A Hybrid Approach for Assessing Research Text Clarity by Combining Semantic and Quantitative Measures

Pranit Prasant Pai[a], Kaashika Agrawal[b], Anushri Anil[c], Archit Saigal[d] and Arti Arya[e]
PES University, Bengaluru, India

Keywords: Large Language Models, Semantic Clarity, Quantitative Clarity, BERT, DistilBERT.

Abstract: The concept of clarity of text is one that can be quite subjective in nature. This work aims to evaluate the clarity of published research in terms of two key components - Semantic Clarity and Quantitative Clarity. Semantic clarity aims to assess how effectively the meaning of the text is structured, articulated, and conveyed to the reader, and quantitative clarity employs a combination of previously defined formulations and metrics to provide measurable insights into the clarity of the text. Semantic Clarity, predicted using a BERT Model achieved a final validation Mean Squared Error of 0.0169, while the Quantitative Clarity, predicted using a DistilBERT Model, achieved a training loss of 6.9776 and validation loss of 3.6322. By integrating these two dimensions, this study seeks to enhance the overall evaluation process and contribute to a more nuanced understanding of research quality.

1 INTRODUCTION

Clarity is essential for effective communication across various fields. While often viewed as subjective, clarity can be quantified using numerical metrics and comparative methods. This study aims to objectively assess clarity in written content, particularly in academic publishing, where clarity and precision significantly impact manuscript acceptance and influence. In this study, we propose a dual approach that incorporates numerical metrics and comparative scoring methods both. Numerical metrics are derived from computational linguistic tools that analyse textual features such as sentence structure, vocabulary usage, and readability indices like the Fog Index(Yaffe, 2022). These tools help assess factors like sentence length, cognitive load(Mikk, 2008), lexical density(To et al., 2013), all of which contribute to textual clarity. Comparative scoring, on the other hand, involves benchmarking a text against high-quality publications from prestigious A-star conferences, providing a reference standard for clarity evaluation. This method also incorporates decision models, such as those suggested by (McElfresh et al., 2021), to address ambiguity and indecisiveness in clarity scoring. The main contributions of this work are listed below:

- Introduced two new aspects of text clarity: S-clarity (semantic clarity using BERT) and Q-clarity (quantitative clarity using established text clarity scores), and developed a combined clarity score (CS_C) that integrates these assessments equally.
- Proposed the concept of an indecisive score to handle ambiguity in clarity assessments.
- Proposed an approach involving A-star conference benchmarking for clarity, using high-quality academic papers as a reference standard for objective clarity scoring.

The motivation behind this study stems from the need to assess the concept of ”clarity” in research communication, prompting the need to bridge the gap between subjective impressions of clarity and objective measurements. By quantifying and comparing the clarity of research papers, we aim to provide a systematic method that enhances clarity assessment, improving the quality of communication not just in academic publishing but across various fields.

[a] https://orcid.org/0009-0000-2521-4345
[b] https://orcid.org/0009-0007-7588-2203
[c] https://orcid.org/0009-0006-6842-4848
[d] https://orcid.org/0009-0008-1689-130X
[e] https://orcid.org/0000-0002-4470-0311

Pai, P. P., Agrawal, K., Anil, A., Saigal, A. and Arya, A.
A Hybrid Approach for Assessing Research Text Clarity by Combining Semantic and Quantitative Measures.
DOI: 10.5220/0013118600003890
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 2*, pages 144-152
ISBN: 978-989-758-737-5; ISSN: 2184-433X
Proceedings Copyright © 2025 by SCITEPRESS – Science and Technology Publications, Lda.

2 RELATED WORK

(Liu et al., 2024) investigate how language models (LMs) handle long contexts and the challenges they face in maintaining coherence over extended text spans. The authors analyse performance on two tasks: multi-document question answering and key-value retrieval, focusing on how the models utilise information from the beginning and the end of the context. It was found that while LMs are effective on short-context texts, their performance seems to degrade as the position of information is varied within a context. Even explicitly long-context models find it difficult to access relevant information in the middle of a context as compared to when the required information is located either at the beginning or the end of the context. The study proposed by us also uses a language model - BERT in order to extract and score the semantics of text in research papers.

(Yu et al., 2024) address the challenge of semantic similarity matching in patent documents which is required for patent classification and related tasks. A combination of an ensemble of BERT-related models and a new text processing method is proposed. It uses four variations of BERT to capture different semantic aspects of patent text. The text processing method involves advanced techniques of preprocessing and segmenting patent documents to improve the quality of input given to the models. Improved semantic clarity matching was observed as compared to existing methods. The ensemble BERT-related models outperformed single-model approaches and showed that the combined approach achieved higher precision and recall in semantic similarity tasks. The proposed study also aims to leverage a fine-tuned BERT model for a similar task.

(Vulić et al., 2020) investigate how well pretrained language models (LMs) capture lexical semantics by employing probing tasks. Probing classifiers were used to assess the extent to which these models encode information about word meanings and their semantic relationships. Probing tasks were designed with a focus on various aspects of lexical semantics. Different pretrained models were studied to understand their lexical semantic knowledge. The authors found that pretrained LMs contain significant information about lexical semantics, but their effectiveness depended on the model and probing task. The study finds that models like BERT capture detailed semantic information better than models like GPT-2, especially for tasks requiring a deeper level of understanding. Similar to this study, BERT model is also used in the proposed work for lexical semantic analysis.

(Yaffe, 2022) critically examines the Fog Index, a readability metric used to assess the complexity of written text. An empirical analysis comparing the Fog Index with other readability metrics was done. Its performance is also evaluated across different genres and audiences. The results suggest that while the Fog Index provides a quick estimation of text readability, its utility may be limited. It was found inconsistent in its correlation with human evaluations, showing that it may not accurately reflect actual comprehensibility in all cases. The author concludes that the Fog Index overlooks some factors but also that it can serve as a preliminary tool for assessing readability along with additional metrics and qualitative assessments for a good, all round evaluation. In the proposed study, the Fog Index is used along with 3 other quantitative metrics to assess the quantitative clarity of a paper which is then combined with the semantic clarity to provide a comprehensive evaluation metric.

(To et al., 2013) explore the relationship between lexical density and readability in English textbooks. Lexical density was evaluated by calculating the ratio of content words to the total number of words in the texts. Additional readability formulae were used to assess the complexity of textbooks. The findings showed a significant correlation between higher lexical density and increased readability levels. Textbooks having greater lexical density were associated with complex structures, which could pose comprehension challenges for students. Enhancing content by using denser vocabulary may hinder certain learners from accessing the content. Thus, it was suggested that balancing lexical density in educational publications is crucial to accommodate student engagement. Comparable to the study, in the proposed work, lexical density is one of the parameters used to calculate quantitative clarity.

(Matthews and Folivi, 2023) examine the influence of sentence length on perception, especially the effect of complexity and word count on readability. In several experiments, the authors presented participants with sentences of differing length and structures, evaluating and analysing their subjective judgements and cognitive processing times. The method combined different metrics to assess participants' perceptions of clarity and ease of comprehension according to the sentences' characteristics. It was found that shorter sentences were generally perceived as clearer and easier to understand, while longer sentences often led to increased cognitive load and varied interpretations. The results propose that sentence length significantly affects readability, with implications for writing and communication practices. The authors conclude that optimising sentence length can complement comprehension and engagement, supporting the

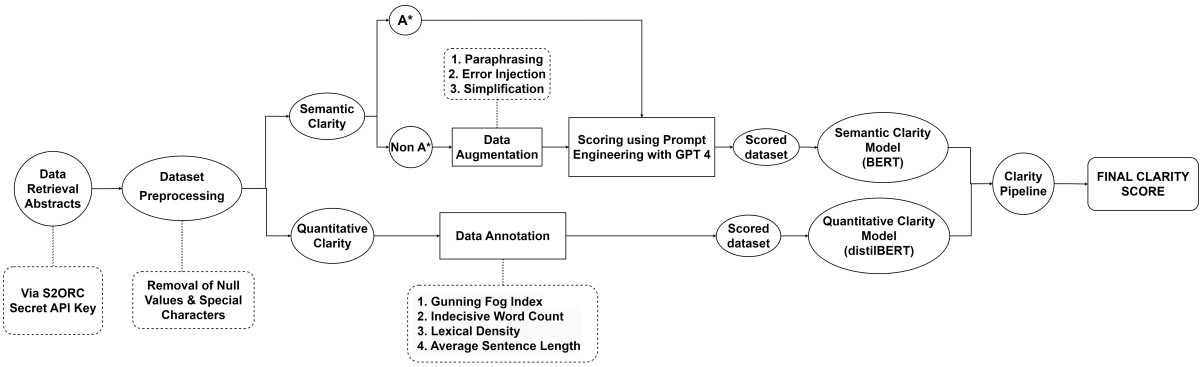

Figure 1: Proposed Approach for assessing the clarity of a research work.

notion that conciseness is essential for effective communication. In this proposed study, to evaluate quantitative clarity mathematically, indecisiveness word count and average sentence length along with other parameters are used.

(Chiang and Lee, 2023) explore the feasibility of using large language models (LLMs) as substitutes for human evaluators in various evaluation tasks. They focus on a) whether LLMs can replace human judgement in accurately assessing the outputs of other models and systems and b) if their evaluations are reliable and rational as compared to human evaluations. The authors find that while LLMs emulate human evaluations for specific tasks involving straightforward metrics or well-defined criteria, they struggle with more nuanced evaluations that require deep understanding. They conclude that LLMs can be used to assist human evaluations in order to provide additional insights and scale evaluation processes. However, they are not completely reliable due to their biases and current limitations.

After an in-depth literature survey, the following gaps are identified in the existing studies:

- While the above mentioned studies provide insights into how language models handle long contexts, they do not explore specific techniques for enhancing context optimization in similar tasks.
- They also evaluate the potential of large language models as alternatives to human evaluators but fail to address the variability in their performance across different domains and the biases that may affect the consistency of evaluations.
- The studies focusing on the quantitative aspects elaborate on how the metrics are used in textbooks and other written texts but have not tried using them to evaluate research quality.

The proposed work aims to address the above issues and uses large language models only for assessing the concept of clarity, which can in turn be used with other criteria or methods in a pipeline in order to come up with a holistic evaluation.

3 PROPOSED METHODOLOGY

In this study, a procedure-oriented approach as seen in Fig 1, was employed to develop a model that is able to predict the clarity of abstracts. The concept of text clarity can be seen as a combination of different aspects, here explored as semantic and quantitative clarity, where the former refers to actual understandability of the text and motives, and the latter refers to clarity as defined by existing established quantitative scores.

3.1 Semantic Clarity

In this study, text is said to be semantically clear if, upon reading through the text, the intentions of the writer are conveyed in an efficient and understandable manner. As this quality of text is difficult to simply define using manual scoring, the concept of A* Conferences was employed and a pre-trained BERT model was fine tuned in order to predict scores for unseen text.

3.1.1 Data Retrieval and Annotation

The process began with retrieval of substantial text data. This data consisted of 1,00,000 abstracts from published research present in the S2ORC dataset(Lo et al., 2019). This dataset is a comprehensive resource that contains both metadata as well as full-text abstracts from published academic papers. Specifically, papers were retrieved using keyword based filtering using the terms “Computer Science” and “Technology”, After cleaning the data and filtering out any null values, the resultant dataset consisted of 80,000 entries.

Next, another filtering process was applied to extract the abstracts of papers published in A* conference venues. These papers are those that are recognized for their high standards and rigorous peer review. The list of A* conferences used to filter the abstracts was identified using the list maintained by CORE (The Computing Research and Education Association of Australasia)(Simon et al., 2023), an organisation that ranks academic conferences based on their quality and impact. Focusing on these high-quality abstracts ensured that the training data represented well-written and assuredly clear content, providing a strong foundation for the model.

A novel approach used in order to label the abstracts with scores involved using prompt engineering with ChatGPT(Marvin et al., 2023). A score, lying between 0 and 1, was assigned to each abstract. This assignment was based on criteria like citation metrics, acceptance rates, and expert opinion. As it would be incorrect to say papers published in non A* venues are semantically unclear, certain purposeful data augmentation was performed on such abstracts. This included randomised paraphrasing, simplification, and deliberate error injection followed by assignment of lower clarity scores. After the labelling of every abstract was completed, the final dataset ensured diversity reflecting both high-quality and lower-quality abstracts.

3.1.2 Model Fine Tuning

For the first step in the model training, A BERT tokenizer ('bert-base-uncased')(Stankevičius and Lukoševičius, 2024) was employed to preprocess the abstracts in the dataset. Initially, the tokenizer converted every abstract into a distinct sequence of tokens. These were transformed into PyTorch tensors. This step was taken in order to ensure that the text was structured in a way that was appropriate for the BERT model. After this, a custom PyTorch Dataset class was developed mainly to pair the tokenized abstracts with their corresponding semantic clarity score. This custom dataset was given to the PyTorch 'DataLoader' in order to handle the processes of batching and shuffling. This optimised the model's exposure to varied sequences seen in the training loop. This step was essential if the model was to be generalised and overfitting was to be prevented as much as possible.

In order to make the pre-trained BERT model ('bert-base-uncased')(Stankevičius and Lukoševičius, 2024) work for regression tasks, it was adjusted with a single regression output head that could output continuous values. For the training process, the AdamW optimizer was used with a learning rate of 1e-5. In the training loop, the tokenized abstracts were given to the BERT model which in turn generated clarity score predictions. To update the model weights periodically, backpropagation was done.

The model was able to achieve a mean squared error of 0.0169 on the validation data and the overall training and testing loss during the epochs can be visualised in Fig 2. In order to test the model's performance further, unseen abstracts from various venues were scored using a custom prediction function. This function first tokenizes the input text, sends it through the trained BERT model, and finally outputs a predicted clarity score.

3.2 Quantitative Clarity

3.2.1 Data Annotations and Formulations

The clarity of textual content in research papers can be assessed using quantitative metrics and scores. In order to demonstrate the same, 80,000 research paper abstracts were subjected to four formulations: Gunning Fog Index, Indecisiveness Word Count, Lexical Density and Average Sentence Length. These formulations are essential in evaluating several key aspects of textual clarity thereby providing a complete understanding of the clarity and readability of the text.

1. Gunning Fog Index. The Gunning Fog Index (Gu and Dodoo, 2020), visualised in equation (1), is a popular readability metric that approximates the number of years of formal education required by a reader to understand the text when read for the first time. The index offers an intuitive way to gauge the complexity of articles. This makes it a popular choice for analysing the readability of text across various fields like education, journalism and legal writing. The index is calculated by taking two key factors into consideration: the average sentence length and the percentage of complex words. Complex words are assumed to be those words that contain three or more syllables

$$\text{Fog Index} = 0.4 \times \left(\frac{W}{S} + 100 \times \frac{C}{W} \right) \quad (1)$$

Where:

- W = Number of words
- S = Number of sentences
- C = Number of complex words (words with 3 or more syllables)

The formulation highlights two important drivers of clarity: sentences that are longer demand more cognitive effort on the reader's part and complex words

require a higher level of linguistic understanding. A higher Gunning Fog Index means that the text is difficult to read and understand. In this study, the index is capped at 20, indicating the upper bound of difficulty.

2. Indecisiveness Word Count. The Indecisiveness Word Count Score measures the presence of words or phrases that indicate the chances of uncertainty or lack of clarity. This includes terms like "maybe", "it could be", "possibly", to name a few. The score was further enhanced by applying position-based and contextual penalties.

- *Position-Based Penalty:* Higher penalties are applied if an indecisive term appears at the beginning or end of the sentence where the impact of its presence is most pronounced. A medium penalty is enforced when the term appears near the start or end of the text, within 20% of the total words.
- *Contextual Penalty:* The presence of assertive words like "certainly", "definitely" and "undoubtedly" in the vicinity of the indecisive term reduces the penalty. If the term is not surrounded by such assertive words, a higher penalty is applied since this reinforces the presence of uncertainty.

In addition to this, the frequency of indecisive words has also been considered. A more severe penalty is imposed if such indecisive terms appear more often within the text. The final score is inversely proportional to the density of these words. This approach provides an indication of how uncertainty reflected in the writing style impacts its clarity.

3. Lexical Density. Lexical density(Amer, 2021) is a measure of the informational content present in the text. It is calculated as a ratio of content words which include nouns, verbs, adjectives and adverbs to the total number of words available. A higher value of lexical density indicates a text that is rich in content, which can either increase clarity by being informative or reduce clarity if the content is highly dense and complex. This formulation as seen in equation (2) helps in understanding the balance between content and readability in the abstracts.

$$\text{Lexical Density} = \frac{C_W}{W} \tag{2}$$

Where:

- C_W = Number of content words (nouns, verbs, adjectives, and adverbs)
- W = Total number of words

4. Average Sentence Length:
Average sentence length is a metric that calculates the average number of words per sentence in the text. The formulation used for the same is seen in equation (3). Shorter sentences are generally easier to understand, while longer sentences can increase the cognitive load on the reader, potentially reducing clarity.

$$\text{Average Sentence Length} = \frac{\sum L_s}{n_s} \tag{3}$$

Where:

- $\sum L_s$ = Total number of words in all sentences
- n_s = Number of sentences

3.2.2 Data Preparation

Data preparation, a critical step in this study, involved the processing of abstracts from several research papers followed by the annotations based on the four distinct textual clarity formulations. Firstly, the text was preprocessed by tokenizing the sentences and words followed by removal of stop words. Subsequently, the four formulations were applied on the abstracts. The final output was a JSON file where each object contained the abstract along with the four scores. This comprehensive data preparation ensured that the abstracts were evaluated in a systematic manner for the evaluation of textual clarity. The data was split into training(90%) and testing(10%) sets.

3.2.3 Model Training

The dataset was first loaded and split into training and testing sets. The text data was tokenized using the DistilBERT tokenizer (Shah et al., 2024), after which a custom dataset was created to handle both the tokenized inputs and the four scores. Mixed Precision Training (AMP) was employed to optimise the memory usage and speed up the computation at the time of fine-tuning DistilBERT. The model was trained using the AdamW optimizer which handled the weight decay seamlessly. A custom linear learning rate scheduler was also used to adjust the learning rates dynamically during the training process. To optimise the performance even further, gradient accumulation was utilised. This allowed processing of larger batches by aggregating gradients over multiple steps before updating the weights of the model. These techniques together helped the model handle the large dataset efficiently and helped ensure a faster convergence. To assess the results, several tests were conducted on the trained model using contrasting pieces of text: one clear and the other unclear.

3.2.4 Final Quantitative Clarity Score

In order to obtain the final quantitative clarity score, the four scores were normalised and assigned weights

based on their importance in determining the textual clarity. The Gunning Fog Index was given a weightage of 0.35 as it is a well established metric that indicates clarity and readability. The Indecisive Word Count Score was assigned a slightly lower weightage of 0.25 as it is already a penalty based score. The relation between clarity and Lexical Density was seen to follow a Gaussian bell shaped distribution. Hence, Lexical Density was normalised using a Gaussian function(Thorpe, 2023) and given a weightage of 0.25 as it balances with the indecisive word count score in indicating how clear the text is. Owing to the potential variability in the Average Sentence Length, a lower weightage of 0.15 was assigned to it. A weighted score combining the four metrics gave the overall quantitative clarity score as seen in equation (4).

$$CS_Q = (0.35 \times G) + (0.25 \times I) + (0.25 \times L) + (0.15 \times A) \quad (4)$$

Where:

- CS_Q = Final Quantitative Clarity Score
- G = Normalised Gunning Fog Index Score
- I = Indecisive Word Count Score
- L = Normalised Lexical Density Score
- A = Normalised Average Sentence Length Score

3.3 Combined Clarity Score

Once both the individual models to predict a semantic and quantitative score for a given abstract were fine tuned and ready for use, a final pipeline was developed. The final semantic clarity model was able to predict clarity by generating a score from the input text. Simultaneously, the final quantitative clarity model evaluated various linguistic attributes such as lexical density, sentence length, the Gunning Fog Index, and the presence of indecisive language and provided a final quantitative clarity score to the text. Finally, an equally weighted combination of these 2 scores was used to calculate the final clarity score. Making use of a custom function to execute both models concurrently(Sodian et al., 2022), the final clarity score was computed by summation of the equally weighted semantic and quantitative clarity scores, as seen in equation (5). The decision to assign equal weights of 0.5 for the semantic and quantitative clarity scores, stems from the fact that both scores assess clarity from different perspectives that are equally essential to the overall assessment. Weighing the two aspects equally avoids bias toward one type of clarity. Using the coefficient of 0.5 effectively averages the scores, ensuring that the final combined clarity score lies on the same scale as the two input scores, that is between 0 and 1.

$$CS_C = (0.5 \times CS_S) + (0.5 \times CS_Q) \quad (5)$$

Where:

- CS_C = Combined Clarity Score
- CS_S = Semantic Clarity Score
- CS_Q = Quantitative Clarity Score

This approach ensured a comprehensive analysis accounting for the content's conveyed meaning as well as its linguistic structure, yielding a robust clarity evaluation.

4 RESULTS AND DISCUSSION

In the semantic clarity part of the pipeline, the BERT model was trained for three epochs and its training and testing loss can be visualised using Fig 2. The Mean Squared Error on the validation data was a promising 0.0169. While for the Quantitative Clarity part of the pipeline, the model was made to run for a total of 5 epochs. As seen in Fig 3, the model gave good results with the training loss being 6.9776 and the validation loss being 3.6322.

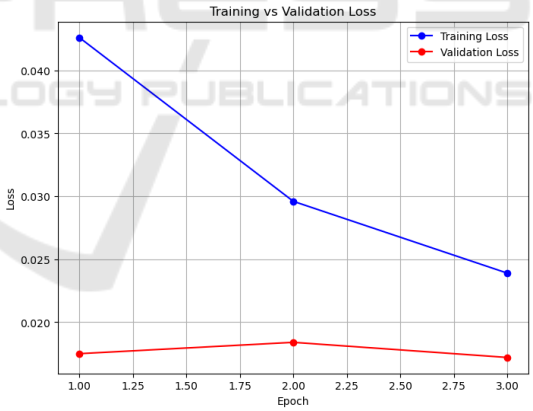

Figure 2: Training and testing loss variation per epoch for BERT Model to predict S-Clarity score.

Currently in literature, textual clarity has not been evaluated using the two aspects used in this work. There has been research undertaken to assess clarity in terms of scores(Assamarqandi et al., 2023) similar to the ones incorporated in the QClarity section of this work, or even using natural language processing techniques separately(Choi, 2024). However combining these two aspects is a novelty presented in the research presented by us.

A plethora of research paper abstracts were scored using the proposed Combined Clarity Model of which

Table 1: Clarity Scores for Various Papers.

Paper	Type	CS_S	CS_Q	CS_C
(Desai and Chin, 2023)	A*	0.920	0.434	0.677
(Liao et al., 2023)	A*	0.860	0.499	0.679
(Kulkarni et al., 2024)	Non-A*	0.865	0.480	0.672
(Goldstein et al., 2024)	A*	0.958	0.488	0.723
(Bagayatkar and Ivin, 2024)	Non-A*	0.748	0.316	0.532
(Shewale et al., 2024)	Non-A*	0.676	0.448	0.562
(Kingma, 2014)	A*	0.858	0.516	0.687
(LeCun et al., 2015)	Non-A*	0.772	0.383	0.577
(He et al., 2016)	A	0.845	0.583	0.714
(Reddy et al., 2022)	Non-A*	0.941	0.475	0.708
(Gargiulo et al., 2017)	Non-A*	0.662	0.526	0.594
(Egele et al., 2021)	A	0.947	0.439	0.693
(Yudistira et al., 2022)	Non-A*	0.767	0.467	0.617
(Cacace et al., 2023)	Non-A*	0.747	0.490	0.619
(Kromidha, 2023)	Non-A*	0.735	0.473	0.604
(Koivisto and Hamari, 2019)	Non-A*	0.711	0.488	0.600
(Jiménez-Luna et al., 2020)	Non-A*	0.696	0.317	0.507
(Huang et al., 2024)	A*	0.956	0.462	0.709
(Cabitza et al., 2023)	A*	0.969	0.494	0.731
(Zulfiqar et al., 2023)	A*	0.942	0.506	0.724

Figure 3: Training and testing loss variation per epoch for DistilBERT Model to predict Q-Clarity score.

20 of the results are tabulated in Table 1. The papers published in A* venues have consistently scored slightly higher indicating the better clarity of research published in such venues. However, publication venue is not the sole parameter on the basis of which clarity can be judged. Hence, there are cases where the constituent scores are higher even if the paper was published in a Non A* venue. The quantitative clarity score has been put together using predefined ranges and justified weighting, hence the score is slightly critical. The two scores considering different aspects of clarity work in tandem to provide a comprehensive combined clarity score.

5 CONCLUSION

The proposed clarity scoring pipeline offers a comprehensive method for evaluating textual clarity by making use of semantic clarity and quantitative clarity BERT-based models. This hybrid methodology successfully differentiates between clear and unclear texts, as demonstrated by the robust results. By assessing both the semantic understanding and structural coherence of text, the system provides a dependable solution for clarity analysis. This makes it a valuable tool for a variety of applications, including evaluating academic abstracts, enhancing the quality of automated writing systems, and conducting in-depth readability assessments.

Future work could focus on various enhancements to further improve the system's assessment quality. Fine-tuning the weighting between semantic and quantitative models could allow greater adaptability to specific types of content, like creative writing or technical documents. Advanced natural language processing (NLP) techniques, such as context-aware embeddings and domain-specific language models, could incorporated in order to refine the pipeline's sensitivity to more clarity features. User feedback loops could also be added to provide dynamic adjustments to the scoring mechanism, thus improving relevance over time. Finally, using more diverse text types and languages in the dataset could increase the

system's generalizability and effectiveness in multilingual and multicultural contexts. This could easily make the system more global and multidisciplinary, making way for broader applications.

REFERENCES

Amer, M. A. B. (2021). Lexical density and readability of secondary stage english textbooks in jordan. *International Journal for Management and Modern Education*, 2(2):11–20.

Assamarqandi, A., Dewi, R., and Daddi, H. (2023). Analyzing clarity and readability of text used in critical reading comprehension classroom at english education department of unismuh makassar. *Journal of Language Testing and Assessment*, 3(2):145–154.

Bagayatkar, A. and Ivin, B. (2024). Survey paper on machine learning and deep learning driven applications using bayesian techniques. In *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, pages 1–7. IEEE.

Cabitza, F., Campagner, A., and Basile, V. (2023). Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6860–6868.

Cacace, J., Caccavale, R., Finzi, A., and Grieco, R. (2023). Combining human guidance and structured task execution during physical human–robot collaboration. *Journal of Intelligent Manufacturing*, 34(7):3053–3067.

Chiang, C.-H. and Lee, H.-y. (2023). Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Choi, J. H. (2024). Measuring clarity in legal text. *"U. Chi. L. Rev."*, 91:1.

Desai, S. and Chin, J. (2023). Ok google, let's learn: Using voice user interfaces for informal self-regulated learning of health topics among younger and older adults. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–21.

Egele, R., Balaprakash, P., Guyon, I., Vishwanath, V., Xia, F., Stevens, R., and Liu, Z. (2021). Agebo-tabular: joint neural architecture and hyperparameter search with autotuned data-parallel training for tabular data. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–14.

Gargiulo, F., Silvestri, S., and Ciampi, M. (2017). A big data architecture for knowledge discovery in pubmed articles. In *2017 IEEE Symposium on Computers and Communications (ISCC)*, pages 82–87. IEEE.

Goldstein, H., Cutler, J. W., Dickstein, D., Pierce, B. C., and Head, A. (2024). Property-based testing in practice. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, pages 1–13.

Gu, S. and Dodoo, R. N. A. (2020). The impact of firm performance on annual report readability: evidence from listed firms in ghana. *Journal of Economics, Business, & Accountancy Ventura*, 22(3):444–454.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Huang, R., Li, M., Yang, D., Shi, J., Chang, X., Ye, Z., Wu, Y., Hong, Z., Huang, J., Liu, J., et al. (2024). Audiogpt: Understanding and generating speech, music, sound, and talking head. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23802–23804.

Jiménez-Luna, J., Grisoni, F., and Schneider, G. (2020). Drug discovery with explainable artificial intelligence. nat mach intell 2: 573–584.

Kingma, D. P. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koivisto, J. and Hamari, J. (2019). The rise of motivational information systems: A review of gamification research. *International journal of information management*, 45:191–210.

Kromidha, E. (2023). Identity mediation strategies for digital inclusion in entrepreneurial finance. *International Journal of Information Management*, 72:102658.

Kulkarni, A. A., Niranjan, D. G., Saju, N., Shenoy, P. R., and Arya, A. (2024). Graph-based fault localization in python projects with class-imbalanced learning. In *International Conference on Engineering Applications of Neural Networks*, pages 354–368. Springer.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

Liao, K., Nie, L., Lin, C., Zheng, Z., and Zhao, Y. (2023). Recrecnet: Rectangling rectified wide-angle images by thin-plate spline model and dof-based curriculum learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10800–10809.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Lo, K., Wang, L. L., Neumann, M., Kinney, R., and Weld, D. S. (2019). S2orc: The semantic scholar open research corpus. *arXiv preprint arXiv:1911.02782*.

Marvin, G., Hellen, N., Jjingo, D., and Nakatumba-Nabende, J. (2023). Prompt engineering in large language models. In *International conference on data intelligence and cognitive informatics*, pages 387–402. Springer.

Matthews, N. and Folivi, F. (2023). Omit needless words: Sentence length perception. *PloS one*, 18(2):e0282146.

McElfresh, D. C., Chan, L., Doyle, K., Sinnott-Armstrong, W., Conitzer, V., Borg, J. S., and Dickerson, J. P. (2021). Indecision modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5975–5983.

Mikk, J. (2008). Sentence length for revealing the cognitive load reversal effect in text comprehension. *Educational Studies*, 34(2):119–127.

Reddy, T., Williams, R., and Breazeal, C. (2022). Levelup–automatic assessment of block-based machine learning projects for ai education. In *2022 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 1–8. IEEE.

Shah, S., Manzoni, S. L., Zaman, F., Es Sabery, F., Epifania, F., and Zoppis, I. F. (2024). Fine-tuning of distil-bert for continual learning in text classification: An experimental analysis. *IEEE Access*.

Shewale, C., Sardeshmukh, A., Shinde, P., Sapkal, O., and Shinde, S. (2024). Intelligent system for crop recommendation and disease identification. In *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, pages 1–5. IEEE.

Simon, Sheard, J., Luxton-Reilly, A., and Szabo, C. (2023). Computing education research in australasia. In *Past, Present and Future of Computing Education Research: A Global Perspective*, pages 373–394. Springer.

Sodian, L., Wen, J. P., Davidson, L., and Loskot, P. (2022). Concurrency and parallelism in speeding up i/o and cpu-bound tasks in python 3.10. In *2022 2nd International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)*, pages 560–564. IEEE.

Stankevičius, L. and Lukoševičius, M. (2024). Extracting sentence embeddings from pretrained transformer models. *arXiv preprint arXiv:2408.08073*.

Thorpe, L. (2023). What are lexical density and lexical diversity? *ReadabilityFormulas.com*.

To, V., Fan, S., and Thomas, D. (2013). Lexical density and readability: A case study of english textbooks. *Internet Journal of Language, Culture and Society*, 37(37):61–71.

Vulić, I., Ponti, E. M., Litschko, R., Glavaš, G., and Korhonen, A. (2020). Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240.

Yaffe, P. (2022). Fog index: is it really worth the trouble? *Ubiquity*, 2022(October):1–4.

Yu, L., Liu, B., Lin, Q., Zhao, X., and Che, C. (2024). Semantic similarity matching for patent documents using ensemble bert-related model and novel text processing method. *arXiv preprint arXiv:2401.06782*.

Yudistira, N., Kavitha, M. S., and Kurita, T. (2022). Weakly-supervised action localization, and action recognition using global–local attention of 3d cnn. *International Journal of Computer Vision*, 130(10):2349–2363.

Zulfiqar, A., Pfaff, B., Tu, W., Antichi, G., and Shahbaz, M. (2023). The slow path needs an accelerator too! *ACM SIGCOMM Computer Communication Review*, 53(1):38–47.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS