Heliyon 9 (2023) e17147

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Deep learning-based application for multilevel sentiment analysis of Indonesian hotel reviews

Retno Kusumaningrum *, Iffa Zainan Nisa, Rahmat Jayanto, Rizka Putri Nawangsari, Adi Wibowo

Department of Informatics, Universitas Diponegoro, Semarang, Indonesia

ARTICLE INFO

Keywords:
Sentiment analysis
Deep learning
Convolutional neural network
Long short-term memory
Sentiment visualization

ABSTRACT

Purpose: In this study, we present a web-based application that retrieves hotel review documents in Indonesian languages from an online travel agent (OTA) and analyses their sentiments from the coarse-grained document to the fine-grained aspect level.
Design: /Methodology/Approach: There are four main stages in this study: development of sentiment analysis model at the document level based on a convolutional neural network (CNN), development of sentiment analysis model at the aspect level based on an improved long short-term memory (LSTM), model deployment for multilevel sentiment analysis in a web-based application, and its performance evaluation. The developed application uses several sentiment visualizations types at coarse-grained and fine-grained levels, such as pie charts, line charts, and bar charts.
Finding: The application's functionality was demonstrated in practice based on three datasets from three OTA websites, which were analyzed and evaluated based on several matrices, namely, the precision, recall, and F1-score. The results revealed that the performance for the F1-score was 0.95 ± 0.03, 0.87 ± 0.02, and 0.92 ± 0.07 for document-level sentiment analysis, aspect-level sentiment analysis, and aspect-polarity detection, respectively.
Originality: The developed application (Sentilytics 1.0) can analyze sentiment at document and aspect levels. The two levels of sentiment analysis are based on two models generated by fine-tuning CNN and LSTM models using specific architectures and domain data (Indonesian hotel reviews).

1. Introduction

Web 2.0 is currently one of the most significant website design types due to the willingness of users to express their opinions on a product or service online via social media or blogs and to review the features of various e-commerce programs as well as applications. One such e-commerce application that hosts many users is the online travel agent (OTA), and one of the most widely used services is hotel booking. Providing review facilities can help other users determine their preferred hotels according to their requirements. In addition, the availability of reviews can be exploited by hotel management to evaluate the facilities and services provided by the hotel. However, the availability of many unsummarized reviews increases the time and resources required for users and hotel management to understand the conditions of a hotel, and manual review summarization is prone to human error [1]. The method required for the

* Corresponding author.
E-mail address: retno@live.undip.ac.id (R. Kusumaningrum).

https://doi.org/10.1016/j.heliyon.2023.e17147
Received 12 October 2022; Received in revised form 3 June 2023; Accepted 8 June 2023
Available online 9 June 2023

automation of this process is referred to as sentiment analysis.

Various studies have been conducted on Indonesian sentiment analysis in various domains [2–16]. However, the number of studies on sentiment analysis for Indonesian documents significantly exceeds the number of developed Indonesian sentiment analysis platforms. Some examples of available Indonesian language sentiment analysis platforms are Sentiment Analysis from SelindoAlpha[1] dan CX Analytics from Prosa.[2] This is not the case with respect to the developed sentiment analysis platforms for English documents, such as Brandwatch,[3] Talkwalker's Quick Search,[4] Repustate,[5] and Rossete.[6] Brandwatch is a sentiment analysis platform that can monitor mentions online, comprehend customer voices, detect fluctuations in sentiment, and measure brand visibility. Talkwalker's Quick Search is a tool for obtaining insights from all brand mentions by automatically analysing social media communications. Repustate is a sentiment analysis platform equipped with features to detect sentiments in slang and emoticons. Rossete can be used to detect in-text emotional hotspots in relation to companies, people, and products.

As previously explained, developing a sentiment analysis platform can be used as part of business intelligence, namely as a tool to capture feedback on the products and services offered [17]. In this case, the developed sentiment analysis platform can assist hotel management in improving the performance and quality of their hotels based on negative reviews and maintaining good hotel performance and quality when getting lots of positive reviews. Furthermore, it can increase hotel occupancy rates, which is expected to help increase the country's revenue from the tourism sector. In addition, although a sentiment analysis platform for English language reviews has been widely developed, it has become less valuable due to the language-dependent characteristics of sentiment analysis; hence, the platform's development will be more optimal when it can capture the specific features of the related language [18].

There are four differences between Indonesian and English related to the sentiment analysis task: language roots, the syntax for adjective phrases, plural form, and active-passive voice. Since Indonesian differs from English because of differences in language roots (i.e., Indonesian comes from Austronesian, whereas English comes from Germany), we need to develop from scratch for this sentiment analysis platform. One of the syntactical differences between Indonesian and English is in the form of adjective phrases, where these adjectives are mainly used to show the polarity of sentiments. Consider the review example "noise from the traffic was unbearable." The word "unbearable" in English is one word, while in Indonesian, it becomes two words by adding the negation word "tidak" ("not") in front of the adjective. Another difference is the plural forms of Indonesian and English. If the English plural forms are formed by adding -s or -es for regular nouns and more complex rules for irregular nouns, in Indonesian, plural forms are formed by repeating words. For example, the word "kamar-kamar" in Indonesian becomes "rooms" in English. The last difference is that passive sentences are often used in Indonesian, while English refers more to the active form, especially when the focus is on an object. In line with Korayem et al. [18], developing a sentiment analysis platform for hotel reviews is still a big challenge.

Theoretically, sentiment analysis can be divided into three levels: document, sentence, and aspect [19]. Document-level sentiment analysis assumes that each document expresses an opinion of a single entity [19]. Thus, if a document consists of opinions on various aspects with various polarities, the sentiment polarity is classified as the dominant polarity class. Sentiment analysis at the document level is simple and can be readily implemented. In addition, the implementation of sentiment analysis at the document level can be rapidly used to provide an overview of an entity's sentiment polarity or overall impression.

Sentence-level sentiment analysis is naturally a more detailed task. The common objective of applying sentence-level sentiment analysis is to improve the prediction results of the overall polarity of the document, which is generally applied based on two approaches, namely, a cascade approach and a joint approach [20]. Both approaches aim to perform a fine-to-coarse sentiment analysis or sentence-to-document sentiment analysis. The sentiment classification results at the sentence level are then used to determine the sentiment polarity of the document.

In contrast to sentence-level sentiment analysis, aspect-based sentiment analysis is an approach implemented to realise fine-grained sentiment analysis, in addition to the detection of an entity polarity or entity aspect in a review document [21]. Aspect-based sentiment analysis is highly useful, given that various aspects of a product or service are referenced by users in a review document, with each aspect being assigned a polarity [22].

Several factors influence the requirements for a sentiment analysis platform for hotel reviews. First, the users exploit the sentiment analysis platform to obtain information on the general condition of a hotel's sentiment polarity. Second, OTA users employ the platform to determine the aspects assigned high ratings. Furthermore, the platform can be used as a reference for users with particular preferences for features or aspects provided by hotels. Third, hotel management may use this platform to identify the features or aspects that require improvements due to low ratings. This study focuses on sentiment analysis for hotel reviews in the Indonesian language at multiple levels, performing both document- and aspect-level sentiment analysis.

In this study, we choose not to apply sentence-level sentiment analysis because, as previously explained, the main objective of sentiment analysis at this level is to find out the polarity of sentiment in a review or what is commonly referred to as a document. The main difference with sentiment analysis at the document level lies only in the processing steps. At the sentence level, each document is initially broken down into sentences to identify the sentiment polarity of each sentence further, and the final result is the polarity that appears the most. Therefore, this condition has been covered by document-level sentiment analysis.

[1] https://selindo.com/sentiment-analysis-bahasa-indonesia/.
[2] https://prosa.ai/solutions/customer-experience.
[3] https://www.brandwatch.com/.
[4] https://www.talkwalker.com/quick-search.
[5] https://www.repustate.com/.
[6] https://www.rosette.com/capability/sentiment-analyzer/.

Therefore, this study aims to develop a sentiment analysis application in a web-based platform that can analyze hotel reviews for both the coarse-grained document level and the fine-grained aspect level of sentiment. There are five aspects: food, room, service, location, and miscellaneous. The backend engine for sentiment analysis at the document level is based on CNN, whereas the backend engine for the aspect level is based on an improved LSTM model. Furthermore, the main contributions of our work can be summarized as follows:

- A benchmark dataset for sentiment analysis of hotel reviews in the Indonesian language
- A pre-trained CNN model for document-level sentiment analysis.
- A pre-trained LSTM model for aspect-level sentiment analysis.
- An application for multilevel sentiment analysis of Indonesian hotel reviews.

The remainder of this study is organised as follows. Section 2 presents state-of-the-art document-based sentiment analysis and aspect-based sentiment analysis. A discussion is presented on applying sentiment analysis to Indonesian documents and documents in other foreign languages. The research methodology is detailed in Section 3. Section 4 describes the results and discussion. It details the developed application called Sentilytics 1.0 and performance evaluation at document and aspect levels. The conclusions and scope of future research are detailed in Section 5.

2. Related works

Sentiment analysis involves processing textual data such as reviews, tweets, and social media posts to obtain sentiment information from documents. Sentiment analysis has been applied to Indonesian-language documents in various domains such as hotels [5,13,14,16,23,24], touristic destinations [6,9], product reviews [3,4], and political figure/presidential elections [11,12,25]. As mentioned previously, there are three sentiment analysis levels: document, sentence, and aspect. Given that the focus of this study was on document-level and aspect-level analyses, the following sub-sections only present discussions on these analysis levels.

2.1. Sentiment analysis at the document level

Sentiment analysis at the document level is the most common level at which review data are classified into positive or negative sentiments, although they contain elements of both polarities. Moraes et al. [26] compared the naïve Bayes (NB), support vector machine (SVM), and artificial neural network (ANN) methods for the document-level sentiment analysis of movies and product review datasets. The results revealed that the ANN method significantly outperformed the other methods, especially for unbalanced data, with an accuracy of 86.5%. Furthermore, Tripathy et al. [27] proposed a hybrid machine learning approach that combines SVM and ANN for the Internet Movie Database (IMDb) and polarity data. A support vector machine selected the optimal features from the training data. Each feature was then inputted to an ANN for processing. This method yielded improved results with an accuracy of 96.4%. However, this method is limited because the data used is generally small.

Several studies have been conducted on sentiment analysis of hotel reviews at the document level [13,16] and sentence level [14]. These studies applied classical machine learning to solve sentiment analysis problems. However, the performance of the models developed in these studies was generally low. Classical machine learning methods are generally accompanied by a feature engineering process such as various word representations [13,14] or the word representation is enhanced using a sentiment shifter [16].

However, the performance of classical machine learning depends on the effectiveness of the handcrafted feature engineering process. The feature engineering process has several limitations, such as (i) it is labour-intensive [28,29] and (ii) costly to obtain satisfactory accuracy, as it requires manual pre-processing; hence, it is time-consuming [30]. This is caused by the inability of shallow learning methods to extract and organize discriminative information from data [29]. These limitations can be overcome by applying deep-learning methods. Implementing a neural network with multiple hidden layers provides a high feature learning capability, which is beneficial for visualizing or classifying data [31]. Convolutional Neural Network (CNN) as a form of deep learning architecture also has the same advantage, i.e., the trained CNN-based model for text classification can recognize patterns in text automatically, such as key phrases [32]. To extract a feature vector from the input word embeddings, CNN-based models employ one-dimensional (1-D) convolution followed by a one-dimensional pooling operation (average or max).

2.2. Sentiment analysis on aspect-level

Aspect-based sentiment analysis (ABSA) is a sentiment analysis method used to identify aspect polarities related to context. This method involves two main tasks, namely, aspect extraction and sentiment classification [33]. The aspect extraction task is called aspect detection, whereas sentiment classification is called aspect-polarity detection. Aspect-based sentiment analysis has been implemented in various foreign languages, such as English [34], Arabic [35], and Burmese [36].

Manik [37] applied ABSA to predict candidate characteristics in the Indonesian presidential election. The findings revealed that the SVM algorithm outperformed the naïve Bayes classifier and k-nearest neighbours (KNN) algorithm. Subsequently, an ABSA study was conducted by Gojali and Khodra [38] for the review's analysis of the restaurant's rating using conditional random fields (CRFs) to predict aspects, and the opinion term achieved an F1-score of 79.4%. For aspect detection, Azhar et al. [39] used a combination of a CNN and extreme gradient boosting (XGBoost) to analyze sentiments in hotel reviews. The model performance was high, with an F1-score of 93.16%. However, only aspects present in the reviews were detected. In contrast, ABSA's objective is to detect an aspect's

polarity in a review.

Neural networks, such as long short-term memory (LSTM), can encode phrases without feature engineering [40]. Moreover, LSTMs can extract sequential information and selectively consider and ignore items [41]. This condition implies that increased the accuracy of ABSA tasks requires more complex architectures. Therefore, it is necessary to develop an approach that can be used to detect aspects and their sentiment polarity in a review simultaneously.

3. Methodology

This study has four main stages in this study, i.e. development of the sentiment analysis model at the document level, development of the sentiment analysis model at the aspect level, model deployment for multi-level sentiment analysis in a web-based application, and performance evaluation. The detailed explanation of those stages is explained in the following sub-sections. An illustration of the research methodology can be seen in Fig. 1.

3.1. Development of sentiment analysis model at the document level

The development process of the sentiment analysis model at the document level was based on a dataset consisting of 2500 review documents. Of these, 1250 data were labelled as positive sentiments and 1250 as negative sentiments. Furthermore, 2500 pre-processed data are divided into training and validation data based on the 10-fold cross-validation concept. Therefore, each fold consists of 250 data, where 125 are positive data samples and the other 125 are negative data samples. The hotel reviews selected as research datasets were sourced from hotels in various regions of Indonesia. Pre-processing involved the following steps: case folding, tokenization, stop-word removal, stemming, and padding.

The architecture of the word2vec model was tuned using the continuous bag of words (CBOW) and skip-gram models with hierarchical softmax and negative sampling evaluation methods, and the vector dimensions were 100, 200, and 300. Subsequently, the training of the word2vec model was performed. The results revealed that the optimal parameter for the architecture of the word2vec model was the skip-gram model, the evaluation method was the hierarchical softmax function, and the vector dimension was 100. A detailed description of the parameter settings for the word2vec training process can be found in Nawangsari et al. [23].

The final step was training the CNN model for document-level sentiment analysis. The architecture applied to this training process

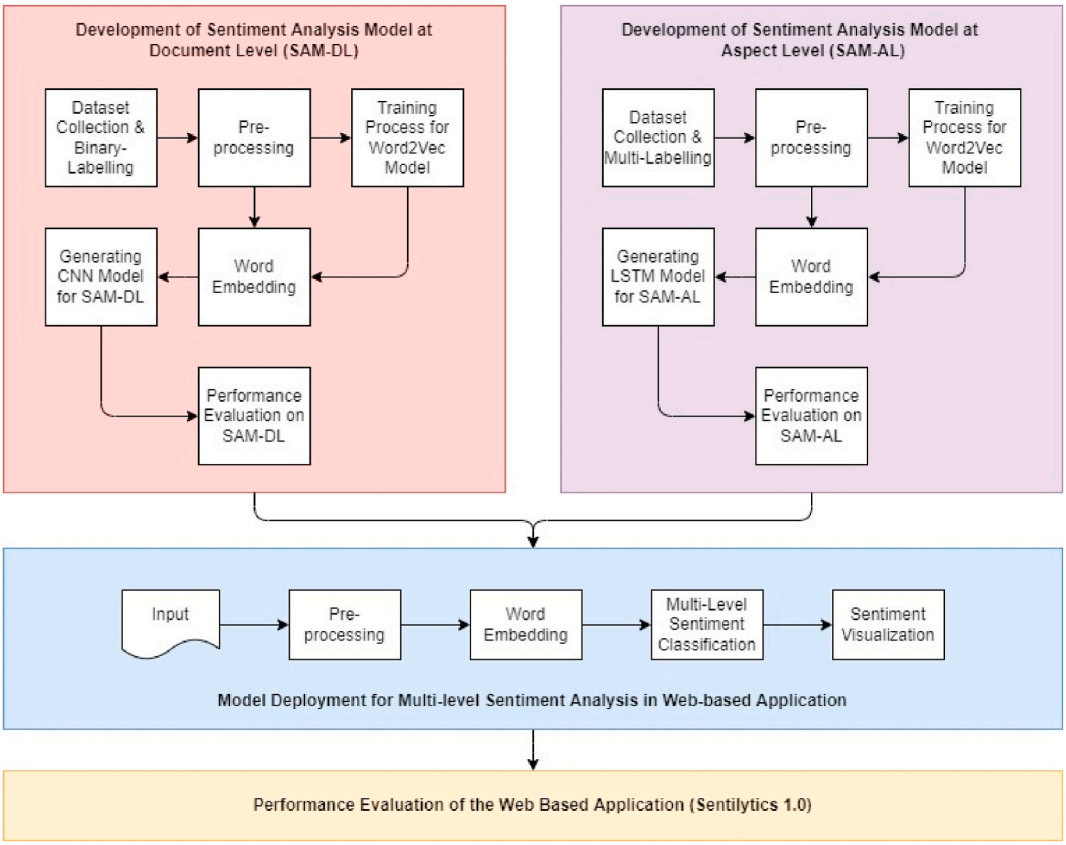

Fig. 1. Research methodology.

is shown in Fig. 2. The results from a previous study revealed that the CNN parameter that yielded the optimal performance was a dropout value of 0.2, the convolution activation function was a rectified linear unit (ReLU), the output activation function was a softmax function, and an Adam optimizer was used. The maximum validation accuracy was 98.16%. Using the same dataset as Nawangsari et al. [23], Muhammad et al. [24] employed LSTM to perform sentiment analysis at the document level with a lower validation accuracy of 85.96%.

3.2. Development of sentiment analysis model at the aspect level

The sentiment analysis model developed in this study was expected to identify five aspects: food, room, service, location, and miscellaneous. The selection of these five aspects was inspired by Pontiki et al. [42]. Their detected aspects for the hotel domain include hotels, rooms, facilities, amenities, service, location, and food and drinks. Because the two aspects of rooms and room amenities refer to hotel rooms, they were merged into one aspect: the room. The aspect of food drinks is simplified to become food, following the habits of Indonesians who do not need to distinguish between food and drink.

Furthermore, the aspects of the hotel and facilities, because they are not too specific, refer to something definite, so they are combined into one aspect, namely miscellaneous. The polarity assigned to each aspect was positive, neutral, or negative. To simplify the data labelling process, 15 binary spaces were used, as shown in Fig. 3.

The labelling process was conducted by three linguistic experts, with previously specified for each aspect as follows: (i). The food aspect accommodates reviews that discuss food-related matters, such as food taste, breakfast time, drink or beverage, variety of dishes, etc.; (ii). The room aspect accommodates reviews relating to the hotel'samenities, facilities, and atmosphere; (iii). The service aspect accommodates the discussion of hotel services, the friendliness of the staff, and others related to service. The location aspect accommodates reviews that discuss location-related matters, such as ease of access, hotel location, and supporting facilities around the hotel. Things that do not fit into the four aspects will enter into the miscellaneous aspect.

As previously explained, the labelling process uses 15 binary spaces, where every three digits represent an aspect, and each digit represents a positive, neutral, and negative polarity. Neutral polarity is obtained when the opinion review is neither positive nor negative or gives positive and negative reviews for the same aspect. Based on the example in Fig. 3, it can be concluded that the review contains two aspects: service and location. The service aspect shows neutral polarity, while the location aspect shows positive polarity.

Developing the sentiment analysis model for the aspect level was based on 5000 review documents with 10,283 aspects. The data consists of 2500 data similar to the data for the formation of SAM-DL (as explained in Section 3.1) and is added with 2500 new data randomly crawled from several hotels representing all levels of star hotels (1 star - 5 stars). In addition, we also added 200 crawled data as the test data. The abovementioned pre-processing steps were employed at this stage: case folding, stop-word removal, stemming, tokenization, and padding. Similarly, the training process of the word2vec model included the same parameters used in Muhammad et al. [24]. In particular, the architecture of the word2vec model was a skip-gram model, the evaluation method was a hierarchical softmax function, and the vector dimension was 300.

Before the training process is carried out in the development of SAM-AL, it is necessary to divide the research dataset. The 5000 preprocessed data are then divided into training and validation data based on the 10-fold cross-validation concept. This validation process aims to obtain the most optimal hyperparameter, commonly known as model selection. Furthermore, the testing process is carried out on 200 test data and is called the assessment model.

The final step was training the aspect-based sentiment analysis model using an improved LSTM. The architecture implemented in

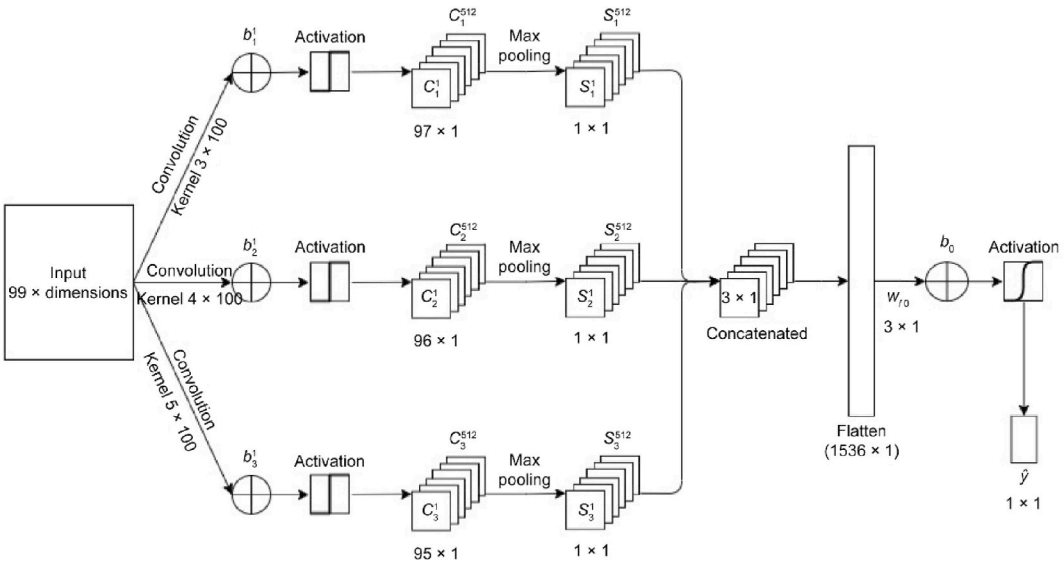

Fig. 2. The CNN architecture for document-level sentiment analysis.

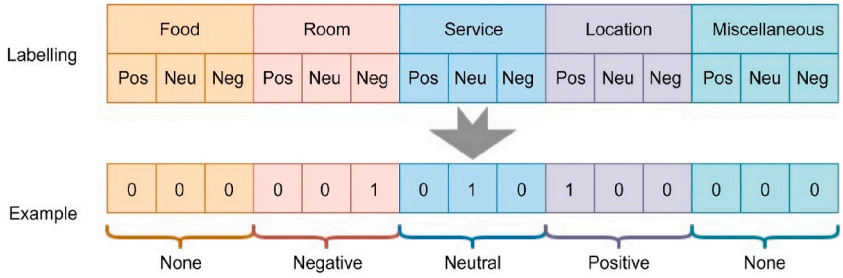

Fig. 3. Aspects and the corresponding polarity labelling.

this training process is shown in Fig. 4. The first layer is the embedding layer, which performs the word-embedding process on the input vector. The first layer produces an output with dimensions of 421 × 300, where 421 is the maximum length of the word vector, and 300 is the length of the embedding matrix. The vector resulting from word embedding is input to the LSTM layers. This layer predicts the class of the input vector as the output of each time step, which continues to the next time step ($t + 1$). The LSTM layer contains 421 cells, making predictions for each input vector. The output of this layer is inputted into the flattened layer to transform n-dimensional data into one-dimensional data. The output of this layer has dimensions of 126,300 × 1. The output is inputted to two fully connected layers. Moreover, this layer undergoes parameter tuning to determine the optimal parameters.

Each layer has a dropout mechanism to minimise the probability of overfitting in the trained model. The dropout value used was 0.5. Subsequently, it was inputted to an output layer with a size of 15 using the sigmoid activation function. This study used the sigmoid activation function instead of the softmax function, given that the classification type was multi-labelling with a value range of 0–1. The softmax function considers the value of each class as dependent on other classes, whereas the sigmoid function considers each class as independent. The output of this layer has dimensions of 15 × 1, corresponding to the class size defined in the previous section. The node value of the output layer ranges from 0 to 1. The following step involves rounding the values to obtain values of 0 or 1. The binary space is divided when 15 binary spaces are formed, according to Fig. 3. Subsequently, it is converted into a class representation by combining three existing binary values. Possible combinations of values are listed in Table 1. There are additional digit combinations for neutral polarities, as presented in Rows 3–5 of Table 1, whereas the non-detected class indicates that the document does not contain related aspects.

The optimal parameters of the LSTM method for aspect-based sentiment analysis are the number of hidden neurons and output activation in Fully connected layers 1 and 2. Combining the LSTM method with two fully connected layers, by which the optimal architecture was realized, yielded optimal results with a micro-average F1-score of 75.28%. In particular, fully connected layer 1 yielded optimal parameters for 1200 neurons with the tanh activation function, and fully connected layer 2 yielded optimal parameters for 600 neurons with the ReLU activation function. Compared with the standard LSTM model, our proposed model (modified LSTM with two fully connected layers) performs better at 10.16% [43].

3.3. Model deployment for multi-level sentiment analysis in web-based application

This stage aims to integrate both sentiment analysis models obtained from previous stages into real time to make practical sentiment analysis based on the submitted file of the OTA's hotel reviews. This stage manifests in the development of a sentiment analysis application, which is called Sentilytics 1.0. As depicted in Fig. 1, several modules in the Sentilytics 1.0 include input data, pre-processing, word embedding, multi-level sentiment analysis, and sentiment visualization.

The first module receives input data in a spreadsheet format with a ".xls" extension. Detailed explanations about this module will be described in the next section. The subsequent two modules are the same steps as the pre-processing and word embedding steps in the sentiment analysis model development for document level and aspect level. The multi-level sentiment analysis stage is the stage of carrying out the feed-forward computing process according to the two architectures described in Figs. 2 and 4 based on the CNN-based model for the document level and the LSTM-based model for the aspect level, respectively. Both models are obtained from the previous stages.

The output of the CNN-based model is the predicted polarity for each document, i.e., positive or negative. In contrast, the output of the LSTM-based model is the detected aspects and assigned polarities (i.e., positive, negative, or neutral). These outputs are subsequently inputted into the visualization module. The document-level visualization consists of two chart types: (i) a pie chart that indicates the percentage of sentiment polarities, and (ii) a line chart that indicates the sentiment timeline of the related hotel, e.g., the number of documents for each sentiment polarity per month. The visualization at the aspect level consists of two chart types: a bar chart that indicates guest perceptions of several aspects (percentage-wise) and a pie chart that indicates the percentage of sentiment polarities for each aspect.

3.4. Performance evaluation of the web-based application (Sentilytics 1.0)

Performance evaluation of Sentilytics 1.0 involved a comparison of the prediction results of the three popular OTA websites from Indonesia based on several metrics, namely, the precision, recall, and F1-score. Evaluation at the document level aims to evaluate the

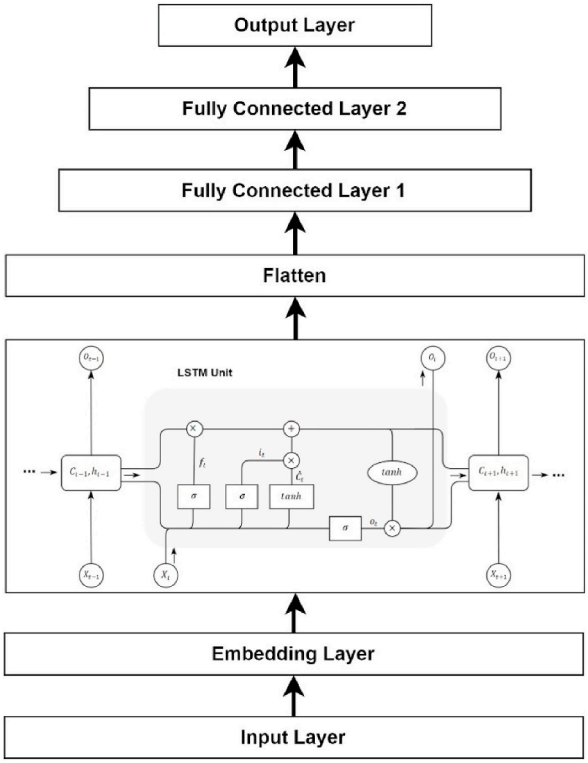

Fig. 4. Architecture of LSTM for aspect-level sentiment analysis.

Table 1
Digit combination for each class.

Detected Class	Digit		
	Positive	Neutral	Negative
Positive	1	0	0
Neutral	0	1	0
	0	1	1
	1	1	0
	1	0	1
Negative	0	0	1
None	0	0	0

performance of predicting a review document's positive or negative class based on CNN-based SAM-DL. In contrast, evaluation at the aspect level consists of the performance of the LSTM model in two tasks, namely aspect detection and aspect polarity detection.

4. Results and discussion

4.1. Sentilytics 1.0: a web-based application for multi-level sentiment analysis

Sentilytics 1.0 is a web-based application for the sentiment analysis of hotel reviews in the Indonesian language at multiple levels, specifically the document and aspect levels, based on deep learning models. This system can receive input reviews from various OTA websites under batch processing conditions, including Traveloka,[7] Pegipegi,[8] and Tiket.com.[9] We randomly collect 100 data for each OTA website. Hotel review data from various OTA websites should be crawled first and saved in a spreadsheet format, consists of two columns (date and review columns). The collected data for this stage was unseen, and we have not used it for training, validation, or testing for model generation.

[7] https://www.traveloka.com/en-id/hotel/.
[8] https://www.pegipegi.com/hotel/.
[9] https://www.pegipegi.com/hotel/.

Furthermore, because the data will be used to evaluate the model that has been developed and implemented as a backend engine on the developed platform, each data is also labelled manually. The labelling process was carried out by two people in the field of linguistics, and then calculating the Kappa statistics value was carried out to get an agreed value between the two labellers. The results of the manual label will be used as the gold standard for calculating the three metrics, including precision, recall, and F1-score. There is no limit to the number of data rows. However, according to the user requirements, the benchmark time execution for a single word is 6.99 ms.

The application interface for input review collection is shown in Fig. 5. This page consists of two buttons. The first button is for browsing the file, and the second is for initiating the sentiment analysis process.

After the user uploads a spreadsheet file and clicks the "START" button, the input data are transmitted to the backend module. The backend performs a sentiment analysis at the document and aspect levels. When the backend receives a request, the input data is checked for its extension, whereas the application only accepts files with the XLS extension. The data were parsed based on the date and review columns. Thereafter, pre-processing was performed by case-folding, tokenization, stop-word removal, stemming, padding, and vectorization. Word embedding was then performed to map each word in the document into a dense vector. Thereafter, the data were inputted to the two models, namely, the CNN model to predict sentiments at the document level and the LSTM model to predict sentiments at the aspect level.

These predicted outputs are subsequently inputted to the visualization module. Figs. 6 and 7 describe the interface for both document- and aspect-level visualization, respectively.

4.2. Performance evaluation of sentiment analysis at document level

The evaluation results for the three OTA websites for the document-level semantic prediction are shown in Table 2. The optimal model was obtained for the OTA-1 and OTA-2 datasets, with an F1-score of 0.97. However, different values were obtained for precision and recall. The recall and F1-score were the same for the entire dataset, with an average value of 0.95 and a standard deviation of 0.03. The average value of the precision was 0.96, and the standard deviation was 0.03. Thus, the sentiment analysis at the document level was appropriately predicted.

Table 2 shows that the recall value is lower than precision due to the high value of False Negative (FN) and low value of False Positive (FP). In the FN example, reviews that should be predicted as a positive class tends to be predicted by the system as a negative class. This is because the system detects many words that are positive but considered negative. For example, the combination of a word with the word *"tidak"* ("no") usually tends to be negative. However, in some reviews, the word *"tidak jauh"* ("not far") should be identified as a positive word instead of a negative. On the other hand, reviews that should be predicted as a negative class but tend to be predicted by the system as a positive class are infrequent. Examples that show the value of FP are the combination of a word with the word *"agak"* ("a bit"), such as *"agak lambat"* ("a bit slow"), *"agak banjir"* ("a bit flooded"), *"agak jutek"* ("a bit bitchy"), etc.

The limitation of the study, precisely sentiment analysis on the document level, is the inability of the system to recognize that the use of negative words followed by negative adjectives in Indonesian will result in positive sentence contexts, but the model predicts a negative sentiment. In future work, we will employ other word embedding techniques that can recognize the context of sentences to handle the limitation.

4.3. Performance evaluation of sentiment analysis at aspect level

4.3.1. Aspect detection

Aspect detection is related to the detection accuracy of user-review aspects as "existing" or "non-existent". The evaluation results at the aspect level for the three OTA websites are shown in Table 3.

Although the OTA-3 dataset demonstrated the lowest performance in the document-level prediction, it demonstrated the highest performance in the aspect-level prediction, with an F1-score of 0.88, precision of 0.92, and recall of 0.85. Concerning the entire dataset, the precision yielded the highest average value of 0.91 ± 0.02.

4.3.2. Aspect polarity detection

Aspect polarity detection determined each aspect's sentiment polarities as positive, neutral, and negative. The evaluation of the aspect polarity detection only applied to the accurately detected aspects.

The average values of the three datasets for each aspect are listed in Table 4. The average F1-score was 0.92, with a relatively large standard deviation of 0.07. This was because three aspects yielded F1-scores below 0.9, namely, the service (0.89), food (0.88), and room (0.84) aspects. Similar results were observed for the precision measurements. The precision was 0.92, with a standard deviation of 0.07. The high standard deviation was influenced by the food and room precision values significantly below 0.9, namely, 0.85 and 0.84, respectively.

Based on Table 4, the sentiment analysis on the location aspect has the best performance. That is, the F1-score value is equal to 1. This aspect is very specific and does not have ambiguity with other aspects. On the contrary, the room aspect has the lowest F1-score value, and this is due to ambiguity in recognizing it as room aspect or other aspects. For example, the review sentence *"handuknya kotor, tidak berwarna putih"* ("towels are dirty, not white"). This sentence will be recognized as an aspect of the room because the toiletries are part of the room or be considered miscellaneous.

Summarizing the previous explanation regarding the detection of aspect polarity, the limitation of this study is that the identified aspects need to be narrower, causing ambiguity during the detection process. Therefore, detecting polarity aspects based on the hotel

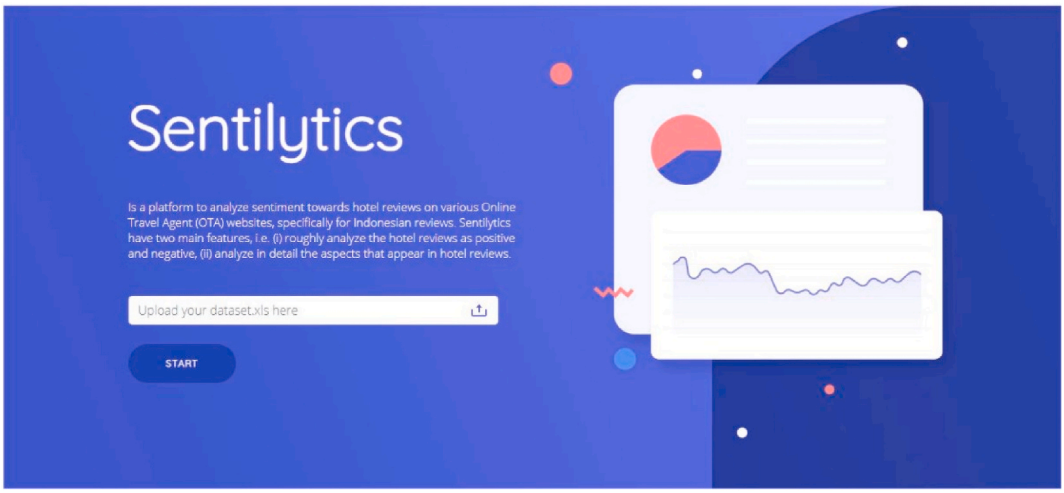

Fig. 5. Input page of sentilytics 1.0.

Fig. 6. Document-level visualization.

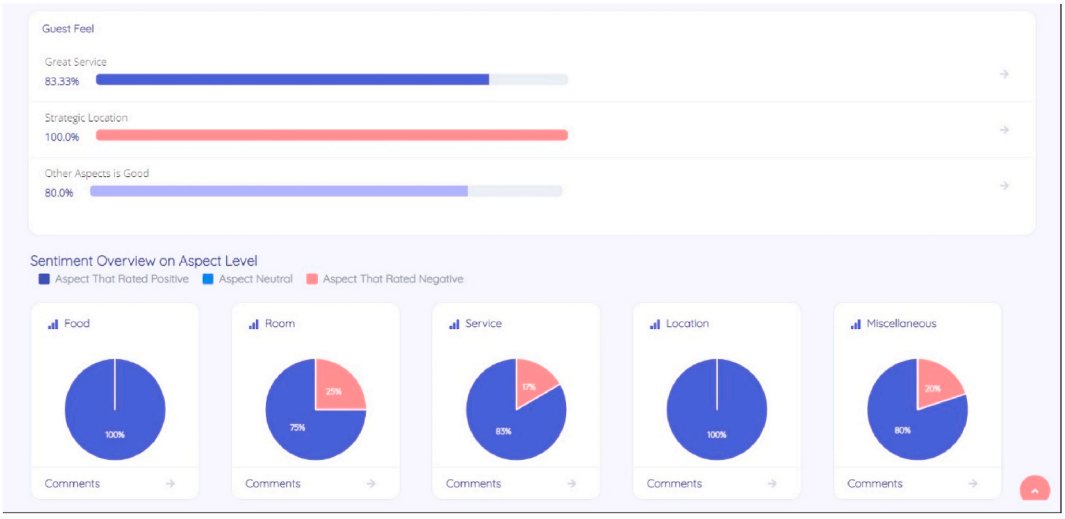

Fig. 7. Aspect-level visualization.

Table 2
Performance evaluation sentiment analysis on document level.

Dataset	Metrics Evaluation		
	Precision	Recall	F1-Score
OTA-1 Dataset	**0.98**	0.96	**0.97**
OTA-2 Dataset	0.97	**0.97**	**0.97**
OTA-3 Dataset	0.92	0.92	0.92
Average	0.96 ± 0.03	0.95 ± 0.03	0.95 ± 0.03

Table 3
Performance evaluation of aspect detection.

Dataset	Metrics Evaluation		
	Precision	Recall	F1-Score
OTA-1 Dataset	0,89	**0,85**	0,87
OTA-2 Dataset	0,91	0,81	0,85
OTA-3 Dataset	**0,92**	**0,85**	**0,88**
Average	0,91 ± 0.02	0,84 ± 0.02	0,87 ± 0.02

Table 4
Performance evaluation of aspect polarity detection.

Aspect	Metric Evaluation		
	Precision	Recall	F1-Score
Food	0.85	0.98	0.88
Room	0.84	0.9	0.84
Service	0.92	0.89	0.89
Location	**1.00**	**1.00**	**1.00**
Miscellaneous	0.97	0.97	0.97
Average	**0.92 ± 0.07**	**0.95 ± 0.05**	**0.92 ± 0.07**

guest cycle, as proposed by Sann and Lai [44], can be an alternative solution to this problem. The guest cycle includes pre-arrival, arrival, occupancy, and departure metrics.

5. Conclusions and future work

This article has presented a web-based application that retrieves hotel review documents from an OTA in Indonesian languages and analyses its sentiment from the coarse-grained document level to the fine-grained aspect level. The sentiment classification model employed in this application is based on deep learning models: a CNN-based classification model for the document level and an improved LSTM-based classification model for the aspect level. The validation performance accuracy for the document-level sentiment analysis was 98.16%, whereas the validation performance for the aspect-level sentiment analysis with respect to the F1-score was 75.28%.

The application uses several sentiment visualization types at both levels. The first visualization type at the document level is a pie chart indicating the percentage of sentiment polarities and a line chart indicating the related hotel's sentiment timeline. The second visualization type at the aspect level is a bar chart indicating the percentage of guest perceptions of several aspects and a pie chart indicating the percentage of sentiment polarities for each aspect.

Furthermore, we have conducted three performance evaluations using three datasets from three popular OTA websites in Indonesia to test unknown datasets. The results show that the F1-scores were 0.95 ± 0.03, 0.87 ± 0.02, and 0.92 ± 0.07 for document-level sentiment analysis, aspect detection, and aspect polarity detection, respectively. Based on these results, the application can be used to obtain an overview of a hotel's sentiment polarity or overall impression. In addition, the application can determine the aspects that obtained high ratings; thus, it is beneficial for a user that requires a reference for particular preferences to features or aspects provided by a hotel. Hotel management can use this platform to identify features or aspects that require improvement due to low ratings.

The aspect detection performance yielded the lowest F1 score based on the test results. However, aspect polarity detection demonstrated high performance. This study was limited in that the evaluation of aspect polarity detection was only performed on the accurately detected aspects; thus, the aspect detection model requires improvement. An effective strategy for future research involves detailing the aspects to be detected to eliminate ambiguity in the detection process. In particular, the guest cycle includes pre-arrival, arrival, occupancy, and departure metrics. In addition, further research can be developed by applying various kinds of the latest word embedding techniques, such as GloVe, FastText, or BERT (Bidirectional Encoder Representations from Transformers).

Author contribution statement

Retno Kusumaningrum: conceived and designed the experiments; performed the experiments; analyzed and interpreted the data; contributed reagents, materials, analysis tools or data; wrote the paper.

Iffa Zainan Nisa: performed the experiments; analyzed and interpreted the data; wrote the paper.

Rahmat Jayanto: performed the experiments; analyzed and interpreted the data; wrote the paper.

Rizka Putri Nawangsari: performed the experiments; analyzed and interpreted the data; wrote the paper.

Adi Wibowo: analyzed and interpreted the data; wrote the paper.

Data availability statement

Data will be made available on request.

Additional information

No additional information is available for this paper.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:Retno Kusumaningrum reports financial support was provided by Ministry of Research and Technology National Research and Innovation Agency. Retno Kusumaningrum has patent #000268163 licensed to Ministry of Law and Human Rights, Indonesia.

Acknowledgments

This work was supported by the Directorate Research and Development, under the Ministry of Research and Technology/National Agency for Research and Innovation, Indonesia [grant number 257-20/UN7.6.1/PP/2021].

References

[1] H.X. Shi, X.J. Li, A sentiment analysis model for hotel reviews based on supervised learning, in: 2011 International Conference on Machine Learning and Cybernetics (ICMLC), 2011, pp. 950–954.
[2] E. Lunando, A. Purwarianti, Indonesian social media sentiment analysis with sarcasm detection, in: 2013 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2013, 2013, pp. 195–198.
[3] W. Maharani, D.H. Widyantoro, M.L. Khodra, SAE: syntactic-based aspect and opinion extraction from product reviews, in: 2015 2nd International Conference on Advanced Informatics: Concepts, Theory and Applications, 2015, pp. 1–6.
[4] E. Wahyudi, R. Kusumaningrum, Aspect based sentiment analysis in e-commerce user reviews using latent dirichlet allocation (LDA) and sentiment lexicon, in: 3rd International Conference on Informatics and Computational Sciences (ICICoS), 2019, pp. 1–6.
[5] J. Fernando, M.L. Khodra, A.A. Septiandri, Aspect and opinion terms extraction using double embeddings and attention mechanism for indonesian hotel reviews, in: 2019 IEEE International Conference on Advanced Informatics: Concepts, Theory, and Applications, ICAICTA 2019, 2019, pp. 1–6.
[6] D.I. Af`idah, R. Kusumaningrum, B. Surarso, Long Short term memory convolutional neural network for indonesian sentiment analysis towards touristic destination reviews, in: 2020 International Seminar on Application for Technology of Information and Communication (ISemantic), 2020, pp. 630–637.
[7] F.W. Kurniawan, W. Maharani, Indonesian twitter sentiment analysis using Word2Vec, in: 2020 International Conference on Data Science and its Applications, ICoDSA 2020, 2020, pp. 31–36.
[8] H. Imaduddin, Widyawan, S. Fauziati, Word embedding comparison for indonesian language sentiment analysis, in: 2019 IEEE International Conference of Artificial Intelligence and Information Technology (ICAIIT), 2019, pp. 426–430.
[9] I.R. Putri, R. Kusumaningrum, Latent dirichlet allocation (LDA) for sentiment analysis toward tourism review in Indonesia, J. Phys. Conf. 801 (12073) (2017) 1–6.
[10] M.F.A. Bashri, R. Kusumaningrum, Sentiment Analysis Using Latent Dirichlet Allocation and Topic Polarity Wordcloud Visualization, in: 5th International Conference on Information and Communication Technology, ICoIC7 2017, 2017.
[11] E.S. Usop, R.R. Isnanto, R. Kusumaningrum, Part of speech features for sentiment classification based on latent Dirichlet allocation, in: 4th Int. Conf. on Information Tech., Computer, and Electrical Engineering (ICITACEE), 2017, pp. 27–30.
[12] W. Budiharto, M. Meiliana, Prediction and analysis of Indonesia presidential election from Twitter using sentiment analysis, J. Big Data 5 (1) (2018) 1–10.
[13] W. Satriaji, R. Kusumaningrum, Effect of synthetic minority oversampling technique (SMOTE), feature representation, and classification algorithm on imbalanced sentiment analysis, in: the 2nd International Conference on Informatics and Computational Sciences, 2018, pp. 99–103.
[14] S. Kurniawan, R. Kusumaningrum, Hierarchical Sentence Sentiment Analysis of Hotel Reviews Using the Naive Bayes Classifier, in: the 2nd International Conference on Informatics and Computational Sciences, 2018, pp. 104–108.
[15] N.M. Elfajr, R. Sarno, Sentiment Analysis Using Weighted Emoticons and SentiWordNet for Indonesian Language, in: 2018 IEEE International Seminar on Application for Technology of Information and Communication (ISemantic), 2018, pp. 234–238.
[16] T.G. Prahasiwi, R. Kusumaningrum, Implementation of negation handling techniques using modified syntactic Rule in Indonesian sentiment analysis, J. Phys. Conf. 1217 (1) (2019) 1–7.
[17] M. Dragoni, M. Federici, A. Rexha, ReUS: a real-time unsupervised system for monitoring opinion streams, Cogn. Comput. 11 (2019) 469–488.
[18] M. Korayem, K. Aljadda, D. Crandall, Sentiment/subjectivity analysis survey for languages other than English, Soc. Netw. Anal, Mining 6 (2016) 75.
[19] B. Liu, Sentiment analysis and opinion mining, Synth. Lect. Hum. Lang. Technol. 5 (1) (2012) 1–167.
[20] C. Zhang, D. Zeng, J. Li, F.-Y. Wang, W. Zuo, Sentiment analysis of Chinese documents: from sentence to document level, J. Am. Soc. Inf. Sci. Technol. 60 (12) (2009) 2474–2487.
[21] N. Liu, B. Shen, Aspect-based sentiment analysis with gated alternate neural network, Knowl. Base Syst. 188 (2020), 105010.
[22] L. Zhuang, K. Schouten, F. Frasincar, SOBA: semi-automated ontology builder for aspect-based sentiment analysis, J. Web Seman. 60 (2020), 100544.
[23] R.P. Nawangsari, R. Kusumaningrum, A. Wibowo, Word2vec for Indonesian sentiment analysis towards hotel reviews: an evaluation study, Proc. Comput. Sci. 157 (2019) 360–366.

[24] P.F. Muhammad, R. Kusumaningrum, A. Wibowo, Sentiment analysis using Word2vec and long short-term memory (LSTM) for Indonesian hotel reviews, Proc. Comput. Sci. 179 (2021) 728–735.
[25] B. Haryanto, Y. Ruldeviyani, F. Rohman, Nugroho, T.N. Julius Dimas, R. Magdalena, Y.F. Muhamad, Facebook analysis of community sentiment on 2019 Indonesian presidential candidates from Facebook opinion data, Proc. Comput. Sci. 161 (2019) 715–722.
[26] R. Moraes, J.F. Valiati, W.P. Gavião Neto, Document-level sentiment classification: an empirical comparison between SVM and ANN, Expert Syst. Appl. 40 (2) (2013) 621–633.
[27] A. Tripathy, A. Anand, S.K. Rath, Document-level sentiment classification using hybrid machine learning approach, Knowl. Inf. Syst. 53 (3) (2017) 805–831.
[28] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, B. Qin, Learning sentiment-specific word embedding for twitter sentiment classification, in: 52nd Annual Meeting of the Association for Computational Linguistics, 2014, pp. 1555–1565.
[29] Y. Bengio, Deep learning of representations: looking forward, Lect. Notes Comput. Sci. 7978 (2013) 1–37.
[30] A.A. Altowayan, L. Tao, Word Embeddings for Arabic Sentiment Analysis, in: 2016 IEEE International Conference on Big Data, 2016, pp. 3820–3825.
[31] W. Wang, Y. Yang, X. Wang, W. Wang, J. Li, Development of convolutional neural network and its application in image classification: a survey, Opt. Eng. 58 (4) (2019), 040901.
[32] S. Soni, S.S. Chouhan, S.S. Rathore, TextConvoNet: a convolutional neural network based architecture for text classification, Appl. Intell. 53 (2023) 14249–14268.
[33] N. Boudad, R. Faizi, R.O.H. Thami, R. Chiheb, Sentiment analysis in Arabic: a review of the literature, Ain Shams Eng. J. 9 (4) (2018) 2479–2490.
[34] M. Shams, N. Khoshavi, A. Baraani-Dastjerdi, LISA: language-independent method for aspect-based sentiment analysis, IEEE Access 8 (2020) 31034–31044.
[35] N.F. Alshammari, A.A. Almansour, Aspect-based sentiment analysis for arabic content in social media, in: 2nd International Conference on Electrical, Communication and Computer Engineering, 2020, pp. 1–6.
[36] C.C. Hnin, N. Naw, A. Win, Aspect Level Opinion Mining for Hotel Reviews in Myanmar Language, in: 2018 IEEE International Conference on Agents (ICA), 2018, pp. 132–135.
[37] L.P. Manik, H.F. Mustika, Z. Akbar, Y.A. Kartika, D.R. Saleh, F.A. Setiawan, I.A. Satya, Aspect-based sentiment analysis on candidate character traits in indonesian presidential election, in: 2020 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET), 2020.
[38] S. Gojali, M.L. Khodra, Aspect based sentiment analysis for review rating prediction, in: 2016 IEEE International Conference on Advanced Informatics: Concepts, Theory And Application, 2016, pp. 1–6.
[39] A.N. Azhar, M.L. Khodra, A.P. Sutiono, Multi-label Aspect Categorization with Convolutional Neural Networks and Extreme Gradient Boosting, in: 2019 International Conference on Electrical Engineering and Informatics, 2019, pp. 35–40.
[40] R.K. Yadav, L. Jiao, M. Goodwin, O.C. Granmo, Positionless aspect based sentiment analysis using attention mechanism, Knowl. Base Syst. 226 (2021), 107136.
[41] A. Ligthart, C. Catal, B. Tekinerdogan, Systematic reviews in sentiment analysis: a tertiary study, Artif. Intell. Rev. 54 (7) (2021) 4997–5053.
[42] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. Al-smadi, G. Eryiğit, SemEval-2016 task 5 : aspect based sentiment analysis, in: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 2016, pp. 19–30.
[43] R. Jayanto, R. Kusumaningrum, A. Wibowo, Aspect-based sentiment analysis for hotel reviews using an improved model of long short-term memory, Int. J. Adv. Intell. Inform. 8 (3) (2022) 391–403.
[44] R. Sann, P.C. Lai, Understanding homophily of service failure within the hotel guest cycle: applying NLP-aspect-based sentiment analysis to the hospitality industry, Int. J. Hospit. Manag. 91 (2020), 102678.