different, as shown in TABLE II. Even though the model quality needs improvements, the model is able to generate some questions-worthy sentences. The results of the human evaluation experiment indicate that the majority (>50%) of respondents rated the fluency, relevance, and answerability of the generated questions as being between 'okay' and 'good' as shown in Fig. 2.

As for future improvement, we suggest the improvement or exploration by utilizing coreference resolution. By applying coreference resolution, the generated question quality will be improved, therefore the question will be more challenging to the students. We will also refine our model based on our experiment results and evaluation to create a balanced number of generated questions between methods available in the model, and increase the quality of fluency, relevance, and answerability.

REFERENCES

[1] I. Ning, N. Agustin, and A. Supriyanto, "Seminar Nasional-Jurusan Administrasi Pendidikan Fakultas Ilmu Pendidikan Universitas Negeri Malang Arah Manajemen Pada Masa Dan Pasca Pandemi Covid-19 PERMASALAHAN PENDIDIKAN DI INDONESIA."

[2] S. 1 Sman, J. Situbondo, and I. Timur, "Analisis Hasil Tes Potensi Skolastik sebagai Indikator Kesiapan Siswa Menghadapi Tes UTBK 2022," *Jurnal Penelitian Pendidikan dan Pembelajaran*, vol. 9, no. 2, 2022, doi: 10.21093/twt.v9i1.4210.

[3] H. L. Koediger and E. J. Marsh, "The positive and negative consequences of multiple-choice testing," *Journal of Experimental Psychology: Learning Memory and Cognition*, vol. 31, no. 5. pp. 1155–1159, Sep. 2005. doi: 10.1037/0278-7393.31.5.1155.

[4] B. Zou, P. Li, L. Pan, and A. T. Aw, "Automatic True/False Question Generation for Educational Purpose," 2022. [Online]. Available: https://wordnet.princeton.edu

[5] C. E. Brassil and B. A. Couch, "Multiple-true-false questions reveal more thoroughly the complexity of student thinking than multiple-choice questions: a Bayesian item response model comparison," *Int J STEM Educ*, vol. 6, no. 1, Dec. 2019, doi: 10.1186/s40594-019-0169-0.

[6] K. S. Khan, D. A. Davies, and J. K. Gupta, "Formative self-assessment using multiple true - False questions on the Internet: Feedback according to confidence about correct knowledge," *Med Teach*, vol. 23, no. 2, pp. 158–163, 2001, doi: 10.1080/01421590031075.

[7] C. Clark, K. Lee, M.-W. Chang, T. Kwiatkowski, M. Collins, and K. Toutanova, "BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions," May 2019, [Online]. Available: http://arxiv.org/abs/1905.10044

[8] T. C. Toppino and H. Ann Brochin, "Learning from Tests: The Case of True-False Examinations," *Journal of Educational Research*, vol. 83, no. 2, pp. 119–124, 1989, doi: 10.1080/00220671.1989.10885940.

[9] H. H. Remmers and E. M. Remmers, "THE NEGATIVE SUGGESTION EFFECT OF TRUE-FALSE EXAMINATION QUESTIONS 1."

[10] J. Qiu and D. Xiong, "Generating Highly Relevant Questions," Oct. 2019, [Online]. Available: http://arxiv.org/abs/1910.03401

[11] N. Duan, D. Tang, P. Chen, and M. Zhou, "Question Generation for Question Answering." [Online]. Available: https://answers.yahoo.com/

[12] X. ; Yao *et al.*, "Semantics-based Question Generation and Implementation," 2012. [Online]. Available: http://www.rug.nl/research/portal.

[13] T. M. Alsubait, "ONTOLOGY-BASED MULTIPLE-CHOICE QUESTION GENERATION."

[14] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." [Online]. Available: https://github.com/tensorflow/tensor2tensor

[15] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," Association for Computational Linguistics, 2015. [Online]. Available: http://aclweb.org/aclwiki/index.php?

[16] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: a large-scale dataset for Fact Extraction and VERification." [Online]. Available: https://github.com/awslabs/fever

[17] V. Nastase, J. Sayyad-Shirabad, M. Sokolova, and S. Szpakowicz, "Learning Noun-Modifier Semantic Relations with Corpus-based and WordNet-based Features." [Online]. Available: www.aaai.org

[18] V. Fathira, "Analysis of EFL Students' Ability in Reading Vocabulary of Synonyms and Antonyms," 2017.

[19] A. M. Fowler and B. S. Dissertation, "NEGATION IN NATURAL LANGUAGE PROCESSING," 2006.

[20] A. Thawani, J. Pujara, P. A. Szekely, and F. Ilievski, "Representing Numbers in NLP: a Survey and a Vision," Mar. 2021, [Online]. Available: http://arxiv.org/abs/2103.13136