IJCCS (Indonesian Journal of Computing and Cybernetics Systems)
Vol.14, No.2, April 2020, pp. 123~134
ISSN (print): 1978-1520, ISSN (online): 2460-7258
DOI: 10.22146/ijccs.51646

Aspect-Based Sentiment Analysis of Online Marketplace Reviews Using Convolutional Neural Network

Mhd. Theo Ari Bangsa*[1], Sigit Priyanta[2], Yohanes Suyanto[3]
[1]Master Program of Computer Science; FMIPA UGM, Yogyakarta, Indonesia
[2,3]Department of Computer Science and Electronics, FMIPA UGM, Yogyakarta, Indonesia
e-mail: ***[1]amhd.theo@gmail.com**, [2]seagatejogja@ugm.ac.id, [3]yanto@ugm.ac.id

Abstrak

Sebagian besar toko online menyediakan fasilitas ulasan produk yang berisi tanggapan-tanggapan terhadap suatu produk. Banyaknya jumlah ulasan menyulitkan para calon pembeli untuk mengambil kesimpulan, sehingga diperlukan analisis sentimen untuk mengekstrak informasi dari ulasan-ulasan tersebut. Kebanyakan analisis sentimen dilakukan pada level dokumen sehingga hasil masih kurang detail karena klasifikasi dilakukan berdasarkan keseluruhan kalimat atau dokumen dan tidak mengidentifikasi aspek yang dibicarakan secara spesifik. Penelitian ini bertujuan untuk melakukan klasifikasi sentimen berbasis aspek terhadap ulasan toko online berbahasa Indonesia menggunakan metode convolutional neural network (CNN) dengan ekstraksi fitur menggunakan Word2Vec. Dataset yang digunakan adalah data ulasan berbahasa Indonesia dari situs bukalapak.com. Hasil pengujian pada sistem yang dibangun memperlihatkan bahwa metode CNN dengan ekstraksi fitur Word2Vec memiliki nilai score yang lebih baik daripada metode naive bayes dengan nilai akurasi 85.54%, presisi 96.12%, recall 88.39%, dan f-measure 92.02%. Klasfikasi tanpa menggunakan preprocessing stemming pada dataset meningkatkan akurasi sebesar 2.77%.

Kata kunci*— analisis sentimen berbasis aspek, convolutional neural network, toko online*

Abstract

Most online stores provide product review facilities that contain responses to a product. The number of reviews makes it difficult for potential customers to make conclusions, so that sentiment analysis is needed to extract information from these reviews. Most sentiment analysis is done at the document level, so the results were still lacking in detail because the classification is based on the entire sentence or document and does not identify the specific aspect discussed. This research aims to classify aspect-based sentiments from online store reviews using the convolutional neural network (CNN) method with the extraction of features using Word2Vec. The dataset used is Indonesian review data from the site bukalapak.com. The test results on the built system showed that CNN's method of Word2Vec feature extraction has a better score than the naive bayes method with an accuracy value of 85.54%, 96.12% precision, 88.39% recall, and f-measure 92.02%. Classification without using stemming preprocessing on the dataset increases the accuracy by 2.77%.

Keywords *— aspect-based sentiment analysis, convolutional neural network, online store*

Received November 19th,2019; Revised April 20th, 2020; Accepted April 30th, 2020

1. INTRODUCTION

The development of Internet technology affects several areas of life, one of which is an online transaction through a website. Most online stores provide product review facilities where customers who have purchased a product on the online store are asked to write the opinion. The reviews that buyers provide can be positive, negative or neutral opinions, and sometimes comments are also specific to certain aspects such as service aspects, prices, deliveries, etc.

The number of reviews available leads to a difficult candidate to make a decision. Therefore it is necessary to classify opinions that can process review data to distinguish negative and positive opinions based on the category of aspects discussed. One method that can be used for classifying opinions is sentiment analysis..

The analysis of sentiment or opinion mining is a computational study of the opinions, sentiments, and emotions expressed in the text. Sentiment analysis is aimed at seeing the tendency of opinions or opinions on an object, whether it tends to be positive or negative [1].

Some of the methods that are often used in conducting sentiment analysis include machine learning methods such as Naive Bayes, Maximum Entropy (ME), and Support Vector Machine (SVM). But the machine learning method has a weakness in selecting a complex feature extraction method and finding a better feature type [2]. Besides, the machine learning method also has a weakness in large data usage, and it affects the performance of the classification accuracy to be generated [3]. At this time, researches implementing deep learning methods on sentiment analysis have been conducted on several studies and getting better performance results [4].

Deep learning is the development of machine learning a "deep" term in deep learning refers to the depth of the architecture of a neural network [5]. Some deep learning methods used for text classification include Recursive Neural Network (RNN) and Convolutional Neural Network (CNN). The CNN method has an advantage over RNN, where the CNN method requires fewer connections and parameters and is easier to do training. CNN's methods have been used in research [6], and [7] for sentiment analysis, where the results of their research have better performance than conventional machine learning methods.

Previous research relating to the analysis of aspects-based sentiment on online stores has been conducted by Fachrina and Widyantoro [8]. They compare the rule base, Naïve Bayes, and SVM methods. For their extraction features use the Unigram feature, POS feature and pattern and rule feature. Besides, research with different data was done by Gojali and Khodra [9] using restaurant data. They compare naïve Bayes, J48, and SVM. The final result is a rating for each aspect specified. Further research on Pratama at al. [10] have used tourism data. The method used is the SVM method with the extraction of TF-IDF features.

Based on the background shown, this study tried to use the Convolutional Neural Network method for aspect-based sentiment analysis on online marketplace reviews. There are two stages of the approach that is the first stage of the classification of aspect categories into six classes and the second stage of the sentiment classification based on each aspect category. The six aspects used are accuracy, quality, service, price, packaging, and delivery, while sentiment polarity is grouped into two classes of positive and negative classes.

2. METHODS

In this section, we discuss the architecture and method used to classify aspect-based sentiment. There are six main stages: data collection, preprocessing, aspect category classification, sentiment classification, aspect category and sentiment sequentially classification, and finally testing and evaluation. The description of the process stages that will be performed in this research can be seen in Figure 1.

The first stage that will be done is the collection of review data from the website www.bukalapak.com, and data retrieval is done by the scrapping technique using the Selenium library in the Python programming language. The scrapping Data is used in the second stage of the preprocessing stage. Preprocessing to clean review data consists of case folding, symbols removal, and normalization of words. In this study also carried out stemming influence testing so that necessary preprocessing output consisting of preprocessing without stemming and that uses stemming. The third stage is the classification of the category of aspects into six categories, namely aspects of accuracy, quality, service, packaging, price, and delivery. Each review can have more than one aspect or Multilabel so the classifications will be done with binary relevance technique. The fourth stage is the sentiment classification phase. Each aspect will be classified into two polarities, namely positive and negative sentiments. The fifth stage is an aspect category and sentiment sequentially classification. Label output on the aspect category classification will be the input for the sentiment classification. The sixth stage is testing and evaluation. The test was conducted with a K-fold cross-validation approach whereby the data was divided as K partitions randomly. Furthermore, experiments conducted by K experiments where each experiment made use of the K partition as data testing and used the rest as training data.

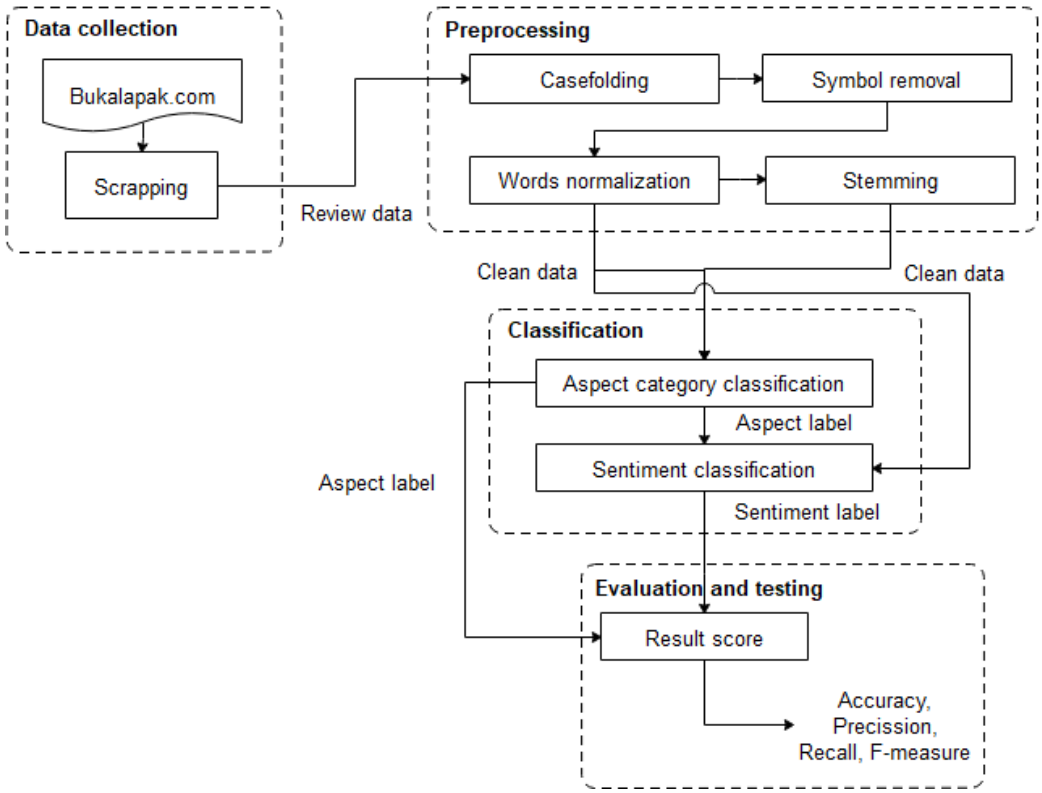

Figure 1 System architecture

2. 1 Data collection

Review dataset is obtained from an online store website www.bukalapak.com. Data captured by only the review data is descriptive text from 98 stores with the electronic domain. The scrapping process is done using the Selenium 3.141.0 library. Data is saved in files CSV format.

The data labelling of the aspect category reviews and sentiments is done manually by the majority voting technique by three people who are the doctoral students of Pendidikan Bahasa Universitas Negeri Jakarta. Each sentence on a review can have more than one aspect category so that the labelling of sentences uses multi-label.

The grouping of aspect categories follows the work conducted by Fachrina and Widyantoro (2017). There are seven categories of aspects, namely, quality, service, communication, price, packaging, and delivery. But in this research aspects of communication combined with the aspect of service, so there are six categories of aspects used.

Table 1 Sample data for aspect category classification

Text Reviews	**Aspect Category**					
	Accuracy	**Quality**	**Service**	**Price**	**Packaging**	**Delivery**
Barang sesuai dengan pictnya. Packingnya rapi dan bagus	1	0	0	0	1	0
Di kira besar ukuranya ternyata kecil. Cukup mahal dengan harga segitu barangnya kecil.	1	0	0	1	0	0
barrang bagus tapi sayang pengiriman lama Namun akan lebih baik apabila cepat sampai	0	1	0	0	0	1

Sentiment labels are grouped into two classes in each aspect category, both positive and negative. Examples of the dataset for sentiment classification can be seen in Table 2 table.

Table 2 Sample data for sentiment classification

Text Reviews	**Aspect Category**					
	Accuracy	**Quality**	**Service**	**Price**	**Packaging**	**Delivery**
Barang sesuai dengan pictnya. Packingnya rapi dan bagus	Pos	-	-	-	Pos	-
Di kira besar ukuranya ternyata kecil. Cukup mahal dengan harga segitu barangnya kecil.	Neg	-	-	Neg	-	-
barrang bagus tapi sayang pengiriman lama Namun akan lebih baik apabila cepat sampai	-	Pos	-	-	-	Neg

2. 2 Data preprocessing

Data preprocessing is to manage review data to make it easier to process later. In this research, there are four stages of preprocessing conducted.

2. 2.1 Casefolding

This process is done by changing any text containing capital letters with lowercase letters. Only the letter ' a ' until the letter ' Z ' is changed, characters other than letters are omitted or considered delimiter.

2. 2.2 Symbol removal

Deletion symbols and punctuation marks are used to remove special characters (#, @,%, $). In addition to the removal of punctuation (.,?! ' ") and numeric number (0-9).

2. 2.3 Word normalization

Normalization of words aims to change the abbreviated word, the word that has the error in writing, and the word is not raw to be a word that corresponds to the great Dictionary of Bahasa Indonesia (KBBI). This process is done by first creating a Slangword dictionary was made manually based on the observation of the dataset used.

2. 2.4 Stemming

Stemming is the process of converting the word into a basic word. The conversion process involves the rules of Bahasa Indonesia. The stemming process uses the Sastrawi 1.0.1 library in the Python programming language with text input that has been performed normalized in the previous word.

2.3 Classification

Classification is done with two classification stages. The first is a classification of aspect categories into six classes. In the second stage, the result of the first stage classification is the input for the second stage of the sentiment classification of each aspect of the first phase. The illustration process of the two-stage classification and sentiment categories can be shown in Figure. 2.

2.3.1 Aspect category classification

Each review can have more than one aspect so that the classification process in this research can be seen as a matter of multi-label classification. The technique used for multi-label classification in this research is to use binary relevance strategy, which is to build the classification model as much as the number of classes separately.

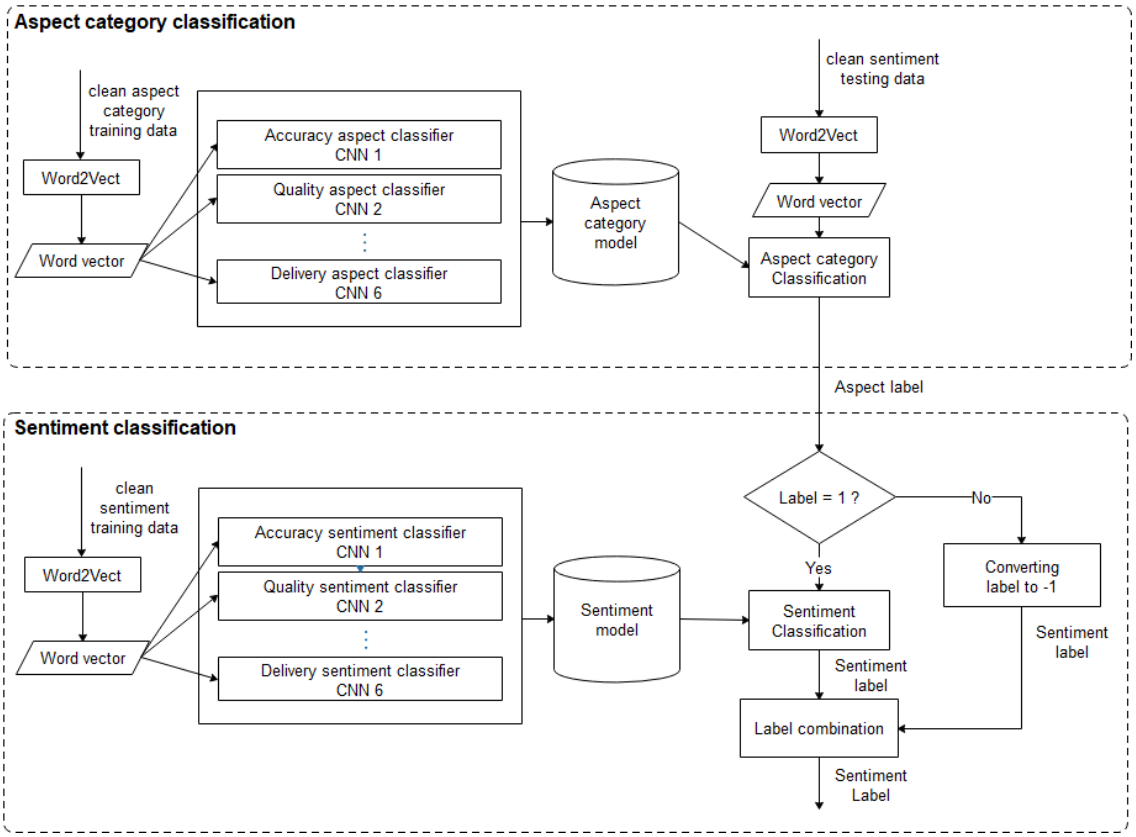

Figure 2 Flow diagram of two step aspect-based sentiment classification

The training process begins with converting the word into vector words using Word2vec, and then the data is trained using the CNN algorithm. Each aspect category has its own classifier, so there are six classifiers used by CNN 1 to CNN 6. For a comparison of the methods in Fig. 3 CNN method can be replaced with Naïve Bayes method while weighted word TF-IDF in place of Word2Vec. Each classifier generates an output of the classification model used for the classification process stage with test data. The output of the classification process is the predictor of the aspect label in the form of a binary class, i.e. "1" If the aspect is detected and the label "0" is otherwise not detected as an aspect. If the predicting label result is 1, then the data will be done in the sentiment classification, but if the predicting label result is 0, it will be converted to-1 as the value false for the final output.

2. 3.2 Sentiment classification

Sentiment classification is done in the same way as aspect category classification. Each aspect has its own individual classification model. Each aspect of the review is classified by two classes of positive and negative classes. The output of the aspect classification into inputs for the sentiment classification.

The training phase of the aspect and sentiment categories is done separately by using each training dataset. While classification data uses test sentiment data for both models. The final output of this two-stage classification process is derived from the label -1 (false) combination of the Aspect model output, and the output of the sentiment model is 0 (negative), and 1 (positive).

2 3.3 Convert word to vector using pretraining Word2Vec

Convert words into Word2vect vector shapes using the Gensim 3.6.0 library. The Library provides functions pertaining to training, storage, as well as the use of the Word2vect model as an expeller feature. The basis for generating vector value on Word2vec is the neural network. There are two neural network architectures to train the Word2vect model, namely a continuous bag of words (CBOW) and skip-gram. This research uses a skip-gram architecture with a vector size of 300 dimensions, and the number of Windows is 10. The reason for the skip-gram architecture selection is that skip-grams work well on a small amount of data and can represent words with a slightly better frequency of occurrence [11]. The vector value generated by Word2vect can represent the meaning of a word. Before Word2vect used to represent vector value first done model creation of word distribution during training. The word distribution used in this research is a corpus of product reviews on Bukalapak's online store as much as 228.612 reviews taken from 98 stores.

2.3.4 Convolutional neural network

CNN architecture used in this research has three layers: convolutional layer, pooling layer, and fully connected layer. CNN's architectural illustration can be seen in Figure. 3.

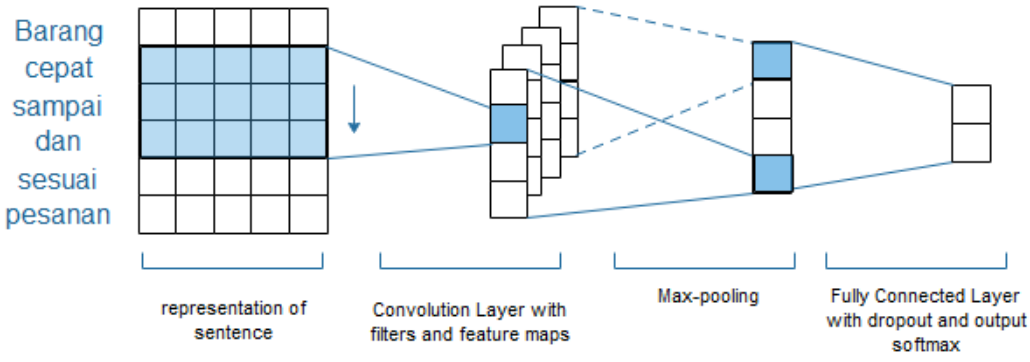

Figure 3 Illustration of CNN architecture

The input layer is the first layer composed of the review sentence with a predefined fixed size. In this research, the input size used is 154 words obtained from the length of the maximum sentence size of the dataset of the reviews used and already through preprocessing.

The input of each text must have the same length so that it needs to be done padding on a sentence that has a word less than the input size. Word Padding is done by adding a number 0 to a sentence that has a size less than the input size set.

After padding, then the word conversion in the input text becomes vector with a fixed size based on the dimensions of the Word2vec model. In this research, the dimension of vector Word2vec had a size of 300. The process of converting the word to vector is that each word input in the text will be checked in the available word dictionary if the word is in the word dictionary, the model will return the word vector value. If the word is not found, then the vector value to be used is a random value with the vector dimension equal to Word2vec.

The convolutional layer is a layer that performs the convolution process against input layers and filters. In this research uses three filters of different sizes. The Filter contains weights that are initialized using randomly obtained values, and the weight of those values will continue to be updated during the training process. In the case of one dimension, the convolution between two vectors $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{f} \in \mathbb{R}^m$ is a vector $\mathbf{c} \in \mathbb{R}^{n-m+1}$, in which each element is $\boldsymbol{c_i}$ counted as a scalar value of the product between Vector f and correspondent of x [7]:

$$c_i = f^T x[i: i + m - 1] \tag{1}$$

The activation function used to output the convolution layer is ReLU. The way ReLU works is to change the negative value to 0 so that the use of this ReLU activation function will generate value on each element of the feature map positively valued. The equation of this function can be shown in equation (2).

$$f(x) = \max(0, x) \tag{2}$$

Feature maps that have been performed ReLU activation function hereinafter carried out the pooling or subsampling process which aims to reduce the size of the matrix by using pooling operation. The pooling method used in this study is the max pooling which the value is taken is the maximum value of the previous convolutional layer. The Max pooling operation is the same as the convolution operation, i.e. using filters run throughout the feature map. However, the filter that is used on the max-pooling process does not have a weighted value. The output of the max-pooling process results in a feature map that is then done by the flatten process so that the feature map transforms into a vector value used as input on a fully connected layer.

The Fully-connected layer is a hidden layer in general on neural networks that form one dimensional neuron. It consists of neurons interconnected with the neurons on the previous and subsequent layers. Each neuron has a ReLU activation function.

The layer Output is the last layer consisting of 2 neurons with the Softmax activation function. Softmax is useful to change the output of the last layer in the neural network into a basic distribution probability. The equation of this function is shown in the Eq. (3) where *K* represents the number of classes and $\boldsymbol{\theta_c}$ is the vector weight associated with the C class [7].

$$P(y = c|\mathrm{h}) = \frac{e^{\mathrm{h}^T \theta_c}}{\sum_{c=1}^{K} e^{\mathrm{h}^T \theta_c}} \tag{3}$$

2. 4 Testing and evaluation

Testing and evaluation are the stages undertaken for testing of the classification model. Testing was conducted using K-fold cross validation using k = 5. Data will be divided into 5 partitions that have the same amount, data consists of 4 training data and 1 data testing. Further calculations to measure the performance of classification with confusion matrix. The test results displayed the values of accuracy, precision, recall, and f-measure.

3. RESULTS AND DISCUSSION

This section discusses the test results of a system that has been built. In the first phase the search performed the best model on CNN architecture. In the second stage carried out the performance testing of CNN method compared to the naïve Bayes method, the third stage of testing was conducted against the size of data. Testing the first stage until the third stage of the stemming process on the dataset is not used because the stemming process in the review dataset is included in the stemming impact testing phase at the fourth stage.

Table 1 Dataset distribution

Aspect category	Sentiment polarity	Amount of data
Accuracy	Positif	2226
	Negatif	992
Quality	Positif	3007
	Negatif	1337
Service	Positif	880
	Negatif	479
Packaging	Positif	806
	Negatif	386
Price	Positif	831
	Negatif	234
Delivery	Positif	1142
	Negatif	458

The amount of data used in this study was 7500 review data. The Data is duplicated into two datasets. The first Dataset is grouped into six classes: accuracy, quality, service, packaging, price, and delivery. Aspect category Data is used for training and testing in the first phase of the aspect category classification. The second Dataset of each aspect category data is divided into two positive and negative classes for sentiment data. Data distribution can be shown in Table 1.

3. 1 CNN hyperparameter testing

Table 4. CNN Hyperparameter.

No.	Hyperparameter	Value
1	Filter size	(2,4,6), (3,5,7), (4,6,8)
2	Number of feature maps	100, 200, 400
3	Drop out probability	0.5
4	Epoch	30, 45, 60
5	Batch size	50
6	Learning rate	0.01

Hyperparameter testing aims to find the optimal hyperparameter for the accuracy performance of CNN's architectural model. Testing was conducted on the Aspect category classification model and the sentiment classification model separately. As the initial configuration of hyperparameters was used following the research conducted by Zhang and Wallace [12] which can be shown in Table 4. Testing was conducted against several hyperparameters epoch, filter sizes, and feature maps.

3. 1.1 Aspect category classification test results with optimal hyperparamter

After testing on several hyperparameters, then obtained the best hyperparameter for the aspect classification, namely filter size (2, 3, 4), the number of feature map 100, and the number of epoch 60. The overall test results with the best hyperparameters gained in previous tests can be shown in Table 5.

Table 5 Test results on Aspect category classification with optimal hyperparameters

Aspect category	Time (minutes)	Accuracy	Precision	Recall	F-Measure
Accuracy	9.73	91.36%	88.45%	91.90%	90.14%
Quality	9.9	74.76%	70.11%	98.39%	81.87%
Service	9.7	95.88%	92.58%	84.03%	88.07%
Packaging	9.67	97.93%	97.10%	89.73%	93.24%
Price	9.68	97.08%	90.80%	88.54%	89.59%
Delivery	9.57	94.85%	87.86%	88.06%	87.94%
Average	**9.71**	**91.98%**	**87.82%**	**90.11%**	**88.47%**

The result of the highest aspect classification category accuracy is classification on packaging aspect with book accuracy 97.93% while the lowest accuracy is owned by quality aspect with 74.76% accuracy value. Frequent classification errors are caused by many words that often appear on one aspect also often appear in many other aspects so that the classification becomes biased as in the text "*Kiriman cepat tiba dan kemasan mantap*". The word "*mantap*" Many appear in the aspect of quality so that the text has a high probability of being detected into the aspect of quality, while the actual target of the text is two targets, namely packaging aspects and delivery aspects**.**

3. 1.2 Sentiment classification test results with optimal hyperparamter

The best hyperparameters for sentiment classification are obtained by filter size (3, 4, 5), the number of feature map 400, and the number of Epoch 60 overall test result with the best hyperparameter obtained in previous tests can be shown on Table 6.

Table 5 Test results on Aspect category classification with optimal hyperparameters

Aspect category	Time (minutes)	Accuracy	Precision	Recall	F-Measure
Accuracy	10.75	95.68%	97.93%	95.77%	96.84%
Quality	14.53	90.79%	95.44%	91.05%	93.19%
Service	4.53	93.38%	97.49%	92.17%	94.74%
Packaging	3.98	94.13%	94.87%	96.53%	95.69%
Price	3.57	89.86%	90.45%	97.24%	93.68%
Delivery	5.35	94.56%	97.40%	94.93%	96.14%
Average	**7.12**	**93.07%**	**95.60%**	**94.62%**	**95.05%**

The highest sentiment classification accuracy result is a classification on the aspect of accuracy with book accuracy of 95.68% while the lowest accuracy is owned by the price aspect with a value of 89.86% accuracy. Fault classification sentiment often appears on the price

aspects with sentences containing negation words such as "*harga tidak murah kualitas okelah fungsi suara baguss...*" On this text, the sentiment classification model predicts the text of a positive price aspect while the actual target is a negative price.

3. 2 Performance testing results of CNN's method on two stages of aspect-based sentiment sequentially

At this stage, the aspect classification testing and sentiment are performed sequentially. There are two data used during training models such as aspect category data to training aspect category and sentiment data for the sentiment model training. However, when testing the data used for both stages is sentiment data. Testing was conducted using 5-fold cross-validation, where the output of the aspect category classification would be the sentiment classification input. Comparison of two-stage classification results sequentially using CNN's and NB method can be seen in Table 7.

Table 7 Comparison of aspect and sentiment category classification results

Aspect category	Methods	Accuracy	Precision	Recall	F-Measure
accuracy	CNN	89.22%	98.61%	92.81%	95.62%
	Naïve bayes	72.87%	92.51%	81.72%	86.76%
quality	CNN	89.73%	95.38%	89.79%	92.49%
	Naïve bayes	76.80%	87.97%	78.93%	83.18%
service	CNN	81.09%	97.50%	82.52%	89.31%
	Naïve bayes	54.01%	90.83%	56.14%	69.37%
packaging	CNN	87.08%	95.85%	89.31%	92.46%
	Naïve bayes	58.90%	89.46%	63.23%	74.04%
price	CNN	89.86%	90.45%	97.24%	93.68%
	Naïve bayes	54.84%	86.32%	63.51%	73.13%
delivery	CNN	85.56%	97.83%	86.46%	91.78%
	Naïve bayes	57.75%	95.04%	62.40%	75.30%
Average	**CNN**	**85.54%**	**96.12%**	**88.39%**	**92.02%**
	Naïve bayes	**62.53%**	**90.36%**	**67.66%**	**76.96%**

The results in Table 7 indicate that the CNN model with the Hyperparameter used was the filter size (2, 3, 4), the number of the feature map 100, and the 60 epoch for the aspect category model, while the filter size (3, 4, 5), the number of the map Featur 400 for the sentiment model, Has a pretty good performance compared to the naïve Bayes model with the average CNN accuracy value of 85.54%.

3. 3 Data Size test Results

In this section, testing is done by varying the amount of data used for the training and testing dataset, and then conducting an analysis of its impact on the outcome of the aspect-based sentiment classification performance with the sequential two stages of classification obtained. Testing influence the size of the dataset is done using the same data distribution as all the tests that have been done. The distribution of datasets is done by dividing data into four sections. The first tests were conducted using 1/4 random data, and the second testing was done using 2/4 data, the third testing using 3/4 data, and subsequent testing using the entire data. The results of the test comparison of the four datasets are shown in Table 8.

Based on testing the size of data that has been done, it can be concluded that the more data used in the training and testing process, the better the performance of the classification of models. This is because by adding data, it will increase the knowledge that is owned by a model so that the model can know more patterns of sentences in the review dataset

Table 8 Comparison of data size testing

Amount of data	Accuracy	Precision	Recall	F-measure
1/4 data	78.05%	92.07%	84.08%	87.68%
2/4 data	84.31%	95.63%	88.05%	91.57%
3/4 data	85.66%	95.98%	88.77%	92.15%
All data	85.54%	96.12%	88.39%	92.02%

3. 4 The stemming impact test results

This test was conducted to see the stemming influence on the review dataset to the accuracy gained using CNN's method with the extraction of the Word2vec feature and also the naïve Bayes method with the extraction of TF-IDF features. Comparison of the stemming influence accuracy results in CNN and naïve Bayes ' methods can be seen in Table 9. The test results are the average of the 5-fold cross-validation of the average 6-aspect score.

Table 9 Stemming influence

Methods	Accuracy	Precision	Recall	F-measure
CNN	85.54%	96.12%	88.39%	92.02%
CNN-*Stemming*	82.77%	96.01%	85.48%	90.22%
NB	62.53%	90.36%	67.66%	76.96%
NB-*Stemming*	61.04%	90.46%	65.61%	75.52%

Based on the results shown in Table 9 can be concluded that the performance of the classification if the dataset carried out extensions stemming from achieving accuracy results less so good on CNN and NB method. The accuracy of CNN's method decreased by 2.77% when using stemming, while the accuracy of the NB method decreased by 1.49%.

4. CONCLUSIONS

Based on the results of the tests, the conclusion of this research is that aspect-based sentiment analysis can be performed using two-stage classification with convolutional neural network methods and feature extraction word2vec. CNN's method with the use of the best hyperparameter combination on the second stage architecture model has the best average accuracy compared to the naïve Bayes method with an accuracy value of 85.54%, 96.12% precision, 88.39% recall, and f-measure 92.02% While the naïve Bayes method has an accuracy value of 62.53%, 90.36% precision, 67.66% recall, and f-measure 76.96%. This is because CNN's method is very good at detecting specific features for both aspects and sentiment categories in the form of words or phrases in a text regardless of the position of the word or phrase. The accuracy of CNN models is also influenced by the amount of data used. By increasing the number of training datasets, it will also increase the accuracy of the resulting value. The use of stemming preprocessing on the two-stage classification process in a sequential way has less good results. Accuracy decreased by 2.77% when using stemming preprocessing on CNN method with Word2vec feature extraction.

This research still has a deficiency that can be improved. Some suggestions for further research are to add a method for determining the expression target opinion, which determines the expression of the opinion of words or phrases associated with the aspect entity. Next, add a method of negation handling in the classification phase of sentiment.

REFERENCES

[1] B. Liu, "Sentiment analysis and opinion mining," *Synth. Lect. Hum. Lang. Technol.*, vol. 5, no. 1, pp. 1–167, 2012.

[2] Z. Su, H. Xu, D. Zhang, and Y. Xu, "Chinese sentiment classification using a neural network tool - Word2vec," in *2014 International Conference on Multisensor Fusion and Information Integration for Intelligent Systems (MFI)*, 2014, pp. 1–6.

[3] M. A. Nasichuddin, "Pengaruh Matriks Filter, Kerangka, dan Pra Pelatihan pada Peningkatan Kinerja Pelatihan CNN Untuk Analisis Sentimen," Universitas Gadjah Mada, 2018.

[4] X. Ouyang, P. Zhou, C. H. Li, and L. Liu, "Sentiment Analysis Using Convolutional Neural Network," in *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*, 2015, pp. 2359–2364.

[5] R. Socher, Y. Bengio, and C. Manning, *Deep Learning for Natural Language Processing (without Magic)*. 2012.

[6] F. Ratnawati and E. Winarko, "Sentiment Analysis of Movie Opinion in Twitter Using Dynamic Convolutional Neural Network Algorithm," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 12, no. 1, p. 1, 2018.

[7] L. Akhtyamova, J. Cardiff, and M. Alexandrov, "Adverse drug extraction in twitter data using convolutional neural network," *Proc. - Int. Work. Database Expert Syst. Appl. DEXA*, vol. 2017-Augus, pp. 88–92, 2017.

[8] Z. Fachrina and D. H. Widyantoro, "Aspect-sentiment classification in opinion mining using the combination of rule-based and machine learning," *Proc. 2017 Int. Conf. Data Softw. Eng. ICoDSE 2017*, pp. 1–6, 2017.

[9] S. Gojali and M. L. Khodra, "Aspect based sentiment analysis for review rating prediction," *4th IGNITE Conf. 2016 Int. Conf. Adv. Informatics Concepts, Theory Appl. ICAICTA 2016*, 2016.

[10] Y. T. Pratama, F. A. Bachtiar, and N. Y. Setiawan, "Analisis Sentimen Opini Pelanggan Terhadap Aspek Pariwisata Pantai Malang Selatan Menggunakan TF-IDF dan Support Vector Machine," vol. 2, no. 12, pp. 6244–6252, 2018.

[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv Prepr. arXiv1301.3781*, 2013.

[12] Y. Zhang and B. Wallace, "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification," *arXiv Prepr. arXiv1510.03820*, 2015.