Journal of Research Administration
Society of Research Administrators International

MODEL DEVELOPMENT OF AUTOMATIC VALIDATION DETECTION ON SURVEY RESPONSES OF COURSE LEARNING PROCESSES USING BERT-BASED MODEL

[1] Irvan Santoso, [2] Edi Abdurachman, [3] Harco Leslie Hendric Spits Warnars, [4] Lili Ayu Wulandhari

[1] Computer Science Department, BINUS Graduate Program – Doctor of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480, [2] Computer Science Department, BINUS Graduate Program – Doctor of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480, [3] Computer Science Department, BINUS Graduate Program – Doctor of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480, [4] Computer Science Department, BINUS Graduate Program – Doctor of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480

[1] isantoso@binus.edu, [2] ediabdurachman@gmail.com, [3] Spits.hendric@binus.ac.id, [4] lili.wulandhari@binus.ac.id

***Abstract*—** *Evaluation of the course learning process is one of the important actions that must be carried out by educational institutions. Evaluation must be carried out thoroughly and in depth on what is felt by students. One form of evaluation that is usually given is in the form of a survey consisting of closed and open-ended questions. The process required to check the validity of student responses is quite time consuming if done manually. Therefore, in this study a model will be proposed to automatically check validation of survey responses to the course learning process using a BERT-based model. The dataset used is survey data taken from one of the educational institutions in Indonesia by considering the relation between closed and open-ended questions towards the same goal. The results obtained from this study are that all BERT-based models can provide a good accuracy value of above 80% in checking the validity of the responses given by students, when compared to manual judgement.*

Index Terms— *Sentiment Analysis, BERT, Course Learning, Education*

INTRODUCTION

Evaluation is an action that is needed and very important in ensuring that something proposed or done has gone according to a predetermined plan or design [1]. Evaluation includes the collection of abundant information or data that has relevance to an object to be examined [2]. The data collection process is followed by the process of converting it into understandable values to be processed by systems or tools that facilitate the evaluation process [3]. In the business world, evaluation is one way to improve the performance of a company or institution so that it can compete and be better than other similar companies or institutions [4]. This also applies to educational institutions, where each educational institution wants to display better selling points in order to achieve maximum intake according to its capacity [5]. One of the important factors in

educational institutions is the learning process which is basically the core of an educational institution and needs to be evaluated [6]. This is done to maintain the consistency and growth of educational institutions regarding technological developments and learning services provided to their students [7].

To obtain appropriate data on what students feel about the learning they receive, each educational institution usually provides a survey of the course learning process that has been experienced [8]. Survey questions will usually cover the course material provided; the way the lecturer teaches; the facilities provided; and others related to the needs of each educational institution [9]. In addition, there are generally two forms of questions given to students, namely closed questions and open-ended questions [10]. Closed questions will usually be in the form of a Likert scale or the like which is given to determine the value of satisfaction felt by students with the learning process while open-ended questions give freedom to students to be able to provide various responses in the form of free text [11]. Student responses to open-ended questions are quite difficult to examine because the sentiments contained in a sentence can vary and require the same understanding in order to give the right value [12].

With a large number of students and the growth in the number of existing intakes each year, checking the validity of the survey on the learning process of this subject becomes difficult to do [13]. The reason is, to ensure the value given by closed questions is the same as open-ended questions is a different process from checking the sentiment value of open-ended questions and this requires more time for educational institutions to work on it. Therefore, in this study an automatic model will be proposed in detecting the validation of responses to the learning process. Responses to closed questions and open-ended questions will be processed and translated into values that can be compared to check the suitability of their values for questions that have the same purpose. For example, if the response to a closed question related to the lecturer's way of teaching has a good or positive value, but in an open-ended question response regarding the same matter it has the opposite value, then it will be categorized as an inappropriate or invalid value. In addition, the model to be developed will use an approach based on BERT as the main model.

Previous Study

Research related to BERT was started by several researchers, namely Devlin, Chang, Lee, and Toutanova [14] with the title "Bert: Pre-training of deep bidirectional transformers for language understanding.". BERT or Bidirectional Encoder Representations from Transformers is a model developed with a deep learning approach. BERT uses transformers, an attention mechanism that studies contextual relations between words in a text so that it allows the system to know the meaning of a sentence. There are several follow-up studies conducted by several other researchers related to the BERT model. One of them is research conducted by Lan, et al. with the title “Albert: A lite bert for self-supervised learning of language representations [15]. The Albert model overcomes difficulties encountered during further model upgrades, due to limited GPU/TPU memory and longer training times. The study proposes two parameter reduction

techniques to reduce memory consumption and increase BERT training speed. Comprehensive empirical evidence shows that the proposed method leads to a much better scalable model compared to the BERT model.

Another study has been conducted by Sanh, Debut, Chaumond, and Wolf with the title "DistillBERT, a distilled version of BERT: smaller, faster, cheaper and lighter" [16]. In this study, DistilBERT produced good performance on a variety of tasks with larger data. The training shows that the size of the BERT model can be reduced by up to 40%, while retaining 97% of language comprehension ability and being 60% faster. Furthermore, Liu, et al. also conducted research on the development of BERT with the title "Roberta: A robustly optimized bert pretraining approach" [17]. In this paper, a BERT pre-training replication study is presented that carefully quantifies the impact of many key hyperparameters and training data measures that ultimately found that the BERT was significantly trained and could match or exceed the performance of every model published thereafter. The model developed uses and pays attention to the previous design model that was produced. In addition, to cover existing research using Indonesian, Wilie, et al. conducted research entitled "IndoNLU: Benchmarks and resources for evaluating Indonesian natural language understanding" which resulted in IndoBERT. IndoBERT is adjusted and optimized with data which is a collection of Indonesian sentences so that it can produce better values in detecting sentences based on Indonesian.

Data Collection

This research will use data from one of the educational institutions in Indonesia which has conducted a survey of the course learning process at its institution. The data used consisted of 32 questions in the form of closed and open-ended questions with four categories of relations, including the effectiveness and benefits of learning; attitude and way of teaching lecturers; the effectiveness of online and offline lectures; and learning media used. An example of one of the relations is described in Table 1. The amount of data used is 15,010 responses given by students which will later be divided into 80% versus 20% for the training and testing process of the system that will be developed.

Table 1. Examples of Relation Closed Questions and Open-ended Questions Related to the Effectiveness and Benefits of Learning

Number	Question Form	Question
1	Closed Question	*Available course materials are clear and concise.*
2	Closed Question	*The course is well organized.*

3	Closed Question	*The course allows adequate development of subject knowledge to achieve its learning objective.*
4	Closed Question	*The lecturer is open-ended towards student's perspectives or ideas.*
31	Open-ended Question	*Your opinion about the benefit of taking this course.*
32	Open-ended Question	*Your objective in taking this course.*

Based on Table 1, this is just one example of the relation that exists in an educational institution. If using data from other educational institutions, there is a possibility that the relations that need to be processed will be different and require adjustments to the analysis. The responses given by students regarding these relations will be processed to compare the responses to closed questions and open-ended questions. If in the process there is a discrepancy, then the system will process it by labelling it invalid whereas if it turns out that the value given matches, then the system will label the response as valid.

As an illustration, the responses given by students to the course learning process will be explained in Table 2 which will later be processed as an example during pre-processing.

Table 2. Examples of Responses Given to Open-Ended Questions

Responses	**Meaning**
Sudah cukup baik saya rasa	It's good enough I think
Perbanyak materi di saat GSLC	Expand material at GSLC
Penjelasan materi sudah cukup jelas,	The explanation of the material is clear

tetapi menurut saya kurang banyak praktik sehingga saya sering lupa akan materi yang telah dibahas	enough, but in my opinion there is not much practice so that I often forget about the material that has been discussed
Cukup baik, hanya saja soalnya sulit sulit	It's good enough, it's just that the task is difficult
Sebenarnya sudah bagus, tapi sepertinya masih bisa lebih baik lagi	Looks like it could be better
tidak terlalu efektif	Not very effective
Kurang baik	Not good
kurang bagus sih karena harus belajar sendiri	It's not good because you must learn it yourself
kadang2 agak susah belajar sendiri	Sometimes it's a bit difficult to study alone
Tugas yang diberikan terkadang tidak relevan dengan materi yang diajarkan (offEMPTYtopic)	The assignments given are sometimes irrelevant to the material being taught
TugasEMPTYtugas yang diberikan dalam GSLC tidak sesuai dengan yang saya inginkan. Dosen memberikan tugas forum yang kurang membahas mengenai Cyber Security.	The assignment given in GSLC was not what I expected. Lecturers give forum assignments that are less related to Cyber Security.

Table 2 is only an example of some of the data used in pre-processing later. The responses given by students described what they felt in undergoing the learning process of existing courses. There are several sentences related to GSLC that emerge from the responses given. GSLC stands for Guide Self Learning Class which is a learning method in one of the educational institutions in

Indonesia. GSLC is carried out online by students and lecturers in charge of the course. Lecturers will provide teaching materials to students through several media that have been provided by the educational institution.

Methodology

In accordance with the objectives of the research conducted, it takes several processes to be built to be able to detect the validity of the responses given to closed and open-ended questions. The flow to be developed will be explained in Fig. 1. Sequentially.

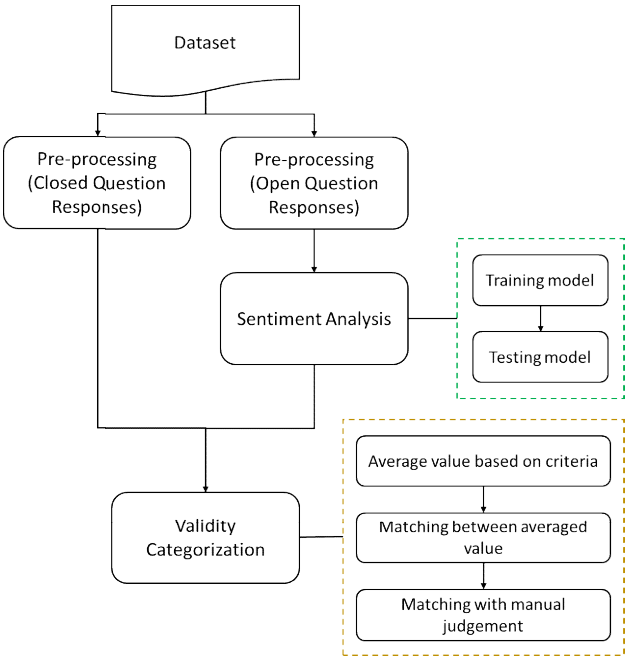

Fig 1. Flow in Checking Response Validity

Based on Fig 1, the process that will be passed first is pre-processing for each response to the question. Pre-processing of closed question responses is only done by categorizing the given Likert scale value, such as if the value is between 1 to 3, it will be categorized as a negative value and values 3 to 6 will be categorized as a positive value. The pre-processing of open-ended question responses will be different from closed question responses because the form of the response given is not a score but free text so that before sentiment analysis is carried out at a later stage, a normalization process is needed first, such as punctuation removal, case folding, spelling correction, and translation to Indonesian for all foreign languages.

Furthermore, sentiment analysis will be carried out by applying the BERT, Albert, DistilBERT, Roberta, and IndoBERT approaches. The approaches or all these models will produce the best model at the training stage which will be used as a model in the testing process to produce negative and positive values. Sentiment analysis results from existing models will be compared to get the highest accuracy value that can be used as the main model in checking

sentiment values. Then in order to be able to compare the values of closed question responses in the form of a Likert scale, a process of converting positive and negative values into numbers according to the criteria for closed question response categories was also carried out. Furthermore, each response value will be averaged, then compared with each other to determine the validity of each response given. If in the matching process it is known whether there is a discrepancy or not the same, then the response will be said to be invalid and vice versa. Then, to ensure the system runs according to the needs and expected output, the final results obtained will automatically be compared with the results obtained based on manual judgment of all existing responses.

Result and Discussion

Implementation is done with the Python programming language based on a predetermined methodology. Each result obtained from each stage will be used as input in the next process until the process of categorizing the validation of the responses given by students. The results of the pre-processing of open-ended questions can be seen in Table 3 which explains the results of punctuation removal and case folding.

Table 3. Example Results of Punctuation Removal and Case Folding

Responses	**Punctuation Removal and Case Folding**
Sudah cukup baik saya rasa	sudah cukup baik saya rasa
Perbanyak materi di saat GSLC	perbanyak materi di saat gslc
Penjelasan materi sudah cukup jelas, tetapi menurut saya kurang banyak praktik sehingga saya sering lupa akan materi yang telah dibahas	penjelasan materi sudah cukup jelas tetapi menurut saya kurang banyak praktik sehingga saya sering lupa akan materi yang telah dibahas
Cukup baik, hanya saja soalnya sulit sulit	cukup baik hanya saja soalnya sulit sulit
Sebenarnya sudah bagus, tapi sepertinya masih	sebenarnya sudah bagus tapi sepertinya masih

bisa lebih baik lagi	bisa lebih baik lagi
tidak terlalu efektif	tidak terlalu efektif
Kurang baik	kurang baik
kurang bagus sih karena harus belajar sendiri	kurang bagus sih karena harus belajar sendiri
kadang2 agak susah belajar sendiri	kadang agak susah belajar sendiri
Tugas yang diberikan terkadang tidak relevan dengan materi yang diajarkan (offEMPTYtopic)	tugas yang diberikan terkadang tidak relevan dengan materi yang diajarkan offemptytopic
TugasEMPTYtugas yang diberikan dalam GSLC tidak sesuai dengan yang saya inginkan. Dosen memberikan tugas forum yang kurang membahas mengenai Cyber Security.	tugasemptytugas yang diberikan dalam gslc tidak sesuai dengan yang saya inginkan dosen memberikan tugas forum yang kurang membahas mengenai cyber security

Furthermore, the pre-processing results for spelling correction and translation into Indonesian for all foreign languages can be seen in Table 4.

Table 4. Example Results of Spelling Correction and Translation

Punctuation Removal and Case Folding	**Spelling Correction and Translation**
sudah cukup baik saya rasa	sudah cukup baik menurut saya
perbanyak materi di saat gslc	perbanyak materi pada waktu gslc
penjelasan materi sudah cukup jelas	penjelasan materinya cukup

tetapi menurut saya kurang banyak praktik sehingga saya sering lupa akan materi yang telah dibahas	jelas tapi menurut saya kurang banyak latihannya sehingga saya sering lupa dengan materi yang sudah dibahas
cukup baik hanya saja soalnya sulit sulit	cukup bagus hanya saja soalnya susah susah
sebenarnya sudah bagus tapi sepertinya masih bisa lebih baik lagi	sebenarnya sudah bagus tapi sepertinya masih bisa lebih baik
tidak terlalu efektif	tidak terlalu efektif
kurang baik	tidak baik
kurang bagus sih karena harus belajar sendiri	itu tidak baik karena anda harus belajar sendiri
kadang agak susah belajar sendiri	terkadang agak sulit untuk belajar sendirian
tugas yang diberikan terkadang tidak relevan dengan materi yang diajarkan offemptytopic	tugas yang diberikan terkadang tidak relevan dengan materi yang diajarkan dari topik kosong
tugasemptytugas yang diberikan dalam gslc tidak sesuai dengan yang saya inginkan dosen memberikan tugas forum yang kurang membahas mengenai cyber security	tugas yang diberikan di gslc tidak seperti yang saya inginkan dosen memberikan tugas forum yang tidak membahas tentang keamanan caber

Based on Table 3 and Table 4, the process carried out at the pre-processing stage of open-ended question responses can be carried out by the system properly. Therefore, the process will continue with the sentiment analysis process, where all pre-processed responses will be processed to obtain the sentiment value. Examples of the results obtained from the sentiment analysis process can be seen in Table 5 which explains the sentiment value per example of the pre-processed sentence.

Table 5. Example Results of Spelling Correction and Translation

Pre-Processing Result	**Sentiment Value**
sudah cukup baik menurut saya	Positive
perbanyak materi pada waktu gslc	Positive
penjelasan materinya cukup jelas tapi menurut saya kurang banyak latihannya sehingga saya sering lupa dengan materi yang sudah dibahas	Negative
cukup bagus hanya saja soalnya susah susah	Positive
sebenarnya sudah bagus tapi sepertinya masih bisa lebih baik	Positive
tidak terlalu efektif	Negative
tidak baik	Negative
itu tidak baik karena anda harus belajar sendiri	Negative
terkadang agak sulit untuk belajar sendirian	Negative
tugas yang diberikan terkadang tidak relevan dengan materi yang diajarkan	Negative

dari topik kosong	
tugas yang diberikan di gslc tidak seperti yang saya inginkan dosen memberikan tugas forum yang tidak membahas tentang keamanan caber	Negative

As a comparison, Table 6 will display the accuracy and time obtained for all models at the training stage and Table 7 will display the accuracy and time obtained for all models at the testing stage.

Table 6. Training Result

Model	**Accuracy**	**Time**	**Dataset**
BERT	87,82	01:01:10	12.008
Albert	83,61	00:48:24	12.008
DistilBERT	86,66	00:29:14	12.008
Roberta	83,23	01:03:01	12.008
IndoBERT	91,24	00:40:41	12.008

Table 7. Testing Result

Model	**Accuracy**	**Time**	**Dataset**
BERT	92,25	00:13:22	3.002
Albert	90,25	00:12:78	3.002
DistilBERT	91,07	00:11:76	3.002
Roberta	89,82	00:13:11	3.002
IndoBERT	93,81	00:12:03	3.002

Based on Table 6 and Table 7, IndoBERT has a better accuracy value when compared to other models or approaches, namely 91.24 for training and 93.81 for testing. While the second model that has a good level of accuracy is BERT with a value of 87.82 for training and 92.25 for testing. When viewed based on the time of testing, there is no significant difference. This is because all models no longer carry out the training process which does take a long time depending on the process per proposed model. In the training process, the DistilBERT model or approach can produce results much faster than other models, and even provide good accuracy. Process-wise, there is a possibility that the accuracy may change depending on the amount of

data and the form of the data being processed. This also does not rule out the possibility that the processing time of each model may change.

Furthermore, after obtaining sentiment values from each model, the process will proceed to the next stage, namely matching between the results of processing closed question responses and the processing results of open-ended responses. At this stage, matching will still be carried out for all models, not only for the model with the best accuracy or time. If the process has been carried out successfully and it is known how many are valid or invalid, then a check or match will be carried out against the results of the manual judgment. The goal is to find out whether the system with the model that has been developed can provide results that are in accordance with research expectations or not. The results of the accuracy of each model after matching with manual judgment will be shown in Table 8.

Table 8. Validity Categorization Accuracy

Model	Accuracy
BERT	85,67
Albert	82,87
DistilBERT	86,64
Roberta	86,58
IndoBERT	87,49

Based on Table 6, IndoBERT continues to provide the best accuracy value in determining the validity of the responses given by students to the course learning process that has been followed with an accuracy value of 87.49. However, when compared to other models or approaches, they are actually not as different as in the testing phase. All models can provide accuracy values above 80% which in fact can produce good values. There is a possibility that IndoBERT can provide a better value compared to other models because the majority of the responses given by students are in the form of Indonesian, where IndoBERT is a model that specializes in processing data in Indonesian.

Conclusion

Education is one of the main keys in the development of strong human resources. Educational institutions are competing in providing and improving their services to make this happen. Every education service provided must be evaluated and become an important key in the development of education itself. However, in the evaluation process that occurs, it takes quite a long process to ensure the data obtained is correct so that it can provide the right decision making for the next development stage. Therefore, this research proposes a model that can be used by educational institutions to speed up the processing or checking of the required data related to the learning process services provided. Implementation is carried out using several models or approaches that have been studied by previous researchers, namely BERT, Albert, DistilBERT, Roberta, and

IndoBERT. The results obtained from each model in its processing are considered good, because it can provide high accuracy values from the training, testing, to matching to manual judgment processes.

For information, the data used is data collected from surveys given to students regarding the learning process of courses in one of the educational institutions in Indonesia. Because the majority of responses given were Indonesian for open-ended questions, IndoBERT was able to provide a better score compared to other models that did not focus on Indonesian. In addition, not all educational institutions have the same pattern of questions so that if the data used comes from other educational institutions, adjustments to the data process are needed, including determining the relation between responses to closed questions and responses to open-ended questions.

REFERENCES

H. Paulheim, "Knowledge graph refinement: A survey of approaches and evaluation methods," *Semantic Web*, vol. 8, no. 3, pp. 489–508, 2016.

K. Järvelin and J. Kekäläinen, "IR evaluation methods for retrieving highly relevant documents," *ACM SIGIR Forum*, vol. 51, no. 2, pp. 243–250, 2017.

T. Grizane and I. Jurgelane, "Social Media Impact on business evaluation," *Procedia Computer Science*, vol. 104, pp. 190–196, 2017.

H. Liao and J. Hitchcock, "Reported credibility techniques in higher education evaluation studies that use qualitative methods: A research synthesis," *Evaluation and Program Planning*, vol. 68, pp. 157–165, 2018.

C. A. Duncan and B. Kern, "Getting competition under control," *Journal of Physical Education, Recreation & Dance*, vol. 91, no. 2, pp. 33–41, 2020.

E. Bardelli, "Teacher evaluation systems: Measures of instructional effectiveness or mechanisms of structural bias?," *Proceedings of the 2022 AERA Annual Meeting*, 2022.

I. Santoso, W. Suparta, A. Trisetyarso, B. Saleh Abbas, and C. Ho Kang, "Analysis of the influence of ICT and public recognition on university credibility," *Intelligent Information and Database Systems*, pp. 541–551, 2019.

X. Wei, N. Saab, and W. Admiraal, "Assessment of cognitive, behavioral, and Affective Learning Outcomes in massive open online courses: A systematic literature review," *Computers & Education*, vol. 163, p. 104097, 2021.

B. Prenkaj, P. Velardi, G. Stilo, D. Distante, and S. Faralli, "A survey of machine learning approaches for student dropout prediction in online courses," *ACM Computing Surveys*, vol. 53, no. 3, pp. 1–34, 2020.

N. E. Simbolon, "EFL students' perceptions of blended learning in English language course: Learning experience and engagement," *Journal on English as a Foreign Language*, vol. 11, no. 1, pp. 152–174, 2021.

S. H. Tan, G. Thibault, A. C. Chew, and P. Rajalingam, "Enabling open-ended questions in team-based learning using automated marking: Impact on student achievement, learning and engagement," *Journal of Computer Assisted Learning*, vol. 38, no. 5, pp. 1347–1359, 2022.

R. Zhou, X. Wang, L. Zhang, and H. Guo, "Who tends to answer open-ended questions in an E-service survey? the contribution of closed-ended answers," *Behaviour & Information Technology*, vol. 36, no. 12, pp. 1274–1284, 2017.

R. Deng, P. Benckendorff, and D. Gannaway, "Understanding learning and teaching in moocs from the perspectives of students and instructors: A review of literature from 2014 to 2016," *Digital Education: Out to the World and Back to the Campus*, pp. 176–181, 2017.

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. "Albert: A lite bert for self-supervised learning of language representations," arXiv preprint arXiv:1909.11942, 2019.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T., "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv preprint arXiv:1910.01108, 2019.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., and Stoyanov, V., "Roberta: A robustly optimized bert pretraining approach," arXiv preprint arXiv:1907.11692, 2019.

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., and Purwarianti, A, "IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding," *arXiv preprint arXiv:2009.05387*, 2020.