Sentiment Summarization Learning Evaluation Using LSTM (Long Short Term Memory) Algorithm

Achmad Yogie Setiawan[1], I Gede Mahendra Darmawiguna[2], Gede Aditra Pradnyana[3]

{ayogiesetiawan@gmail.com[1],mahendra.darmawiguna@undiksha.ac.id[2], gede.aditra@undiksha.ac.id[3]}

Informatics Engineering, Faculty of Engineering and Vocational, Universitas Pendidikan Ganesha, Singaraja, Indonesia[1,2,3]

Abstract. Lecturer learning evaluation is a text that contains student reviews related to lecturer learning performance. Learning evaluation is used as a lecturer's self-reflection material to improve the learning services provided in the next lesson. The evaluations are many in number, making it difficult for lecturers to analyze. Sentiment analysis techniques are needed to classify student evaluations. The evaluation that has been classified still leaves a long and convoluted text. Text summarization is one solution to summarize a long text into a dense and informative text. Text summarization is helpful to save time searching for the text's gist. There are two methods in text summarization, extractive and abstractive methods. This study applied an abstract method because the data used was an evaluation of lecturer learning whose reviews were written by students. The algorithm used for sentiment classification and text summarization used the Long Short Term Memory (LSTM) algorithm. The sentiment classification results were evaluated using a confusion matrix, namely testing the model with evaluation data. While the summary results were evaluated using ROUGE, which compared the summary results from the system with a manual summary by experts. In testing the confusion matrix system, the accuracy value was 0.902, and the f-measure value was 0.921. In the Recall-Oriented Understudy for Gisting Evaluation (ROGUE) test, the positive evaluation scored 0.16, and the negative evaluation scored 0.2. The developed tokenizer has not stored the tokens resulting from the training process. As a result, the prediction results when loading the model were not as good as when training was finished.

Keywords: Sentiment Analysis, Summarization, LSTM, ROGUE.

1 Introduction

Information is one of the essential needs that cannot be separated from human life. Reading is one way for humans to obtain information, such as reading books, e-mails, news, magazines, articles, journals, etc. In improving the learning performance of a lecturer, a learning evaluation is needed to determine the student's learning experience in the course. Learning evaluation is used as a lecturer's self-reflection material to improve the learning services provided in the next lesson.

Ganesha University of Education (Undiksha) provides online questionnaires for each subject through the Academic Information System, which students can fill out at the end of each semester. The online questionnaire provided at SIAK Undiksha is in the form of respondent

IConVET 2021, November 27, Singaraja, Indonesia

DOI 10.4108/eai.27-11-2021.2315533

entries in criticism and suggestions. In the questionnaire, students shared their experiences while being taught by lecturers in the subjects being taught, or also called lecturer learning evaluations. Lecturer learning evaluation is a text that contains student reviews related to the lecturer's learning performance.

The large number of students taught by lecturers in each course makes lecturers spend more time reading each evaluation. Evaluations containing criticism and suggestions make it difficult for lecturers to analyze whether students were satisfied with the learning services provided. To streamline the lecturer's time, it is necessary to have a system that can automatically assist lecturers in conducting the learning evaluation analysis process. Lecturer learning evaluation was classified into positive and negative evaluation forms, the classification is called opinion mining or sentiment analysis.

Sentiment analysis has been widely used by service or product providers as review material for the services or products they offer. Sentiment analysis is done by classifying reviews into positive and negative. Learning evaluations that have been classified into positive and negative evaluations still leave a long and convoluted text. The evaluation text makes lecturers need more time to find the existing essence. Therefore, there is a summary section in some types of text to shorten the time searching for the text's essence. Summaries are an effective way of presenting an extended essay yet concisely. Lecturer learning evaluation is one type of text that does not have a summary.

Therefore, an automatic text summarization technology is needed to simplify an extended lecturer learning evaluation information into simple information without losing the essence of the text. Automatic text summarization technology is a solution to overcome these problems. Automated text summarization is the process of generating text derived from one or more texts containing important information. Text summarization can be done in 2 ways: extractive and abstract. In this study, the researcher used an abstract method, done by making and compiling new sentences containing the essence of information from the summarized document as done by humans.

2 Sentiment summarization

2.1 Theoretical review on sentiment summarization

Sentiment Analysis is the process of automatically studying text data to generate the information contained in the sentence. Sentiment analysis is usually used to see the view of a sentence on a problem, whether the opinion is positive or negative. An example of the use of sentiment analysis is the identification of market trends and market opinions on an object of goods [1]. Sentiment Analysis is part of Natural Language Processing (NLP) science and moves on a continuum starting from the text classification stage to the stage of reviewing its polarity [2]. Text summarization is one of the branches of natural language processing (Natural Language Processing). Automatic text document summarization or automatic text summarization is a way to extract information from one or more text documents [3].

The structure of Long Short Term Memory (LSTM) is almost the same as that of the Recurrent Neural Network (RNN), which has the form of a series of repetitive modules from an artificial neural network. The difference is that the repeating module in the RNN has a single layer like the single layer Tanh. In comparison, the LSTM circuit has four interacting layers [4]. Cell state activity is controlled by layers called gates. Gates consists of a sigmoid neural network layer and a unidirectional multiplication operation so that the LSTM can delete or add information into the cell. The heart of Long Short Term Memory is state cells which are horizontal lines run across the top of the diagram. This cell state runs straight down the entire chain with some small linear interactions. This causes information to flow easily without any changes. The sigmoid layer (σ) outputs a value between 0 and 1. A value of 0 means that it does not allow any information to enter the cell, while a value of 1 means that it allows any information to enter the cell [5].

2.2 Previous works on sentiment summarization

In 2018, Ikhwan Nizwar Akhmad et al. from Sebelas Maret University created an automatic multi-document summary system using Log-Likelihood Ratio (LLR) and Maximal Marginal Relevance (MMR) for Indonesian articles. The data used in this study were an Indonesian language disease article from DokterSehat/AloDokter. The results of this study indicated that the topic signature (and its accuracy) greatly affected the results of automatic summarization with the method used. However, in this study, the extractive summary method has a weakness in forming a coherent summary. Each sentence has not formed a relationship with one another [6]. Abstractive summarization was chosen based on several factors, including the evaluation of lecturer learning in the Informatics Engineering Education Study Program, Ganesha Education University, where the dataset was written by students whose writing grammar is not guaranteed according to enhanced spelling as in official articles. Therefore, automatic summarization with an abstract method was considered suitable for use in this study.

In 2019, Rike Adelia et al. from Telkom University once made an Indonesian Abstractive Text Summarization Using Bidirectional Gated Recurrent Units. The research was conducted using the abstract method. The dataset used in this research was 500 Indonesian language journal documents from various sources. The results obtained from this study indicated that the model can learn and understand the words contained in the dataset and can produce a summary with the core words of the text. However, the weakness in this study was using the Bidirectional Gated Recurrent Unit, which results in a poor grammatical structure. The evaluation score of the two scenarios was not higher than the score of the model in English. This is caused by the size of the text and grammatical factors [7]. In 2019, Nurrohmat et al. created a system of Sentiment Analysis of Novel Review Using Long Short-Term Memory Method. The dataset used in this study was a review of Indonesian-language novels taken from the goodreads.com site. The method used in this research was the method of Long Short-Term Memory and Naïve Bayes. This study compared the LSTM method with the Naïve Bayes method based on the calculation of the values of accuracy, precision, recall, f-measure. In this study, the researchers stated that the Long Short-Term Memory method has better accuracy than the Naïve Bayes method [8].

In 2020, Lionovan et al. have created a Topic Classification and Sentiment Analysis System for University Feedback Questionnaires using the Long Short-Term Memory Method. The dataset used in this study was a feedback questionnaire at Petra Christian University. The method used in this research was the Long Short-Term Memory method. This research developed a topic classification system and sentiment analysis using Word2vec and Long Short-Term Memory. In this study, the researchers stated that the number and variation of comments could affect the average value of accuracy in the LSTM model. In addition, the researcher also mentioned that the steaming method, when carrying out the preprocessing process, can increase the average value of accuracy in the LSTM model [9]. In 2020 Yuliska, et al. had made a Literature Review of Methods, Applications, and Datasets for Automatic Text Document Summarization for Indonesian Text. In this study, the researcher stated that the summary of text documents was automatically dominated by extractive techniques. Summarizing Indonesian text documents was also dominated by unsupervised methods, while supervised methods such as machine learning and deep learning are still very rare [10]. In 2019, Alpina et al. had created a Pointer Generator and Coverage Weighting system to improve abstract summarization. The dataset used in this study was CNN/Daily Mail in English. The method used in this research was abstract. In this study, the researcher stated that the proposed model produced several summaries that were quite similar to the summaries made by experts. However, some shortcomings were identified where the summary results were not significant enough when viewed from a grammatical point of view. The summary results of this proposed model could be improved by taking into account other information in the training process[11]. In 2019, Ivanedra et al. created a system for implementing the Recurrent Neural Network Method in Text Summarization with Abstract Techniques. The dataset used in this research was 4515 English articles or news from Hindu, Indian Times, and Guardian. The method used in this research was abstract. In this study, the researcher stated that the difference in the number of datasets used as training material was very influential because more vocabulary would certainly make the program more understandable and more accurate in creating summaries. This study also identified some shortcomings where some of the summary results were still not by the context [12]. In 2018, Yoko et al. had created an Abstractive Automated Summarizing System Using a Recurrent Neural Network. The dataset used in this research was articles or news in the Indonesian language from Kompas/Detik. The method used in this research was abstract. In this study, the researcher stated that a good summary of the topic of the news summarized was similar to the topic of the news in the training data even though the words in the news are different from the training data. But in this study, there are shortcomings, namely, errors made by the system in the form of a summary with words related to the subject or location of the news [13]. In 2018, Kurniawan et al. created an INDO SUM system: A New Benchmark Dataset for Indonesian Text Summarization. The dataset used in this research was 20 thousand articles or news in English from various sources. This research produced a benchmark extractive summary dataset, namely INDOSUM. ROUGE is the evaluation standard for automated summarizing technologies. In addition, the researcher also mentioned that it is necessary to focus on developing a summary with the latest neural model for abstract summarization [14].

Based on the results of the previous research, sentiment analysis using the LSTM method was suitable for classifying learning evaluations. Furthermore, while automatic summarization using extractive methods has been widely used, the use of automatic summarization using abstractive methods was rarely used, especially those using Indonesian language datasets. Abstractive summarization was chosen based on several factors, including the dataset used in this study was the evaluation of lecturer learning in the Informatics Engineering Education Study Program, Ganesha Education University, where the dataset was written by students whose writing grammar is not guaranteed according to enhanced spelling as in official articles. In previous studies of abstraction summary such as those conducted by Adelia using the Bidirectional Gated Recurrent Unit resulted in poor grammatical structures, the evaluation score of the two scenarios was not higher than the score of the model in English. This is due to the size of the text and grammatical factors. This study developed an abstract summary using a Long Short Term Memory (LSTM) artificial neural network. LSTM is one type of Recurrent Neural Network (RNN) that has been modified by adding a memory cell so that it can store information for a long time [15].

2.3 Dataset for sentiment summarization

This study used data that has been collected in a study entitled "Development of a Sentiment Analysis System for Performance Evaluation of Ganesha University Lecturers with the Naive Bayes Method" as a dataset for the final project, namely the development of a sentiment analysis system for evaluating the performance of lecturers at the Ganesha University of Education with the Naive Bayes method. The dataset has also been labeled in the form of positive or negative sentiment on each evaluation. Researchers used the dataset by adding a new label, namely a manual summary to adjust to the developed system.

The data were an evaluation of learning at Ganesha University of Education in the period 2014 to 2018 in various departments. The total number of learning evaluations was 22,780 with five data attributes: lecturer's NIP, lecturer's name, lecturer's faculty, sentiment, and evaluation. The sentiment label has been filled in the research entitled "Development of a Sentiment Analysis System for Performance Evaluation of Ganesha University Lecturers with the Naive Bayes Method" in developing the system. The researcher took two of the five labels to adapt the dataset to the developed system, namely, sentiment and evaluation. Subsequently, several evaluations were deleted as they were considered duplicates. After deletion, 17,099 evaluations were left. Furthermore, the dataset was made with one additional label, namely a summary, making the total labels in the dataset three.

3 Proposed model and method

3.1 System flowchart

First, datasets that already have a sentiment label were given an additional label, namely summary. Furthermore, the dataset was entered into the system, and the preprocessing process was carried out through several processes, namely cleaning, case folding, stopwords, stemming, max length, and tokenization. The dataset was separated into training data and testing data with a ratio of 8 to 2.

In the sentiment analysis section, resampling was carried out for the training data so that the number of positive evaluations and negative evaluations was the same to increase the accuracy of the prediction results so that they were not biased towards data with more labels. The next steps were creating and training the model, which was followed by an evaluation of the confusion matrix. Furthermore, predictions were made to try out the model. If the prediction results issue a value equal to zero, the system would predict the sentiment in the evaluation is negative; otherwise, the sentiment is positive. In the summarization section, a model was created and the training was carried out. And then, it proceeded with inference modeling until the model could predict the summary.

The evaluations that have been carried out were sentiment classification and summarized, then separated between positive and negative evaluations. It then passed the summary string separately and detected the summary string until the string became compact. It produced two summary outputs, namely for positive and negative labels. For more details, see the following **Figure 1**.

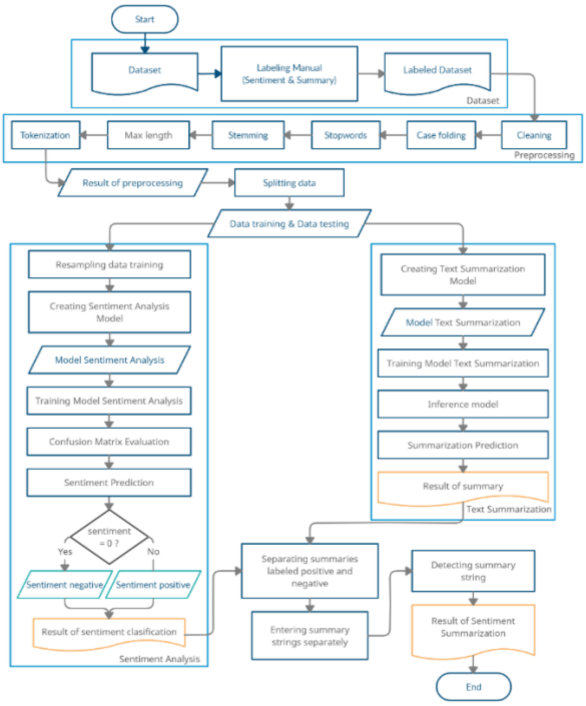

Fig 1. System flowchart.

3.2 Pre-processing

In the preprocessing stage, the first step was cleaning, which removed symbols, characters, HTML tags, text in brackets, citations, and changed abbreviations into actual words. Case folding is a process where the input text was changed to all letters into lowercase letters. The stopwords list contained common words that commonly appear. During the matching, stopwords input the words with a list of words in stopwords. If they match, the word would be deleted. Stemming is the process of changing the input word into a word into its basic word form in order to reduce the number of words contained in the token later. Max length is the process of estimating the distribution of words in the evaluation and summary. The majority result in the distribution was used as the maximum limit for sentence length. Tokenization is a process of checking the space. If there was a space, the space would be deleted and the words before and after the space were separated. The tokenization stage generated a list of tokens. Splitting the data is the process of dividing the dataset into training data with a size of 80% and testing data with a size of 20%. Training data was used for the process of training the model in order to generate predictions. Data testing was used for the evaluation process and to test the accuracy of model predictions.

3.3 LSTM flowchart

The algorithm on the LSTM layer is described as follows; first, the value of the previous cell state will enter the current cell. If the current cell was the first cell in the LSTM layer, then the previous cell state was 0. At the same time, information would enter and pass through the forget gate. This gate would be calculated with a sigmoid function where if the result was 0, then the information was forgotten, while if the result was 1, then the information would be remembered. The value of the forget gate would be multiplied by the previous cell state value to produce the cell state value. The information that passed through the input gate would be calculated with the sigmoid function. The result was multiplied by the result of the calculation of the information with the tanh function. Then, this value would be multiplied by the result value of the input gate. The multiplication result was then added with the cell state value to produce a new cell state value. The value was duplicated in two; first, it would be forwarded to the next cell and second, the two new cell state values would be calculated tanh function produces the tanh value. Information that passes through the output gate would be calculated with a sigmoid function, and the result would be multiplied by the value of tanh to produce the hidden state value in the current cell which would be passed to the next cell. The process of running the data was repeated until it reached the last cell in the LSTM layer. Information that passed through the output gate would be calculated with a sigmoid function, and the result would be multiplied by the value of tanh to produce the hidden state value in the current cell, which passed to the next cell. The process of running the data was repeated until it reached the last cell in the LSTM layer. Information that passed through the output gate was calculated with a sigmoid function, and the result was multiplied by the value of tanh to produce the hidden state value in the current cell passed to the next cell. The process of running the data was repeated until it reached the last cell in the LSTM layer. For more details, see the following **Figure 2**.

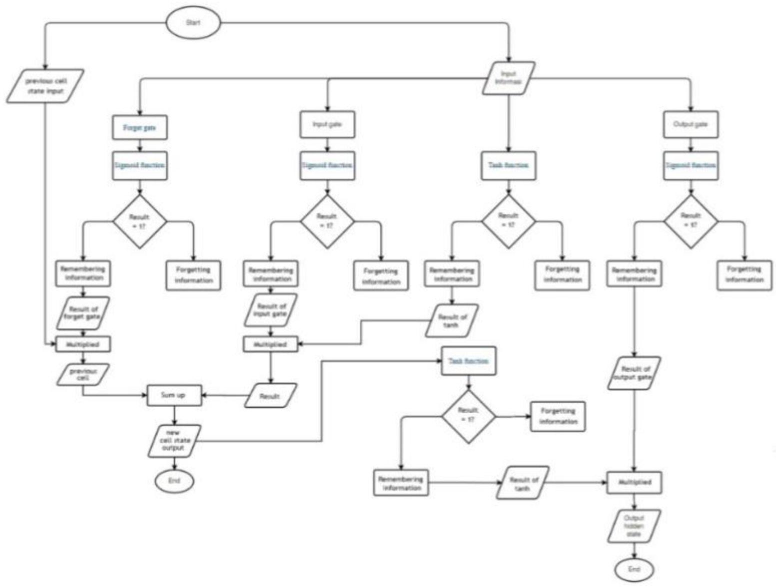

Fig. 2. LSTM flowchart.

3.4 Model sentiment analysis

The activation function used was softmax, which is a mathematical function that converts a number vector into a probability vector where each value is proportional to the relative scale of the vector value. The result of the transformation into a probability value between 0-1 and suitable for multiclass. In addition, softmax is also the only activation function that is recommended for loss function categorical cross-entropy. The loss function used was categorical cross-entropy or also called softmax loss because one sample dataset can have several classes or labels. Categorical cross-entropy generated a one-hot array containing the probability matches for each category. The optimizer used was Adam because it was suitable for solving problems related to data and parameters. In determining the weight value, Adam used the second value of the gradient. Optimizer was used to improve model accuracy during training. The metric used accuracy which calculates the percentage of system predictions (yPred) that match the actual predictions (yTrue). If the system prediction matched the actual prediction, it was considered accurate. Then, it calculated accuracy by dividing the accurately predicted number that was recorded by the total number of records. It used dropout which serves to reduce overfitting and improve model performance. The step took the overfitting model and trained its sub-models by removing units randomly for each training batch. Repeatedly eliminating random units, dropouts force units to become stronger, learning their own features, regardless of the unit.

The sentiment analysis model has two layers of dense layers. The dense layer of the first layer serves to accommodate the output of each LSTM cell which was then connected to the dense layer of the second layer, which consists of two units, each unit representing each classification label, namely positive and negative. For more details, see the following **Figure 3**.

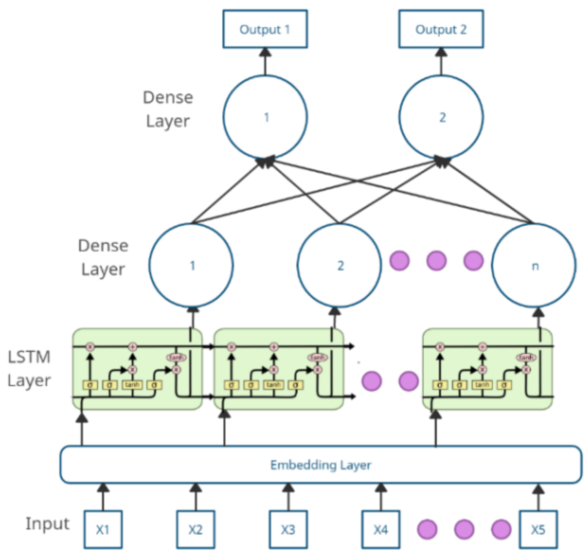

Fig 3. Model sentiment analysis.

3.5 Model summarization

In the same way that the model for Sentiment Analysis is utilized, the activation function is softmax, the metrics are accuracy, and dropout is employed. The difference is that the loss function used was sparse categorical cross-entropy because each class or label was mutually exclusive (each sample was owned by one class or label). Sparse categorical cross-entropy returned the index category integer of the most suitable category. In addition, the optimizer used was RMSProp, which is an algorithm that regulated the learning rate based on the average value of the weight. In finding the weight value, RMSProp used the first value in the gradient.

The input vector in the LSTM layer for the summary model was determined based on the maximum limit value for the number of words in the evaluation contained in the previous max length stage, which was 40. Meanwhile, the output vector was determined based on the maximum limit value for the number of words in the summary, which was 30. Same as the LSTM layer in the sentiment model, which used Keras' default gates, which consisted of 3 gates, namely forget gates, input gates, and output gates. For more details, see the following **Figure 4**.

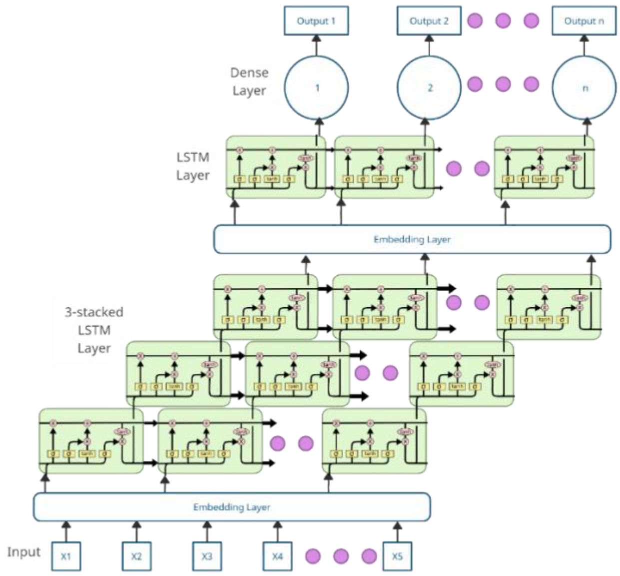

Fig 4. Model summarization

4 Results and discussion

4.2 Test confusion matrix

Table 1. Confusion matrix.

	Negative Prediction	Positive Prediction
Negative Actual	1080 (True Negative)	105 (False Positive)
Positive Actual	201 (False Negative)	1880 (True Positive)

After completing data testing from 3266 evaluations, it was found that there were 1080 evaluations predicted to have negative sentiments and matched the actual label. There were 201 evaluations predicted to be negative and did not match the actual label. Furthermore, 105 evaluations were predicted to have positive sentiments and did not match the actual label. Finally, 1880 evaluations were predicted to be positive and match the actual label. From the table data above, it can be seen the values of precision, recall, accuracy, and f-measure as follows.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Precision = \frac{1889}{1889 + 120} = 0{,}940 \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$Recall = \frac{1889}{1889 + 192} = 0{,}907 \tag{5}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$Accuracy = \frac{1889 + 1065}{1889 + 1065 + 120 + 192} = 0{,}904 \tag{7}$$

$$F - measure = \frac{2 \times precision \times recall}{precision + recall} \tag{8}$$

$$F - measure = \frac{2 \times 0{,}940 \times 0{,}907}{0{,}940 + 0{,}907} = 0{,}923 \tag{9}$$

The confusion matrix generated from the metrics library can be seen in the following table.

Table 2. Classification report.

	Precision	Recall	F1-score
Sentiment Negative	0.847255	0.898734	0.872236
Sentiment Positive	0.940269	0.907737	0.923716
Accuracy	0.904470	0.904470	0.904470

Accuracy is a measure of how accurately the model can classify correctly. Accuracy contains the ratio of correct predictions (positive and negative) to the overall data. In other words, accuracy is the level of closeness of the predicted value to the actual (actual) value. Precision measures the level of accuracy between the requested data and the prediction results provided by the model. Precision contains the ratio of correct positive predictions to the overall positive predicted results. Of all the positive classes that have been correctly predicted, how many data that are truly positive recalled a measure of the success of the model in retrieving information. Recall contains the ratio of true positive predictions compared to the total number of true positive data.

4.3 Test recall-oriented understudy for gisting evaluation (ROGUE)

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is the method for automatically evaluating the results of text summaries that are most often used. ROUGE counts the number of n-gram words that match between the system summary and the reference summary. In this study, the ROGUE trial process was carried out by providing 20 evaluations consisting of ten positive sentiment evaluations and ten negative sentiment evaluations, which were then summarized by linguists, respectively. Furthermore, the summary made by the linguist is compared with the summary generated by the system.

Table 3. Positive evaluation.

No.	Evaluation (Bahasa)	Evaluation (English)
1	Dalam mengajar di kelas, bapak sudah mengajar dengan sangat baik, yaitu dengan cara menjelaskan ataupun mencontohkan langsung di depan kelas sehingg mudah di ikuti dan di pahami oleh mahasiswa selain itu dalam mengajar sangat disiplin tetapi enjoy.	In teaching in class, you have taught very well, namely by explaining or giving examples directly in front of the class so that it is easy for students to follow and understand. Besides, your teaching is very disciplined but enjoyable.
2	Ketika mengajar, Pak Wiguna sangat sistematis ketika memaparkan materi, khususnya dalam materi OOP. Sehingga kami paham dengan materi yang diberikan meskipun kadang dibuat bingung dengan pertanyaan yang sering beliau berikan. Saran saya kepada beliau adalah lebih bersabar lagi ketika menghadapi mahasiswa yang kurang disiplin ketika berada di kelas	During his teaching, Mr. Wiguna is very systematic when presenting material, especially in OOP material. So that we understand the material given even though sometimes we are confused by the questions he often gives. My advice to him is to be more patient when dealing with students who are less disciplined in class

3	Bapak sudah mengajarkan materi dengan baik, dan juga mengajari kita disiplin. semoga kedepannya lebih ditingkatkan lagi	You have taught the material well and also taught us discipline. I hope it will be improved in the future
4	Sudah bagus cara mengajarnya, mungkin perlu ditingkatkan lagi menjadi lebih seru dan kreatif sehingga kami semakin semangat untuk belajar, terima kasih	It's a good way of teaching, maybe it needs to be improved to be more fun and creative so that we are more enthusiastic to learn, thank you
5	Saran:Selalu jadi dosen yang menyenangkan di kelas kami pak Kritik:Menurut saya bapak terlalu cepat dalam menerangkan sesuatu dimohonkan agar bapak lebih santai sehingga para mahasiswa memahaminya	Suggestion: Always be a pleasant lecturer in our class, Sir. Criticism: I think you are too fast in explaining something, I beg you to be more relaxed so that students understand it
6	Dalam proses pembelajar sudah sangat baik, dan untuk sistem uas tergolong banyak jadi dimohon untuk pengurangannya. Terimakasih	In the learning process, it has been very good, and for the exam system, it is quite large, so it is requested to reduce it. Thank you
7	Cara mengajar dikelas kami bapak sudah cukup baik dan mudah dipahami oleh mahasiswa. Bapak orangnya friendly, memberikan kami motivasi. Saran untuk kedepannya bisa lebih dekat lagi dengan mahasiswa. Semangat dosen panutanku????	The way you teach in our class is quite good and easy for students to understand. You are a friendly person, giving us motivation. Suggestions for the future can be closer to students. Keep it up, my role model teacher????
8	Cara mengajar sudah cukup baik dan menggunakan metode yg gampang dimengerti oleh mahasiswa, kedepannya hanya perlu ditingkatkan lagi agar semua mahasiswa merata bisa memahaminya????????	The teaching method is quite good and uses methods that are easily understood by students, in the future, it only needs to be improved so that all students can understand it evenly
9	Sudah cukup baik dalam mengajar dan memberi materi. Sehingga bisa dimengerti dan diikuti. Diharapkan untuk bisa lebih baik lagi kedepannya. Terimakasih	It is quite good in teaching and giving material. So that it can be understood and followed. It is hoped that it will be even better in the future. Thank you
10	Cara mengajar bapak selama ini sudah sangat baik dan Selama saya diajarkan oleh bapak saya mendapatkan ilmu baru yang bermanfaat. juga dalam pemberian materi sangat menyenangkan dan tidak membuat saya bosan. Semoga kedepannya bisa lebih baik lagi dalam pengajarannya. Terimakasih	Your way of teaching so far has been very good, and as long as I was taught, I got new useful knowledge. Also in giving the material is very fun and does not make me bored. Hopefully, in the future, he can do better in his teaching. Thank you

Table 4. Negative evaluation.

No.	Evaluation (Bahasa)	Evaluation (English)
1	Saran saya agar lebih ditingkatkan lagi dalam membawakan materi yang lebih kompleks sehingga mahasiswa bisa memahami dengan cepat,kritik saya supaya materi yang diajarkan	My suggestion is to be further improved in presenting more complex material so that students can understand quickly. My criticism is that the material being taught should be more interesting so that

	dibawakan lebih menarik agar mahasiswa semangat dalam mengikuti pembelajaran	students are enthusiastic in participating in learning
2	Cara mengajarnya sudah tersusun secara rapih, namun terkadang terlalu cepat. Sarannya sebaiknya memastikan tidak ada mahasiswa yang tertinggal dalam 1 tahapan materi.	The way of teaching has been arranged neatly, but sometimes it is too fast. The advice would be to ensure that no student is left behind in one stage of the material
3	Saran saya adalah mungkin untuk wifi di lab dasar agar lebih ditingkatkan agar saat mengerjakan uts maupun uas oracle tidak terdapat kendala mengenai jaringan internet. Kritik saya adalah menurut saya tugas project yang diberikan untuk uas terlalu banyak	My suggestion is that it is possible for the Wi-Fi in the basic lab to be further improved so that when working on UTs and oracle exams, there are no problems regarding the internet network. My criticism is that I think the project assignments given for the exam are too many
4	Memberikan lebih banyak lagi contoh-contoh latihan tentang meteri yang sedang di berikan, agar mahasiswa dapat lebih memahami materi yang di sedang di berikan	Provide more examples of exercises about the material being given so that students can better understand the material that is being given
5	Saran saya dalam melaksanakan perkuliaha agar menjelaskan materi dengan detail dan menyarankan buku untuk pemblajaran	My advice in carrying out lectures is to explain the material in detail and suggest books for learning
6	Saran : pada saat proses pembelajaran sedang berlangsung interaksi antara peserta didik harus sering dilakukan guna mempermudah peserta didik dalam memahami materi. Dan jika peserta didik belum paham akan materi tersebut agar dilalukan tanya jawab agar peserta didik dengan baik memahami materi yang diajarkan. Kritik : pada saat proses pembelajaran yang berlangsung sudah baik. Akan tetapi saat proses pembelajaran berlangsung peserta didik lebih di kontrol seperti melakukan tanya jawab, mengajukan pertanyaan apabila ada yg belum dimengerti, agar peserta didik bisa memahami materi dengan baik.	Suggestion: when the learning process is in progress, the interaction between students should be done frequently to make it easier for students to understand the material. And if students do not understand the material so that questions and answers are carried out so that students understand the material being taught. Criticism: the learning process was conducted well. However, when the learning process takes place, students are more controlled, such as conducting questions and answers, asking questions if there is something that has not been understood so that students can understand the material well.
7	Cara mengajarnya sudah tersusun secara rapih, namun terkadang terlalu cepat. Sarannya sebaiknya memastikan tidak ada mahasiswa yang tertinggal dalam 1 tahapan materi.	The way of teaching has been arranged neatly, but sometimes it is too fast. The advice should be to ensure that no student is left behind in one stage of the material
8	Kritik saya pada mata kuliah ini adalah dosen masih sedikit meluangkan waktunya untuk memberikan kesempatan kepada mahasiswa untuk berdiskusi. Kemudian terdapat batasan antara hubungan dosen dengan mahasiswa. Saran saya	My criticism of this course is that the lecturers still take a little time to provide opportunities for students to discuss. Then there are boundaries between the relationship between lecturers and

	sebelum memulai perkuliahan dosen harus nya menyusun rancangan perkuliahan terlebih dahulu dimana dosen seharusnya memberikan waktu lebih banyak kepada mahasiswa untuk mendiskusikan tgs.	students. My advice before starting lectures is that the lecturer should prepare a lecture plan first where the lecturer should give students more time to discuss assignments
9	Om Swastyastu, Saya ingin memberi sedikit saran kepada bapak pada saat dikelas terkadang terlalu cepat menjelaskan dan saya kadang sulit untuk mengikutinya. Jadi untuk kedepannya mungkin bisa lebih dipelankan lagi untuk pengajaran dikelas. Sisanya sudah cukup baik menurut saya, cukup mudah dimengerti namun terkadang juga membuat bingung dikelas karena terlalu cepat menurut saya.	Om Swastyastu, I want to give you a little advice when in class, sometimes it is too fast to explain and sometimes it is difficult for me to follow it. So in the future, it might be slower for teaching in class. The rest is good enough in my opinion, quite easy to understand but sometimes it also confuses the class because it's too fast in my opinion.
10	Saran dan kritik saya terhadap dosen saya yang mengajar mata kuliah Pemrograman Berorientasi Objek yaitu bapak I Gede Mahendra Darmawiguna, S.Kom.,M.Sc. adalah yang pertama saran saya adalah bapak supaya lebih tegas dalam mengajar ke mahasiswanya dan kritk saya adalah ketika bapak mengajarkan materi tentang sesuatu diharapkan bapak mengajarkannya secara pelan-pelan dan jangan cepat-cepat agar supaya mudah dimengerti oleh para mahasiswanya	My suggestions and criticisms towards my lecturer who teaches Object-Oriented Programming courses, namely Mr. I Gede Mahendra Darmawiguna, S.Kom., M.Sc. The first is that my advice is for you to be more assertive in teaching your students and my criticism is that when you teach material about something, it is hoped that you teach it slowly and don't rush so that it is easily understood by the students.

Table 5. Test ROGUE on positive evaluation.

	Model	**Reference**
Bahasa	pembelajaran yang disampaikan materidiajarkan banyak contoh dirasa sebaiknya lebih tegas dan agar mahasiswa aktif dengan baik	Cara mengajar sudah baik, mudah dipahami, dan menyenangkan. Selain itu, bapak disiplin dan friendly sehingga mahasiswa termotivasi untuk belajar
English	The learning delivered by the material is taught by many examples. It is felt that it should be more assertive and so that students are active properly.	The way of teaching is good, easy to understand, and fun. In addition, the father is disciplined and friendly so that students are motivated to learn.

$$ROUGE - 1 = \frac{3}{18} = 0{,}16$$

In the negative evaluation, only 3 out of 18 words were the same between the summary prediction made by the system and the summary made by the linguist. The same words include "bagus," "dan," and "mahasiswa," or in English, "good, "and," and "student."

The small ROGUE value was obtained because the summary writing technique was different between the researcher in the dataset trained in the model and the summary made by the linguist. The difference in the summarizing technique makes the same number of words slightly smaller, resulting in a small ROGUE value.

Table 6. Test ROGUE on Negative Evaluation

	Model	Reference
Bahasa	sebaiknya sudah harus tegas lebih terutama proses pembelajaran metode mengajar lagi dalam kritik dan slide banyak saran kedisiplinan	Penjelasan materi terlalu cepat, kurang diskusi, dan terlalu banyak memberikan tugas
English	it's better if you have to be more assertive, especially the learning process of teaching methods again in the criticism and slides, there are lots of disciplinary suggestions	Explanation of material is too fast, lack of discussion, and giving too many assignments

$$ROUGE - 1 = \frac{2}{10} = 0{,}2$$

In the positive evaluation, only 2 out of 10 words were the same between the summary prediction made by the system and the summary made by the linguist. The same words include “banyak” and “dan” or in English “a lot” and "and."

5 Conclusion

In testing the confusion matrix system, the accuracy value was 0.902, and the f-measure value was 0.921. This value means that the system was able to produce sentiment predictions that were in accordance with the training data. In the Recall-Oriented Understudy for Gisting Evaluation (ROGUE) test, the positive evaluation scored 0.16, and the negative evaluation scored 0.2. This value means that the system cannot generate summary predictions that match the expert's summary optimally.

The small ROGUE value obtained was also caused by the fact that the number of epochs during the training of the summarization model was still small, namely, only 50, with the result that the accuracy value was only 0.8993 and the loss was 0.3404. The size of the dataset makes the training process take a long time. In this study, researchers with limited resources and hardware were only able to do training with 50 epochs. The developed tokenizer has not been able to store tokens from the training process. If the system was reopened by uploading the stored model, then the system would have difficulty in determining the token for each word. These problems have an impact on the summary prediction process carried out by the summarization model, where user input must pass through the tokenizer before it can be processed. As a result, the tokens on each word became random and did not match what the previous model had learned.

In future research, it will be better to adjust the summary technique with experts who become references, so that the model can predict the choice of the same important sentence as the reference. Also, it is expected to conduct the test with a larger number of datasets because the more data used, the better the system will be in studying the data patterns used. Therefore, the predictions generated by the system will be more accurate. The summary labeling for training data can be done by the same person or technique as during the ROGUE trial; thus, the writing style is more similar. The performance of models should be compared with different parameters to find the best model that can produce the most accurate predictions. It can also analyze other algorithms that can be better at predicting sentiment classification and learning evaluation summaries. In addition, it can also develop a tokenizer that can store tokens after the training process so that the token value does not change and can be reused without having to repeat the training process. The system prediction results can also be compared using other deep learning frameworks such as PyTorch, MXNet, or ONNX.

References

[1] Rozi, I. F., Pramono, S. H, Dahlan, E. A., “Implementasi Opinion Mining (Analisis Sentimen) untuk Ekstraksi Data Opini Publik pada Perguruan Tinggi,” in Journal EECCIS, 2012.

[2] Kouloumpis, E., Wilson, T., Moore, J., “Twitter Sentiment Analysis: The Good the Bad and the OMG!,” in ICWSM, 2011.

[3] Yuliska. 2020. “Literatur Review Terhadap Metode, Aplikasi dan Dataset Peringkasan Dokumen Teks Otomatis untuk Teks Berbahasa Indonesia”. IT Journal Research and Development (ITJRD). Vol.5, No.1.

[4] Sanjeevi, M. 2018. “Chapter 10.1: DeepNLP — LSTM (Long Short Term Memory) Networks with Math.”. https://medium.com/deep-math-machine-learning-ai/chapter-10-1-deepnlp-lstm-long-short-term-memory-networks-with-math-21477f8e4235. Accessed on January 22, 2021 at 09.05 WITA.

[5] Colah. 2015. “Understanding LSTM Networks”. https://colah.github.io/posts/2015-08-Understanding-LSTMs/. Accessed on January 22, 2021 at 12.32 WITA.

[6] Akhmad, I. N. 2018. “Peringkasan Multidokumen Otomatis dengan Menggunakan Log-Likelihood Ratio (LLR) dan Maximal Marginal Relevance (MMR) untuk Artikel Bahasa Indonesia”. Journal Linguistik Komputasional. Vol.1, No.1. Available at http://inacl.id/journal/index.php/jlk/article/view/6/7. Accessed on December 15, 2020.

[7] R. Adelia, S. Suyanto and U.N. Wisesty, “Indonesian Abstractive Text Summarization Using Bidirectional Gated Recurrent Unit,” International Conference on Computer Science and Computational Intelligence, 2019

[8] Nurrohmat, M. A., “Sentiment Analysis of Novel Review Using Long Short-Term Memory Method,” in IJCCS (Indonesian Journal of Computing and Cybernetics Systems), 2019.

[9] Lionovan, D. A., 2020. “Klasifikasi Topik dan Analisa Sentimen Terhadap Kuesioner Umpan Balik Universitas Menggunakan Metode Long Short-Term Memory”. Journal Infra. Vol.9, No.2. Available at http://publication.petra.ac.id/index.php/teknik-informatika/article/view/10512/9355. Accessed November 19, 2020.

[10] Y. Yuliska and T. Sakai, “A Comparative Study of Deep Learning Approaches for QueryFocused Extractive Multi-Document Summarization,” in International Conference on Information and Computer Technologies, 2019, p.p. 153-157.

[11] Alpiani, A. S. 2019. “Pointer Generator dan Coverage Weighting untuk Memperbaiki Peringkasan Abstraktif”. Ind. Journal on Computing. Vol.4, No.2. Available at https://socj.telkomuniversity.ac.id/ojs/index.php/indojc/article/view/341/148. Accessed on December 15, 2020.

[12] Ivanedra, K. 2019. “Implementasi Metode Recurrent Neural Network Pada Text”. Journal Teknologi Informasi dan Ilmu Komputer (JTIIK). Vol.6, No.4. Available at http://jtiik.ub.ac.id/index.php/jtiik/article/view/1067. Accessed on December 15, 2020.

[13] Yoko, K. 2018. “Sistem Peringkas Otomatis Abstraktif Dengan Menggunakan Recurrent Neural Network”. Journal of Computer Science and Information Systems. Vol.2, No.1. Available at https://journal.untar.ac.id/index.php/computatio/article/view/1481. Accessed on December 15, 2020.

[14] Kurniawan, K., Louvan, S. 2018. INDO SUM: A New Benchmark Dataset for Indonesian Text Summarization. Proceedings of 2018 International Conference on Asian Language Processing (IALP), Bandung: 15-17 November 2018.

[15] Zaman, L., Sumpeno, S., Hariadi, M, “Analisis Kinerja LSTM dan GRU sebagai Model Generatif untuk Tari Remo,” in JNTETI, 2018.