






# Sentiment Analysis of Student Evaluations of Teaching Using Deep Learning Approach

Edi Sutoyo<sup>1</sup> , Ahmad Almaarif<sup>1</sup> , and Iwan Tri Riyadi Yanto<sup>2</sup> 

<sup>1</sup> Department of Information Systems, Telkom University, Bandung, West Java, Indonesia  
{edisutoyo, ahmadalmaarif}@telkomuniversity.ac.id

<sup>2</sup> Departement of Information System, Universitas Ahmad Dahlan, Yogyakarta, Indonesia  
yanto.itr@is.uad.ac.id

**Abstract.** One part of improving and developing lecturer pedagogical competence and evaluating the substance of the courses taught in the lecture process is to use Lecturer Evaluation Effectiveness by Students (EDOM). The EDOM is a feedback questionnaire that can be used by educators to assess and evaluate pedagogical competencies that they have and aims to make continuous improvements to various matters related to the teaching process in the classroom. In addition to questions in the form of a Likert scale, EDOM also accommodates open-ended questions. However, this type of question tends not to be considered in the feedback assessment and is not explored more deeply to gain insight. Therefore, in this study text mining was carried out to find out the sentiments given by students, whether positive or negative sentiments using Convolutional Neural Network (CNN). The experimental results showed that CNN was able to achieve accuracy, precision, recall, and F1-Score of 87.95%, 87%, 78%, and 81%, respectively.

**Keywords:** Sentiment analysis · Opinion mining · Teaching evaluation · Convolutional Neural Network · Deep learning

## 1 Introduction

Evaluation is one of a series of activities in improving the quality, performance, or productivity of an institution in implementing its program. The focus of the evaluation is the individual, that is, the learning achievement achieved by the group or class. Through the evaluation will be obtained information about what has been achieved and what has not been achieved. Furthermore, this information is used to improve the program. The main function of evaluation, in this case, is to provide information that is useful for the decision-maker to determine the policy to be taken based on the evaluation that has been done [1].

In educational institution entities, one of the usual evaluations is giving feedback on evaluating lecturer performance. Feedback is a statement sent to the entity about its past behavior from which the entity can analyze future and current behavior to achieve the expected results. Feedback plays an important role in education and learning by helping to adopt new knowledge and prevent recurring mistakes. Feedback is a process

that helps organizations to monitor, evaluate, and manage the overall work environment. Good feedback practices provide useful information for organizations in improving teaching and learning experiences. At the end of the semester, in every university, it is common to conduct a lecturer evaluation process to determine the effect of lecturer teaching on students. Good teaching will certainly greatly help students to achieve good learning. The quality of teaching and academic standards certainly needs to be evaluated and improved because higher education is an important activity [2].

Quality of teaching, academic standards and student satisfaction are closely related to the match between expectations and the reality of the quality of educational services obtained from universities. If this educational institution treats students as customers by evaluating the gap between expectations and reality felt by students on the quality of educational services, the university can prepare an appropriate strategic plan to improve its quality. Evaluation of student satisfaction can be used to determine factors of quality of educational services that need to be improved, maintained, or even reduced will result in misallocation of resources such as funds, labor, and time. Incorrect allocation of resources causes ineffective efforts to improve quality and reduce student satisfaction.

Teaching and learning processes and the creation of a conducive learning atmosphere are carried out through innovative learning processes using electronic facilities/infrastructure and the latest methods in the teaching and learning process. To ensure a good teaching and learning process, regular monitoring of the implementation of learning activities is carried out, both regarding the frequency of attendance of lecturers/students and the suitability of the lecture substance discussed with the Semester Learning Plan. Evaluating and monitoring using Lecturer Evaluation Effectiveness by Students (EDOM) is conducted once per semester, in addition to evaluating lecturer performance, it is also used as a tool to evaluate how far learning targets have been met. The evaluation aims to improve the quality of learning including evaluating the way the material is delivered and the presence of lecturers, the material content of each course, student motivation and difficulties arising in the process of interaction between lecturers and students.

This feedback is distributed to students, while the result of the EDOM is intended to be information for stakeholders to improve the quality of learning. In the personal structure of lecturers, EDOM is expected to be a reflection as well as a means to improve the lecturers and develop potential of the lecturers. Whereas in the level of management/governance the results of EDOM are expected to be information that can be used as a reference in developing work programs related to improving the quality of the learning process and lecturer performance.

The feedback mechanism on EDOM, questions can be categorized in the form of scale questions (score based on Likert scale) and open-ended questions. The Likert Scale is a psychometric response scale mainly used in questionnaires to obtain participant preferences or the level of agreement with statements or sets of statements. Students are asked to indicate the level of the agreement through statements given by the ordinal scale. Likert scale-based score questions are given to students and asked to answer these questions using a ranking-based scale. These types of questions do not usually adequately represent student sentiment, while open-ended questions are sometimes not considered in feedback assessments and are not explored deeper to gain insight.

Sentiment Analysis (SA) has a significant role in various fields, including education because it can be used to assess the effectiveness of learning technology from student feedback [3]. Learning outcomes can be assessed in two ways, directly or indirectly. Direct assessment considers the results of student work such as midterm exam, final exam, assignments, quizzes, and project reports. Meanwhile, indirect assessment is based on feedback made by students about learning experiences and the quality of teaching. This indirect assessment can be done by conducting student feedback sentiment analysis. This sentiment analysis analyzes student feedback, whether it is a formal survey conducted at the end of the semester or informal comments on social media, which can be used to find out student opinions, evaluate lecturers' performance and to identify areas that can be improved through continuous improvement actions [3].

One of the most important processes in sentiment analysis is the determining of feature selection used to sentiment classifiers that classify sentiments. Algorithms in machine learning are commonly used as an approach to feature selection [4–6]. However, the approach of using machine learning in some cases still has not reached high accuracy. Accordingly, several researchers propose an approach using deep learning for improving the accuracy and computational time [7–10]. Therefore, in this study, one of the deep learning algorithms called Convolutional Neural Network (CNN) is used for sentiment analysis of student evaluations of teaching.

The remainder of this article is explained as follows: Sect. 2 explains the related work and also theoretical background. Section 3 describes the dataset used, elaborates on the results of experiments, and evaluating performance. And finally, this research is concluded in Sect. 4.

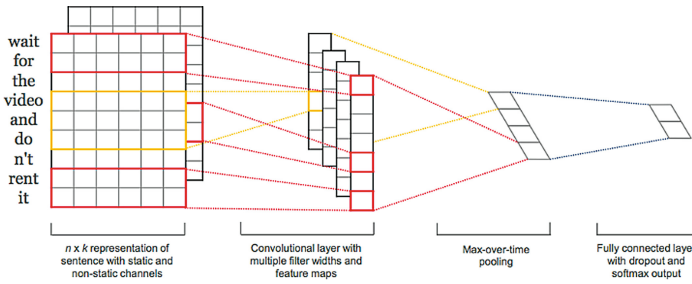
## 2 Theoretical Background

Deep learning is about learning several levels of representation and abstraction which help to understand data such as images, sounds, and text. Convolutional neural network (CNN) is one of the deep learning methods that can be applied to classify text documents. Along with the development of computing using the Graphical Processing Unit (GPU) to make the process of training the model on the CNN algorithm also becomes faster [11]. In this study, we apply the convolutional neural network method to classify texts based on positive and negative polarity which includes sentiment analysis.

### 2.1 Convolutional Neural Network (CNN)

The first layer stores words in a low dimensional vector. The next layer runs convolutions using multiple filter sizes. For example, shifting 3, 4, or 5 words at a time. Next, max-pool the results of the convolutional layer into a long feature vector, add a dropout regularization, and classify the results using the SoftMax layer [12]. In this experiment, the architectural pattern of the model is shown as in Fig. 1.

CNN is a mathematical construct that is usually composed of 3 types of layers, namely convolution, pooling, and a fully connected layer. The first two layers, convolution, and pooling do feature extraction, while the third layer, places extracted features into the final output, such as classification [13].



**Fig. 1.** The architecture of the multichannel CNN [7]

Convolution is a special type of linear operation used for feature extraction, where an array of numbers, called the kernel is implemented at the input, which is an array of numbers called tensors. Calculations are performed on each kernel matrix element and tensor input and summed to get an output called a feature map. This stage is repeated until a feature map is obtained that represents the characteristics of the tensor input. Therefore, the kernels of other sizes can have different extraction results. Convolution propagation calculations are shown in Eq. 1.

$$a_{ij}^{(k)} = \sum_{s=0}^{m-1} \sum_{t=0}^{n-1} W_{st}^{(k)} x_{(i+s)(j+t)} + b^{(k)} \quad (1)$$

where

$x$ : is the input

$a^k$ : after convolution

$k$ : kernel index (weight filter)

$W$ : kernel (weight filter)

$b$ : bias

Pooling is also called down sampling which reduces the dimensionality of the field while maintaining important information. The pooling operation that is often used is max pooling, which is capable of extracting from the filter features input filter, then selecting the highest value in each filter and discarding other values. Pooling has several types, in this case, 1D global max pooling is used. 1D Global max pooling takes a vector and calculates the maximum value of all values for each input channel.

The Fully Connected (FC) Layer is a fully connected layer of neurons. All neurons in the FC layer have a relationship with all activations in the previous layer. This layer is placed before the results of the CNN classification. The Dropout layer is also applied to CNN to overcome overfitting. The dropout technique overcomes overfitting by temporarily removing the contribution to neuron activation during the forward pass, and any weight updates are not applied to the neurons when the backward pass.

The output of feature maps from the final convolution or pooling layer is usually flattened as it is transformed into a vector dimension and connected to one/more fully connected layers (FC) or also called dense layers, where each input connected to each output has a weight. After going through feature extraction at the convolution layer and reducing the sample at the pooling layer, then the output is mapped by the FC layer to

the final output of the network, such as the probability of each class on the classification task. The final FC layer usually has the same number of nodes as the number of classes.

The output of the penultimate convolutional and pooling layers  $x$  is passed to a fully connected softmax layer. It computes the probability distribution using Eq. 2.

$$\begin{aligned} P(y = j|x, s, b) &= \text{softmax}_j(x^T w + b) \\ &= \frac{e^{x^T w_j + b_j}}{\sum_{k=1}^K e^{x^T w_k + b_k}} \end{aligned} \quad (2)$$

where  $w_k$  is the weight and  $b_k$  is the bias of the  $k$ -th class.

## 2.2 Sentiment Analysis

Sentiment Analysis is a method used to process various opinions given by consumers or experts through various media, regarding a product, service, or an agency. Sentiment Analysis is a method used to understand, extract opinion data, and process textual data automatically to get a sentiment contained in an opinion. In sentiment analysis, there are 3 types of opinions, namely positive, neutral, and negative opinions, so that the 3 types of opinions can be used by companies to determine consumer responses to a product or service [14]. But many researchers only use 2 types of sentiments, namely positive and neutral [16–18].

Sentiment analysis is one of the fields of Natural Language Processing (NLP) that builds systems for recognizing and extracting opinions in text form. Information in the form of text is currently widely available on the internet in the format of forums, blogs, social media, and sites containing reviews. The data can explain public opinion about products, brands, services, politics, or other topics. Companies, governments, and other fields then use these data to make marketing analyzes, product reviews, product feedback, and community services.

## 3 Result and Discussion

### 3.1 System Architecture

In general, the system architecture consists of three parts: the first is preprocessing for data collection, preparing data, and labeling, then the second is the formation of a data matrix as input, and the formation of a Convolutional Neural Network model as a classification feature and the third is the result of evaluation and validation. The system architecture is designed as shown in Fig. 2.

**Data Collection.** The dataset used in this study was sourced from student feedback comments on the EDOM application at Department of Information Systems, Telkom University. Preprocessing is crucial in the process of determining sentiment with CNN and the classification of tweets becomes more accurate. Preprocessing is also used to get clean data. The preprocessing stage consists of several processes namely Case Folding, Symbol Removal, Tokenization, Slang word Conversion, and Stopword Removal [19].

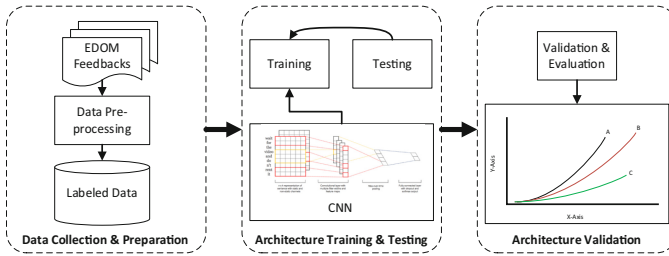


Fig. 2. System architecture

Due to existing limitations, the amount of feedback from EDOM that was successfully obtained was 1800 feedback. The feedback data is then labeled manually with positive and negative sentiments. Figure 3 illustrates the amount of feedback for each sentiment polarity, i.e. for positive sentiment is 1,300 feedback represented with 1 while negative sentiment is 420 labeled with 0.

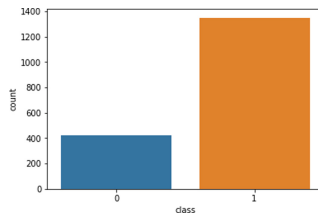


Fig. 3. Sentiment polarity distribution on the EDOM feedback dataset

**Architecture Training & Testing.** Convolutional Neural Network is a network model that accepts matrix input. The application of data in the form of images is done by using pixel numbers in the image as the contents of the matrix. This is different from text data that does not have numbers so it needs to be converted into numbers. Kim [7] conducted research by comparing the initialization of numbers used for text data on the Convolutional Neural Network method. The process of converting text into numbers is done by using pre-trained word vectors or Word2Vec.

To avoid overfitting models, a dropout layer is needed which functions to randomly choose neurons not to be used during the training process [20]. The probability of the dropout layer used in this study is 0.5. The training process continues with pooling on the feature map. The pooling method used is MaxPooling, which takes the maximum value  $\hat{c} = \max(c)$  as a feature based on a filter. The purpose of the pooling process is to get the most important features that represent other features for each feature map. The features that have been obtained from the pooling results are used for the classification process at the fully connected layer and are trained with the Backpropagation algorithm [21]. Each feature will be connected to each neuron in the fully-connected layer with independent weights and a total of 50 neurons. Each neuron from the fully-connected

layer becomes the input for the softmax layer. The number of neurons in the softmax layer is based on the number of sentiments used, 2 neurons for 2 sentiments and 3 neurons for 3 sentiments. The softmax function is used to get the probability values that have been normalized in all classes. In conducting training, an optimizer is needed to reduce the loss function or errors contained in neurons and filters. Optimizer is done by updating the neurons and filters on each iteration. In this study, Adam optimizer is used in conducting Deep Convolutional Neural Network training [22].

**Architecture Validation.** At this stage an evaluation of the architecture that has been formed. The architecture needs to be seen its performance in carrying out the desired task. There are several ways of evaluation that can be done in classification tasks on an architecture, including through accuracy, precision, recall, and F1-score. All evaluation metrics are from the confusion matrix.

**3.2 Result**

After all the designs have been completed and the model has been built, the next step is to use the model to produce dividing the dataset into 2 parts, namely training data and validation data of 80% and 20%, respectively. The next step is to test the accuracy of the model by testing data and using validation data so that performance can be determined.

In this study, we trained and evaluated convolution neural network (CNN) for several numbers of epochs and batch sizes, we trained the model using Adam optimizer [22] with a learning rate of 0.001 and batch sizes of 32, 64, 128, and 256. An early stopping technique is used to optimize the number of training iterations or epochs automatically. It has been supported by TensorFlow to determine when to stop training the model by monitoring validation errors. This method is used in order to avoid the overfitting of hyper-parameters by tuning the number of epochs. Dropout is also used to overcome overfitting on the model used. In the following Table 1 is the result of the training and validation process by considering the results of accuracy and loss with the number of epochs is 5.

**Table 1.** The result of the training and validation process with 10 epochs

No. of epoch	Loss	Accuracy	Validation loss	Validation accuracy
1	0.5692	0.7611	0.5342	0.7646
2	0.5302	0.7611	0.5041	0.7646
3	0.4723	0.7611	0.4441	0.7646
4	0.3812	0.8152	0.3583	0.8531
<b>5</b>	<b>0.2686</b>	<b>0.8975</b>	<b>0.2916</b>	<b>0.8795</b>

In the training and validation process, overfitting is observed for model experiments. Figure 4 shows the training accuracy and loss results for the CNN model used in this study. The results show that, if the error rates on the training dataset and dataset tests are low and the difference in error rates tends to be not far and stable, it means that it can be determined that the result is a good fit.

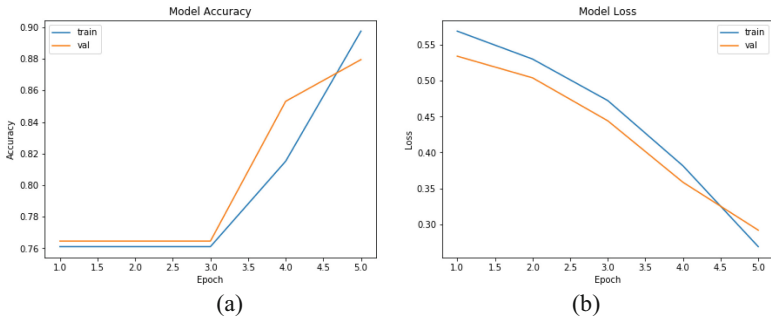


Fig. 4. Training and validation accuracy (a) and training and validation loss (b) results

Based on the experimental results for the number of epochs = 5, the performance evaluation of the results obtained is 87.95% for accuracy (as shown in Table 1). Detailed performance evaluation from the point of view of the confusion matrix can be depicted in Fig. 5.

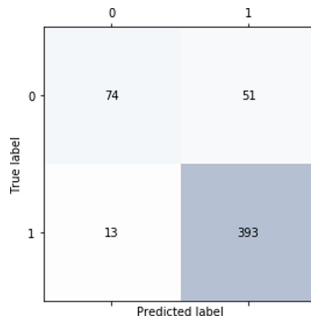


Fig. 5. Performance evaluation from the confusion matrix

From the results of the matrix confusion, other evaluation results such as precision, recall, and F1-score can be calculated. Table 2 shows the results for each positive and negative sentiment, namely class 1 and 0, respectively. While the results for macro average precision, macro average recall, and macro average F1-score are 87%, 78%, and 81%, respectively.



**Table 2.** Results detail of the performance evaluation

Class	Precision	Recall	F1-Score
0	83%	73%	77%
1	92%	95%	94%

## 4 Conclusion

The feedback mechanism on EDOM, questions can be categorized in the form of scale questions (score based on Likert scale) and open-ended questions. Likert scale-based score questions are given to students and asked to answer these questions using a ranking-based scale. These types of questions do not usually adequately represent student sentiment, while open-ended questions are sometimes not considered in feedback assessments and are not explored deeper to gain insight. Therefore, in this study text mining was carried out to find out the sentiments given by students' feedbacks, whether positive or negative sentiments using Convolutional Neural Network (CNN). To evaluate performance, accuracy, recall, precision, and F1-Score is calculated and the evaluation has shown positive results. Based on the results of experiments have shown that CNN is able to achieve accuracy, precision, recall, and F1-Score of 87.95%, 87%, 78%, and 81%, respectively. Suggestions for future research are to increase the amount of data for positive and negative sentiment and increase the accuracy of the algorithm used.

## References

1. Hunston, S.: Evaluation and organization in a sample of written academic discourse. In: *Advances in Written Text Analysis*. pp. 205–232. Routledge (2002). <https://doi.org/10.4324/9780203422656-16>.
2. Chen, Y., Hoshower, L.B.: Student evaluation of teaching effectiveness: an assessment of student perception and motivation. *Assess. Eval. High. Educ.* **28**, 71–88 (2003). <https://doi.org/10.1080/02602930301683>
3. Rani, S., Kumar, P.: A sentiment analysis system to improve teaching and learning. *Computer (Long. Beach. Calif.)* **50**, 36–43 (2017)
4. Sharma, A., Dey, S.: A comparative study of feature selection and machine learning techniques for sentiment analysis. In: *Proceedings of the 2012 ACM Research in Applied Computation Symposium*, pp. 1–7 (2012)
5. Neethu, M.S., Rajasree, R.: Sentiment analysis in twitter using machine learning techniques. In: *2013 4th International Conference on Computing, Communication and Networking Technologies (ICCCNT 2013)* (2013). <https://doi.org/10.1109/ICCCNT.2013.6726818>
6. Aninditya, A., Hasibuan, M.A., Sutoyo, E.: Text mining approach using TF-IDF and naive bayes for classification of exam questions based on cognitive level of bloom's taxonomy. In: *2019 IEEE International Conference on Internet of Things and Intelligence System (IoT&IS)*, pp. 112–117. IEEE (2019)
7. Kim, Y.: Convolutional neural networks for sentence classification. *arXiv Preprint arXiv:1408.5882* (2014)

8. Severyn, A., Moschitti, A.: Twitter sentiment analysis with deep convolutional neural networks. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 959–962. ACM (2015)
9. Vateekul, P., Koomsubha, T.: A study of sentiment analysis using deep learning techniques on Thai twitter data. In: 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), pp. 1–6 (2016)
10. Ay Karakuş, B., Talo, M., Hallaç, I.R., Aydin, G.: Evaluating deep learning models for sentiment classification. *Concurr. Comput.* **30**, 1–14 (2018). <https://doi.org/10.1002/cpe.4783>
11. Kalchbrenner, N., Grefenstette, E., Blunsom, P.: A convolutional neural network for modelling sentences. arXiv preprint [arXiv:1404.2188](https://arxiv.org/abs/1404.2188) (2014). <https://doi.org/10.3115/v1/p14-1062>
12. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* **12**, 2493–2537 (2011)
13. Yamashita, R., Nishio, M., Do, R.K.G., Togashi, K.: Convolutional neural networks: an overview and application in radiology. *Insights Imaging* **9**(4), 611–629 (2018). <https://doi.org/10.1007/s13244-018-0639-9>
14. Srivastava, A., Singh, V., Drall, G.S.: Sentiment analysis of twitter data. *Int. J. Healthc. Inf. Syst. Informatics.* **14**, 1–16 (2019). <https://doi.org/10.4018/ijhisi.2019040101>
15. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation, vol. 5, pp. 1320–1326 (2010). <https://doi.org/10.17148/ijarce.2016.51274>
16. Anstead, N., Loughlin, B.O.: Social media analysis and public opinion: the 2010 UK general election. *J. Comput. Commun.* **20**, 204–220 (2015). <https://doi.org/10.1111/jcc4.12102>
17. Li, Q., Li, S., Hu, J., Zhang, S., Hu, J.: Tourism review sentiment classification using a bidirectional recurrent neural network with an attention mechanism and topic-enriched word vectors. *Sustain* **10** (2018). <https://doi.org/10.3390/su10093313>
18. Sutoyo, E., Almaarif, A.: Twitter sentiment analysis of the relocation of Indonesia’s Capital City. *Bull. Electr. Eng. Inform.* **9**, 1620–1630 (2020). <https://doi.org/10.11591/eei.v9i4.2352>
19. Liu, B., Liu, B.: Sentiment analysis and subjectivity. In: Handbook of Natural Language Processing, 2nd edn., Taylor & Francis Group, Boca Raton (2010)
20. Srivastava, N., et al.: Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014)
21. Hecht-Nielsen, R.: Theory of the backpropagation neural network. In: *Neural Networks for Perception*, pp. 65–93. Elsevier (1992). <https://doi.org/10.1016/b978-0-12-741252-8.50010-8>
22. Kingma, D.P., Ba, J.L.: Adam: a method for stochastic optimization. In: 3rd International Conference on Learning Representations (ICLR 2015) - Conference Track Proceedings, pp. 1–15 (2014)