer.

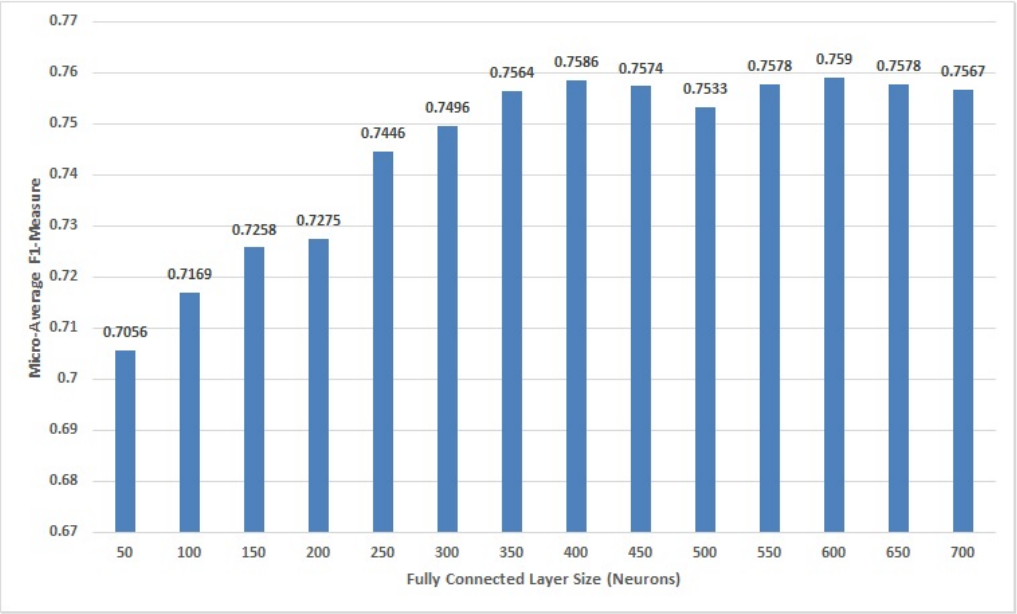


**Fig. 6.** Impact of Size of Fully Connected Layer 2 on Micro-average F1-measure

In the fourth research scenario, the performance of the model was grouped based on the activation function used at fully connected layer 2. As in the first fully connected layer, the tanh, ReLU, and sigmoid activation functions were considered. The results of this scenario are displayed in Fig. 7.

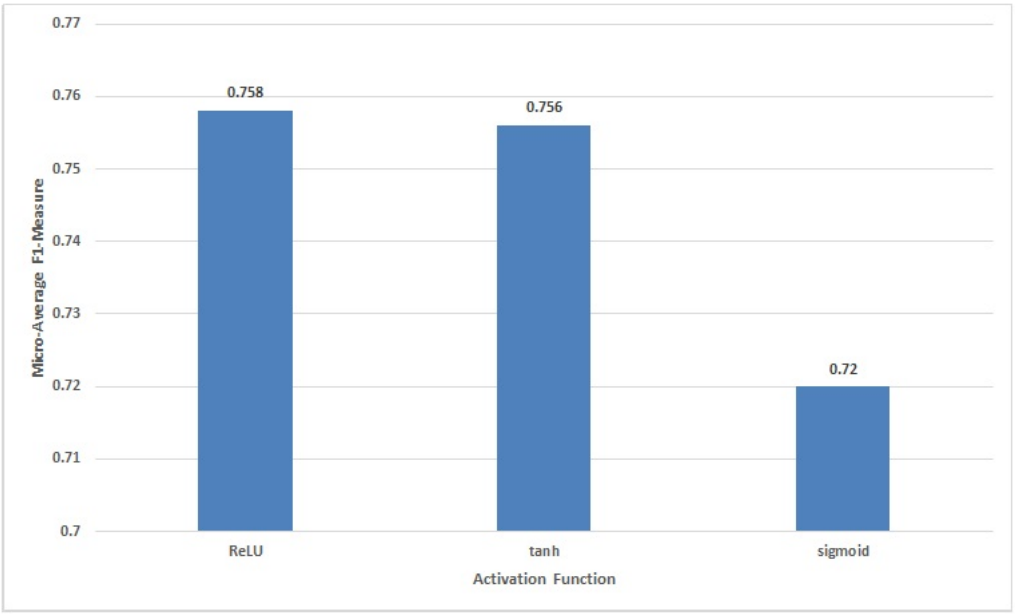


**Fig. 7.** Impact of Activation Function of Fully Connected Layer 2 on Micro-Average F1-Measure

Fig. 7 reveals that ReLU, with a micro-average F1-measure value of 0.758, is the best activation function at fully connected layer 2. The least favourable activation function is the sigmoid function, which results in a micro-average F1-measure of 0.72. The ReLU activation function in this architecture can limit the lower value of the resulting neuron to 0, which corresponds to the desired result range [31].

The best combination of parameters derived from the four scenarios is utilised for the construction of the ideal model. The best parameters are a neuron count of 1,200 for fully connected layer 1 using the tanh activation function and a neuron count of 600 for fully connected layer 2 with a neuron count

of 600 using the ReLU activation function. The constructed model achieved a micro-average F1-measure of 0.7528 when tested using the same data.

During the testing phase, several misclassifications were detected in the model derived through this process. The first is attributed to the inappropriate use of the aspect and sentiment pairs. For example, "Saat pertama masuk lobby, bau pekat rokok, tapi di kamar, tidak ada bau rokok sama sekali, bersih, luas. Beberapa jam saya turun ke lobby, bau rokok sdhbtdk ada lagi, wangi. Proses check-in, check-out cepat. We'll be back soon. Thanks Epic!" (When you first enter the lobby, there is a thick smell of cigarettes, but in the room, there was no cigarette smell at all, clean, spacious. A few hours later, I went down to the lobby, and there was no smell of cigarettes anymore, fragrant. Check-in, check-out is fast. We'll be back soon. Thanks Epic!). This review should register positive sentiments in the "kamar" (room) aspect and negative sentiments in other aspects. However, neutral sentiments were recognised in the room aspect as "bau" (smells) and "rokok" (cigarettes) in the training data are more often associated with the "kamar" (room) aspect rather than the "lainnya" (miscellaneous) aspect.

The second source of misclassification was the inappropriate use of a foreign language mixed with the Indonesian language. For example, "Staff-nya ramah banget, breakfast-nya enak" (The staff is very friendly, the breakfast is delicious). This review could ganer a positive sentiment for the "makanan" (food) and "layanan" (service) aspects. However, it was recognised as a positive sentiment for the "layanan" (service) aspect, but the model did not recognise the "makanan" (food) aspect. The inappropriate integration of a foreign language with the Indonesian language diminished the quality of the pre-processing of data, particularly when it came to stemming and stop-word removal. In the example above, the word "breakfast-nya" cannot be appropriately recognised, as this word cannot be stemmed in Indonesian.

The third source of misclassification was the size of the training dataset, which was relatively small. Due to the lack of training data, the model only recognised limited patterns. The limited training data not only limited pattern recognition but also caused an imbalance in data for each aspect, thereby affecting the model's performance in recognising patterns for certain aspects. Hence, the model was good at recognising one aspect but not the other aspects.

In this study, we also compared our proposed model's results to those of the baseline LSTM architecture. Both architectures were trained and tested with the same 5,000 training and 200 testing data elements. According to the test results exhibited in Table 4, the performance of our proposed model surpassed that of the baseline model. The most significant difference came from the proposed model's two additional fully connected layers before the output layer. The sigmoid activation function was used in the output layer. As demonstrated in Table 2, the installation of additional fully connected layers enhanced the model's capacity for data pattern identification.

**Table 4.** F1-measure comparison between baseline and proposed model

| Model | Micro-average F1-measure |
|---|---|
| Baseline Model | 0.6512 |
| Proposed Model | 0.7528 |

## 4. Conclusion

A system for recognising hotel quality based on reviews can be established by obtaining review data, pre-processing the data, using a model for recognising the data, then obtaining the aspects and sentiments of the reviews. Aspect and sentiment data can be processed to be more understandable and appealing to customers. The best model architecture was realised through a combination of LSTM and two fully connected layers (fully connected layer 1 with 1,200 neurons using the tanh activation function and fully connected layer 2 with 600 neurons using the ReLU activation function). The proposed model, with a micro-average F1-measure of 0.7528, outperformed the baseline model by 0.1016 (10.16%) in the F1-measure.

The performance of this model can be improved using other RNNs such as GRU, Bi-GRU, Bi-LSTM, or vanilla RNN. The use of other feature extraction methods such as GloVe or FastText can enhance the model's word identification ability, and the use of a larger training set will improve the model's pattern identification ability.

## Acknowledgment

The authors thank Directorate Research and Development, Ministry of Research and Technology/National Agency for Research and Innovation, Indonesia for supporting this research.

## Declarations

**Author contribution.** All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.
**Funding statement.** This research was funded by Directorate Research and Development, Ministry of Research and Technology/National Agency for Research and Innovation, Indonesia under grant of Basic Research, fiscal year 2020, Number 257-20/UN7.6.1/PP/2020.
**Conflict of interest.** The authors declare no conflict of interest.
**Additional information.** No additional information is available for this paper.

## References

[1] D. Ekawati and M. L. Khodra, "Aspect-based sentiment analysis for Indonesian restaurant reviews," in *Proc. of 2017 International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA)*, 2017, pp. 1–6. doi: 10.1109/ICAICTA.2017.8090963.

[2] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, "SemEval-2014 Task 4: Aspect based sentiment analysis," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 2014, pp. 27–35. doi: 10.3115/v1/S14-2004.

[3] J. Fernando, M. L. Khodra, and A. A. Septiandri, "Aspect and opinion terms extraction using double embeddings and attention mechanism for Indonesian hotel reviews," 2019. doi: 10.1109/ICAICTA.2019.8904124.

[4] A. A. Septiandri and A. P. Sutiono, "Aspect and Opinion Term Extraction for Aspect Based Sentiment Analysis of Hotel Reviews Using Transfer Learning," 2019. doi : 10.48550/arXiv.1909.11879.

[5] W. Wang, S. J. Pan, D. Dahlmeier, and X. Xiao, "Coupled multi-layer attentions for co-extraction of aspect and opinion terms," in *Proc. of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017, pp. 3316–3322. doi : 10.1609/aaai.v31i1.10974.

[6] H. Xu, B. Liu, L. Shu, and P. S. Yu, "Double embeddings and CNN-based sequence labeling for aspect extraction," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2018, pp. 592–598. doi: 10.18653/v1/P18-2094.

[7] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol.*, 2019, pp. 4171–4186. doi : 10.48550/arXiv.1810.04805.

[8] C. Fiarni, H. Maharani, and R. Pratama, "Sentiment analysis system for Indonesia online retail shop review using hierarchy Naive Bayes technique," 2016. doi: 10.1109/ICoICT.2016.7571912.

[9] S. Gojali and M. L. Khodra, "Aspect Based Sentiment Analysis for Review Rating Prediction," in *2016 International Conference On Advanced Informatics: Concepts, Theory And Application*, 2016, pp. 1–6. doi: 10.1109/ICAICTA.2016.7803110.

[10] D. H. Sasmita, A. F. Wicaksono, S. Louvan, and M. Adriani, "Unsupervised aspect-based

sentiment analysis on Indonesian restaurant reviews," in *Proceeding of 2017 International Conference on Asian Language Processing (IALP)*, 2017, pp. 383–386. doi: 10.1109/IALP.2017.8300623.

[11] A. Cahyadi and M. L. Khodra, "Aspect-based sentiment analysis using convolutional neural network and bidirectional long short-term memory," in *Proceeding of 2018 5th International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA)*, 2018, pp. 124–129. doi: 10.1109/ICAICTA.2018.8541300.

[12] Z. Fachrina and D. H. Widyantoro, "Aspect-sentiment classification in opinion mining using the combination of rule-based and machine learning," 2017. doi: 10.1109/ICODSE.2017.8285850.

[13] A. Ilmania, Abdurrahman, S. Cahyawijaya, and A. Purwarianti, "Aspect detection and sentiment classification using deep neural network for Indonesian aspect-based sentiment analysis," in *Proceeding of 2018 International Conference on Asian Language Processing*, 2018, pp. 62–67. doi: 10.1109/IALP.2018.8629181.

[14] L. Qi and L. Chen, "A linear-chain CRF-based learning approach for web opinion mining," in *Proceeding of the International Conference on Web Information Systems Engineering*, 2010, pp. 128–141. doi: 10.1007/978-3-642-17616-6_13.

[15] F. Tang, L. Fu, B. Yao, and W. Xu, "Aspect based fine-grained sentiment analysis for online reviews," *Inf. Sci. (Ny).*, vol. 488, pp. 190–204, 2019, doi: 10.1016/j.ins.2019.02.064.

[16] M. Al-Smadi, O. Qawasmeh, M. Al-Ayyoub, Y. Jararweh, and B. Gupta, "Deep recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels' reviews," *J. Comput. Sci.*, vol. 27, pp. 386–393, 2018, doi: 10.1016/j.jocs.2017.11.006.

[17] N. Akhtar, N. Zubair, A. Kumar, and T. Ahmad, "Aspect based Sentiment Oriented Summarization of Hotel Reviews," *Procedia Comput. Sci.*, vol. 115, pp. 563–571, 2017, doi: 10.1016/j.procs.2017.09.115.

[18] O. Alqaryouti, N. Siyam, A. A. Monem, and K. Shaalan, "Aspect-based sentiment analysis using smart government review data," *Appl. Comput. Informatics*, 2020, doi: 10.1016/j.aci.2019.11.003.

[19] N. Liu and B. Shen, "ReMemNN: A novel memory neural network for powerful interaction in aspect-based sentiment analysis," *Neurocomputing*, vol. 395, pp. 66–67, 2020, doi: 10.1016/j.neucom.2020.02.018.

[20] N. Liu and B. Shen, "Aspect-based sentiment analysis with gated alternate neural network," *Knowledge-Based Syst.*, vol. 188, p. 105010, 2020, doi: 10.1016/j.knosys.2019.105010.

[21] D. Tang, B. Qin, and T. Liu, "Aspect level sentiment classification with deep memory network," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 214–224. doi : 10.18653/v1/D16-1021

[22] R. P. Nawangsari, R. Kusumaningrum, and A. Wibowo, "Word2vec for Indonesian sentiment analysis towards hotel reviews: An evaluation study," *Procedia Comput. Sci.*, vol. 157, pp. 360–366, 2019, doi: 10.1016/j.procs.2019.08.178.

[23] E. Haddi, X. Liu, and Y. Shi, "The role of text pre-processing in sentiment analysis," *Procedia Comput. Sci.*, vol. 17, pp. 26–32, 2013, doi: 10.1016/j.procs.2013.05.005.

[24] A. Kulkarni and S. Mundhe, "A theoretical review on text mining: Tools, techniques, applications and future challenges," *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 4, no. 11, pp. 19225–19230, 2016, doi: 10.15680/IJIRCCE.2016. 0411037.

[25] S. Kannan and V. Gurusamy, "Preprocessing Techniques for text mining," *Int. J. Comput. Sci. Commun. Networks*, vol. 5, no. 1, pp. 7–16, 2014. Available at : Google Scholar.

[26] T. Mikolov, G. Corrado, K. Chen, and J. Dean, "Efficient estimation of word representations in vector space," in *ICLR*, 2013, pp. 1–12. doi : 10.48550/arXiv.1301.3781

[27] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[28] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based LSTM for aspect-level sentiment classification," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016, pp. 606–615. doi: 10.18653/v1/d16-1058.

[29] Y. Yang and X. Liu, "A re-examination of text categorization methods," 1999. doi : 10.1145/312624.312647

[30] I. Shafi, J. Ahmad, S. I. Shah, and F. M. Kashif, "Impact of varying neurons and hidden layers in neural network architecture for a time frequency application," 2006. doi: 10.1109/INMIC.2006.358160.

[31] C. E. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation Functions: Comparison of trends in Practice and Research for Deep Learning," in *Proceeding of the 2nd International Conference on Computational Sciences and Technologies*, 2020, pp. 124–133. doi :10.48550/arXiv.1811.03378.