PERBANDINGAN PENDEKATAN DIRECT SCORING DAN SIMILARITY-BASED SCORING DALAM SISTEM PENILAIAN JAWABAN SINGKAT OTOMATIS

SKRIPSI

Diajukan Untuk Memenuhi Sebagian dari Syarat Memperoleh Gelar Sarjana Komputer Program Studi Ilmu Komputer

Oleh

Bayu Wicaksono

2106836

PROGRAM STUDI ILMU KOMPUTER

FAKULTAS PENDIDIKAN MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS PENDIDIKAN INDONESIA

2025

PERBANDINGAN PENDEKATAN DIRECT SCORING DAN SIMILARITY-BASED SCORING DALAM SISTEM PENILAIAN JAWABAN SINGKAT OTOMATIS

Oleh
Bayu Wicaksono
2106836

Sebuah Skripsi yang Diajukan untuk Memenuhi Salah Satu Syarat Memperoleh Gelar Sarjana Komputer di Fakultas Pendidikan Matematika dan Ilmu Pengetahuan Alam

BAYU WICAKSONO

PERBANDINGAN PENDEKATAN DIRECT SCORING DAN SIMILARITY-BASED SCORING DALAM SISTEM PENILAIAN JAWABAN SINGKAT OTOMATIS

Disetujui dan disahkan oleh pembimbing:

Pembimbing I

Dr. Rasim, S.T., M.T.

NIP 197407252006041002

Pembimbing II

Yaya Wihardi, S.Kom., M.Kom.

NIP 198903252015041001

Mengetahui,

Ketua Program Studi Ilmu Komputer

Dr. Muhamad Nursalman, M.T.

NIP 197909292006041002

PERNYATAAN BEBAS PLAGIARISME

Saya yang bertanda tangan di bawah ini:

Nama	: Bayu Wicaksono
NIM	: 2106836
Program Studi	: Ilmu Komputer
Judul Karya	: Perbandingan Pendekatan *Direct Scoring* dan *Similarity-Based Scoring* dalam Sistem Penilaian Jawaban Singkat Otomatis

Dengan ini menyatakan bahwa karya tulis ini merupakan hasil kerja saya sendiri. Saya menjamin bahwa seluruh isi karya ini, baik sebagian maupun keseluruhan, bukan merupakan plagiarisme dari karya orang lain, kecuali pada bagian yang telah dinyatakan dan disebutkan sumbernya dengan jelas. Jika di kemudian hari ditemukan pelanggaran terhadap etika akademik atau unsur plagiarisme, saya bersedia menerima sanksi sesuai peraturan yang berlaku di Universitas Pendidikan Indonesia.

Bandung, 30 Juni 2025

Yang Membuat Pernyataan

Bayu Wicaksono

2106836

KATA PENGANTAR

Puji dan syukur penulis panjatkan kepada Allah SWT karena telah memberikan berkah dan rahmat-Nya. Shalawat dan salam semoga terlimpahcurahkan kepada baginda tercinta kita yaitu Nabi Muhammad SAW. Tanpa pertolongan-Nya, tentu penulis tidak akan sanggup untuk menyelesaikan skripsi yang berjudul "Perbandingan Pendekatan *Direct Scoring* dan *Similarity-Based Scoring* dalam Sistem Penilaian Jawaban Singkat Otomatis" ini tepat pada waktunya.

Penulisan skripsi ini memiliki tujuan sebagai salah satu syarat untuk memeroleh gelar Sarjana Ilmu Komputer (S.Kom) pada jenjang studi Strata-1 pada Program Studi Ilmu Komputer di Universitas Pendidikan Indonesia.

Dalam penyusunan skripsi ini, penulis telah berusaha semaksimal kemampuan yang dimiliki. Namun tidak bisa dipungkiri sebagai manusia, penulis pun tidak luput dari kesalahan. Kesalahan tersebut bisa berupa isi, tata bahasa, tanda baca, dan lain-lain. Oleh karena itu, penulis mengharapkan kritik dan saran dari pembaca, agar skripsi ini dapat menjadi lebih baik lagi. Dengan demikian, semoga skripsi ini bisa bermanfaat bagi penulis dan pembaca. Terima kasih

Bandung, 30 Juni 2025

Bayu Wicaksono

UCAPAN TERIMA KASIH

Penyusunan skripsi ini tidak lepas dari bantuan, bimbingan, dan dukungan dari berbagai pihak. Oleh karena itu, pada kesempatan ini saya ingin mengucapkan terima kasih yang sebesar-besarnya kepada:

1. Kedua orang tua serta kedua kakak penulis yang selalu memberikan doa, dukungan, bantuan, serta semangat kepada penulis dalam menjalankan perkuliahan dan penulisan skripsi ini.
2. Bapak Dr. Rasim, S.T., M.T. selaku dosen pembimbing I sekaligus dosen pembimbing akademik yang senantiasa membimbing dan memberi masukan yang bermanfaat untuk penulis.
3. Bapak Yaya Wihardi, M. Kom. selaku dosen pembimbing II yang senantiasa membimbing dan memberi masukan serta saran yang bermanfaat selama penulisan skripsi ini.
4. Bapak Dr. Muhammad Nursalman, M.T. selaku ketua Program Studi Ilmu Komputer Universitas Pendidikan Indonesia.
5. Seluruh jajaran dosen serta staff Departemen Pendidikan Ilmu Komputer Universitas Pendidikan Indonesia yang tidak bisa penulis tuliskan satu-persatu yang telah senantiasa membantu, mengarahkan, serta membimbing penulis selama menempuh pendidikan Sarjana.
6. Teman-teman Ilmu Komputer angkatan 2021 yang telah berjuang bersama penulis dalam kegiatan perkuliahan maupun luar perkuliahan.
7. Serta seluruh pihak lain yang tidak dapat penulis sebutkan yang telah membantu serta memberikan dukungan dalam perkuliahan serta penulisan skripsi ini hingga selesai.

Akhir kata, penulis berharap agar skripsi ini dapat memberikan manfaat serta menambah ilmu bagi para pembaca. Sekadar tulisan tidak dapat menggambarkan rasa terima kasih penulis kepada seluruh pihak yang telah mendukung. Semoga Tuhan Yang Maha Esa senantiasa memberikan kebahagiaan dan kesehatan untuk sekarang dan seterusnya. Aamiin ya rabbal alamin.

PERBANDINGAN PENDEKATAN DIRECT SCORING DAN SIMILARITY-BASED SCORING DALAM SISTEM PENILAIAN JAWABAN SINGKAT OTOMATIS

Oleh
Bayu Wicaksono
2106836

ABSTRAK

Dalam era pendidikan digital, kebutuhan akan sistem penilaian otomatis untuk jawaban teks pendek semakin meningkat. *Automatic Short Answer Scoring* (ASAS) bertujuan untuk mengotomasi proses penilaian ini dengan pendekatan yang efisien dan konsisten. Dua pendekatan yang umum digunakan dalam ASAS adalah *direct scoring* dan *similarity-based scoring*. Meskipun kedua pendekatan ini sudah banyak digunakan, penelitian sebelumnya cenderung fokus terhadap metrik seperti RMSE dan *Pearson Correlation* dalam menilai performa model. Penelitian ini bertujuan untuk melakukan analisis yang lebih mendalam dengan membandingkan kedua pendekatan tersebut pada dua skenario evaluasi, yaitu *specific-prompt* dan *cross-prompt*, dengan menilai akurasi dan stabilitas model. Dataset yang digunakan adalah dataset Rahutomo. Hasil analisis menunjukkan bahwa *direct scoring* lebih unggul dibandingkan *similarity-based scoring*. Pada skenario *specific-prompt*, diperoleh RMSE sebesar 0.0817 dan korelasi *Pearson* 0.9504, sedangkan pada *cross-prompt*, diperoleh RMSE sebesar 0.0917 dan korelasi *Pearson* 0.9286. Penelitian ini memberikan wawasan yang lebih komprehensif tentang performa model dengan tidak hanya mengandalkan metrik evaluasi, tetapi juga dengan melihat distribusi residual dan outlier, yang memberikan gambaran lebih lengkap mengenai stabilitas model.

Kata kunci: *Automatic Short Answer Scoring; Cross-Prompt; Direct Scoring;* Outlier*; Similarity-Based Scoring; Specific-Prompt*

COMPARISON OF DIRECT SCORING AND SIMILARITY-BASED SCORING APPROACHES IN AUTOMATIC SHORT ANSWER SCORING

Arranged by
Bayu Wicaksono
2106836

ABSTRACT

In the era of digital education, the need for automated scoring systems for short text answers has been steadily increasing. Automatic Short Answer Scoring (ASAS) aims to automate this assessment process with efficient and consistent approaches. Two commonly used approaches in ASAS are direct scoring and similarity-based scoring. Although these two approaches have been widely used, previous research has mostly focused on metrics like RMSE and Pearson Correlation to assess model performance. This study aims to provide a more in-depth analysis by comparing both approaches in two evaluation scenarios, specific-prompt and cross-prompt, by evaluating the accuracy and stability of the models. The dataset used in this study is the Rahutomo dataset. The results of the analysis show that direct scoring outperforms similarity-based scoring. In the specific-prompt scenario, an RMSE of 0.0817 and a Pearson Correlation of 0.9504 were obtained, while in the cross-prompt scenario, the RMSE was 0.0917 and the Pearson Correlation was 0.9286. This study provides a more comprehensive insight into model performance by not only relying on evaluation metrics but also examining the distribution of residuals and outliers, which offers a more complete picture of model stability.

Keywords: *Automatic Short Answer Scoring; Cross-Prompt; Direct Scoring;* Outlier*; Similarity-Based Scoring; Specific-Prompt*

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR PUSTAKA

Bonthu, S., Rama Sree, S., & Krishna Prasad, M. H. M. (2023). Improving the performance of automatic short answer grading using transfer learning and augmentation. *Engineering Applications of Artificial Intelligence*, *123*(September 2022), 106292. https://doi.org/10.1016/j.engappai.2023.106292

Cai, S., Shu, Y., Chen, G., Ooi, B. C., Wang, W., & Zhang, M. (2019). *Effective and Efficient Dropout for Deep Convolutional Neural Networks*. 1–12. http://arxiv.org/abs/1904.03392

Chamidah, N., Yulianti, E., & Budi, I. (2023). Evaluating the Impact of Sentence Tokenization on Indonesian Automated Essay Scoring Using Pretrained Sentence Embeddings. *Revue d'Intelligence Artificielle*, *37*(5), 1101–1108. https://doi.org/10.18280/ria.370502

Choi, R. Y., Coyner, A. S., Kalpathy-Cramer, J., Chiang, M. F., & Peter Campbell, J. (2020). Introduction to machine learning, neural networks, and deep learning. *Translational Vision Science and Technology*, *9*(2), 1–12. https://doi.org/10.1167/tvst.9.2.14

Cohn, C. (2020). *BERT Efficacy on Scientific and Medical Datasets: A Systematic Literature Review*. https://via.library.depaul.edu/cdm_etd/24%0Ahttps://www.proquest.com/dissertations-theses/bert-efficacy-on-scientific-medical-datasets/docview/2476867924/se-2?accountid=10639

Denis Rothman. (2024). *Transformers for Natural Language Processing and Computer Vision*. Packt Publishing.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*(Mlm), 4171–4186.

Efendi, R., Junaidi, A., & Rizki, A. M. (2024). Penentuan Pusat Klaster Secara

Otomatis Pada Algoritma Density Peaks Clustering Berbasis Metode Inter Quartile Range. *Jurnal Informatika Dan Teknik Elektro Terapan*, *12*(3). https://doi.org/10.23960/jitet.v12i3.4997

Engel Novita Ramadani, & Dina Fitria Handayani. (2024). Instrumen Penilaian Hasil Pembelajaran Kognitif Pada Tes Objektif. *Jurnal Pendidikan Dan Ilmu Sosial (Jupendis)*, *2*(4), 86–96. https://doi.org/10.54066/jupendis.v2i4.2159

Ferreira Mello, R., Pereira Junior, C., Rodrigues, L., Pereira, F. D., Cabral, L., Costa, N., Ramalho, G., & Gasevic, D. (2025). Automatic Short Answer Grading in the LLM Era: Does GPT-4 with Prompt Engineering beat Traditional Models? *15th International Conference on Learning Analytics and Knowledge, LAK 2025*, 93–103. https://doi.org/10.1145/3706468.3706481

Galal, O., Abdel-Gawad, A. H., & Farouk, M. (2024). Rethinking of BERT sentence embedding for text classification. *Neural Computing and Applications*, *36*(32), 20245–20258. https://doi.org/10.1007/s00521-024-10212-3

Haidir, M. H., & Purwarianti, A. (2020). Short Answer Grading Using Contextual Word Embedding and Linear Regression. *Jurnal Linguistik Komputasional*, *3*(2), 54–61. https://inacl.id/journal/index.php/jlk/article/view/38

Hao, Y., Dong, L., Wei, F., & Xu, K. (2019). Visualizing and understanding the effectiveness of BERT. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 4143–4152. https://doi.org/10.18653/v1/d19-1424

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-Decem*, 770–778. https://doi.org/10.1109/CVPR.2016.90

Karanikolas, N., Manga, E., Samaridi, N., Tousidou, E., & Vassilakopoulos, M. (2023). Large Language Models versus Natural Language Understanding and Generation. *ACM International Conference Proceeding Series*, *4*, 278–290. https://doi.org/10.1145/3635059.3635104

Kaya, M., & Cicekli, I. (2024). A Hybrid Approach for Automated Short Answer

Grading. *IEEE Access*, *12*(May), 96332–96341. https://doi.org/10.1109/ACCESS.2024.3420890

Khatoon, S., Alshamari, M. A., Asif, A., Hasan, M. M., Abdou, S., Elsayed, K. M., & Rashwan, M. (2021). Development of social media analytics system for emergency event detection and crisismanagement. *Computers, Materials and Continua*, *68*(3), 3079–3100. https://doi.org/10.32604/cmc.2021.017371

Kim, S., & Jo, M. (2024). Is GPT-4 Alone Sufficient for Automated Essay Scoring?: A Comparative Judgment Approach Based on Rater Cognition. *L@S 2024 - Proceedings of the 11th ACM Conference on Learning @ Scale*, 315–319. https://doi.org/10.1145/3657604.3664703

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP. *COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference*, 757–770. https://doi.org/10.18653/v1/2020.coling-main.66

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Liu, Z., Xu, Z., Jin, J., Shen, Z., & Darrell, T. (2023). Dropout Reduces Underfitting. *Proceedings of Machine Learning Research*, *202*, 21715–21729.

Locke, S., Bashall, A., Al-Adely, S., Moore, J., Wilson, A., & Kitchen, G. B. (2021). Natural language processing in medicine: A review. In *Trends in Anaesthesia and Critical Care* (Vol. 38, pp. 4–9). Elsevier Ltd. https://doi.org/10.1016/j.tacc.2021.02.007

Mahmood, S. A., & Abdulsamad, M. A. (2024). Automatic assessment of short answer questions: Review. *Edelweiss Applied Science and Technology*, *8*(6), 9158–9176. https://doi.org/10.55214/25768484.v8i6.3956

Mardini G, I. D., Quintero M, C. G., Viloria N, C. A., Percybrooks B, W. S., Robles N, H. S., & Villalba R, K. (2024). A deep-learning-based grading system (ASAG) for reading comprehension assessment by using aphorisms as open-answer-questions. *Education and Information Technologies*, *29*(4), 4565–4590. https://doi.org/10.1007/s10639-023-11890-7

Maria, A., Vasquez, B., Utz, H. F., & Peter, H. (2016). Outlier detection methods

for generalized lattices : a case study on the transition from ANOVA to REML. *Theoretical and Applied Genetics*. https://doi.org/10.1007/s00122-016-2666-6

Mohler, M., & Mihalcea, R. (2009). Text-to-text semantic similarity for automatic short answer grading. *EACL 2009 - 12th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings*, *April*, 567–575. https://doi.org/10.3115/1609067.1609130

Nicolson, A., & Paliwal, K. K. (2020). Masked multi-head self-attention for causal speech enhancement. *Speech Communication*, *125*, 80–96. https://doi.org/10.1016/j.specom.2020.10.004

Pires, T., Schlinger, E., & Garrette, D. (2020). How multilingual is multilingual BERT? *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 4996–5001. https://doi.org/10.18653/v1/p19-1493

Qiu, X. P., Sun, T. X., Xu, Y. G., Shao, Y. F., Dai, N., & Huang, X. J. (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, *63*(10), 1872–1897. https://doi.org/10.1007/s11431-020-1647-3

Rahutomo, F., Ari Roshinta, T., Rohadi, E., Siradjuddin, I., Ariyanto, R., Setiawan, A., & Adhisuwignjo, S. (2018). Open Problems in Indonesian Automatic Essay Scoring System. *International Journal of Engineering & Technology*, *7*(4.44), 156. https://doi.org/10.14419/ijet.v7i4.44.26974

Raj, J. A., Qian, L., & Ibrahim, Z. (2024). *Fine-tuning -- a Transfer Learning approach*. http://arxiv.org/abs/2411.03941

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 3982–3992. https://doi.org/10.18653/v1/d19-1410

Salam, M. A., El-Fatah, M. A., & Hassan, N. F. (2022). Automatic grading for Arabic short answer questions using optimized deep learning model. In *PLoS ONE* (Vol. 17, Issue 8 August). https://doi.org/10.1371/journal.pone.0272269

Salim, H. R., De, C., Pratamaputra, N. D., & Suhartono, D. (2022). Indonesian automatic short answer grading system. *Bulletin of Electrical Engineering and Informatics*, *11*(3), 1586–1603. https://doi.org/10.11591/eei.v11i3.3531

Santoso, R. R., Megasari, R., & Hambali, Y. A. (2020). Implementasi Metode Machinelearning. *Jurnal Aplikasi Dan Teori Ilmu Komputer*, *3*(2), 85–97. https://ejournal.upi.edu/index.php/JATIKOM

Sarker, I. H. (2021). Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. In *SN Computer Science* (Vol. 2, Issue 6, pp. 1–20). Springer Singapore. https://doi.org/10.1007/s42979-021-00815-1

Song, Y., Wang, J., Liang, Z., Liu, Z., & Jiang, T. (2020). *Utilizing BERT Intermediate Layers for Aspect Based Sentiment Analysis and Natural Language Inference*. http://arxiv.org/abs/2002.04815

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*, 1929–1958.

Tsukagoshi, H., Sasano, R., & Takeda, K. (2021). DefSent: Sentence Embeddings using Definition Sentences. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, *2*, 411–418. https://doi.org/10.18653/v1/2021.acl-short.52

Tunstall, L., Von Werra, L., & Wolf, T. (2022). Natural Language Processing with Transformers (Revised Edition). In *O'Reilly Media* (Vol. 19, Issue 1). https://www.oreilly.com/library/view/natural-language-processing/9781098136789/

van der Goot, R., Müller-Eberstein, M., & Plank, B. (2022). Frustratingly Easy Performance Improvements for Low-resource Setups: A Tale on BERT and Segment Embeddings. *2022 Language Resources and Evaluation Conference, LREC 2022*, *June*, 1418–1427.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomes, A. N., Kaiser, L., & Polusukhin, I. (2017). Attention Is All You Need. *31st Conference on Neural Information Processing Systems*, *Nips*.

https://doi.org/10.1145/3583780.3615497

Verma, V. (2025). *A Comprehensive Framework for Residual Analysis in Regression and Machine Learning*. *January*. https://doi.org/10.52783/jisem.v10i31s.4958

Wijanto, M. C., & Yong, H. S. (2024). Combining Balancing Dataset and SentenceTransformers to Improve Short Answer Grading Performance. *Applied Sciences (Switzerland)*, *14*(11). https://doi.org/10.3390/app14114532