applied sciences

MDPI

Article

Automated Assessment of Inferences Using Pre-Trained Language Models

Yongseok Yoo

School of Computer Science and Engineering, Soongsil University, Seoul 06978, Republic of Korea; yyoo@ssu.ac.kr; Tel.: +82-2-820-0678

Abstract: Inference plays a key role in reading comprehension. However, assessing inference in reading is a complex process that relies on the judgment of trained experts. In this study, we explore objective and automated methods for assessing inference in readers' responses using natural language processing. Specifically, classifiers were trained to detect inference from a pair of input texts and reader responses by fine-tuning three widely used pre-trained language models. The effects of the model size and pre-training strategy on the accuracy of inference classification were investigated. The highest F1 score of 0.92 was achieved via fine-tuning the robustly optimized 12-layer BERT model (RoBERTa-base). Fine-tuning the larger 24-layer model (RoBERTa-large) did not improve the classification accuracy. Error analysis provides insight into the relative difficulty of classifying inference subtypes. The proposed method demonstrates the feasibility of the automated quantification of inference during reading, and offers potential to facilitate individualized reading instructions.

Keywords: natural language processing; inference; reading; think aloud; language models

check for updates

Citation: Yoo, Y. Automated Assessment of Inferences Using Pre-Trained Language Models. *Appl. Sci.* **2024**, *14*, 3657. https://doi.org/10.3390/app14093657

Academic Editors: Cristina Portales, João M. F. Rodrigues and Pedro J. S. Cardoso

Received: 25 March 2024
Revised: 21 April 2024
Accepted: 23 April 2024
Published: 25 April 2024

1. Introduction

1.1. Automating the Assessment of Reading Comprehension Skills

Natural language processing (NLP) can improve productivity in many areas. This is especially true for organizations that deal with large volumes of textual data [1]. For example, NLP helps users to identify and extract key entities, facts, and relationships from text to build knowledge bases [2]. These knowledge bases become valuable resources for quick information access and decision making. In addition, in customer service, sentence classification can be used to automatically categorize customer inquiries, complaints, or feedback [3]. This enables quicker routing to the appropriate department or personnel, thereby reducing response times and improving customer satisfaction. Furthermore, sentiment analysis, a special case of sentence classification, can automatically identify positive, negative, or neutral sentiments in customer feedback or social media posts [4]. This helps organizations to quickly gauge public opinion and adjust their strategies accordingly.

Recent research has increasingly focused on using NLP to augment or even replace human expert judgments. In the legal field, NLP-powered tools are accelerating case preparation by efficiently sifting through large volumes of legal documents [5]. Similarly, in the financial sector, NLP-based systems are enhancing the ability of analysts to predict market trends by analyzing large and complex datasets [6], a task that is beyond human capabilities. Similarly, in education, automated essay grading systems provide instant feedback [7], thus allowing educators to spend more time on individualized instructions.

This study focuses on automating the assessment of reading comprehension skills through the development and application of advanced NLP in education. Reading is a fundamental skill for the acquisition of knowledge [8]. Assessing reading skills is challenging because reading involves multiple cognitive processes, such as letter–sound correspondence, phonological memory, word recognition, sentence processing, and comprehension [9]. These individual skills are assessed using standardized tests [10]. The objectivity of standardized tests makes it possible to compare the reading skills of students from different backgrounds using the same criteria. However, standardized tests often focus on specific types of reading skills, potentially neglecting broader aspects of literacy, such as critical thinking and engagement with diverse background knowledge [11]. To gain a holistic understanding of a student's reading skills, the interpretation of test results and personalized feedback from human experts are essential.

Machine learning models are actively used to improve the accuracy of reading level assessments. In [12], Petersen and Ostendorf used support vector machines to combine lexical and syntactic features of a given text to assess its reading level. The corpus used for this study was *Weekly Readers*, which consisted of 2400 articles from an educational magazine designed for children of ages 7–10 in the United States. Their model outperformed the traditional Flesch–Kincaid score [13], which is based on sentence length and the average syllable count. In [14], Boonthum-Denecke et al. used latent semantic analysis (LSA) [15] and word matching to first identify students' reading strategies and then estimate their reading comprehension scores. The correlation between predicted and actual reading scores was low (<0.5), indicating that estimating reading comprehension skills is a difficult task. In [16], Allen et al. used a linguistic measure of coherence, called the Coh-Matrix [17], and LSA to assess the reading comprehension of high school students. Their predictions were compared with standardized scores, using the Gates-MacGinitie reading skill test, level 10/12 [18], and the correlation between them was low (<0.5).

However, existing studies using machine learning models combine known linguistic features to improve the overall accuracy of reading assessments. In contrast, this study investigates a way to use pre-trained language models to assess a specific cognitive task (inference) for reading. This would support human experts in assessing reading skills by providing tools and platforms that streamline assessment processes using NLP.

1.2. Contributions of the Study

Specifically, this study was motivated by the fact that inference has been shown to play a key role in reading comprehension, serving as a critical component in the construction of meaning from text [19–21]. Inference allows readers to fill in gaps in explicit textual information, facilitating the deeper understanding and integration of knowledge. Previous research emphasizes its importance not only for comprehending literal content, but also for engaging with the text at a deeper level [20], allowing for the application of prior knowledge and the anticipation of subsequent narrative developments [21]. Enhancing inferential skills could significantly improve reading comprehension outcomes, thereby advancing literacy education practices.

However, the evaluation of inferences during reading is a complex process and is subject to the judgment of human experts. Currently, readers' cognitive processes during reading are monitored by the think-aloud protocol [22–24]. First, readers read a given text and verbally report their thoughts (Figure 1A). Then, the readers' responses are transcribed and evaluated by multiple evaluators (Figure 1B). This traditional method relies heavily on qualitative analysis and the interpretive insights of the raters, which leads to inter-rater variability and limits the scalability of the evaluation. This variability and the high resource requirements underscore the need for the development of more objective and automated methods for assessing inference in readers' responses [25].

Thus, this study attempts to address these challenges by using NLP to automatically evaluate reader responses. Specifically, we formulate the problem as sentence classification, where a classifier is trained to classify a pair of input texts and reader responses as inferential or non-inferential (Figure 1C). The classifier is fine-tuned from pre-trained language models that have been trained with large text corpora. These pre-trained models efficiently produce the contextualized representations of a given text and are a good starting point for developing a task-specific language model [26]. Further training on smaller, more focused datasets which are relevant to the task at hand (inference classification) would allow for the automated quantification of readers' responses to a given text.

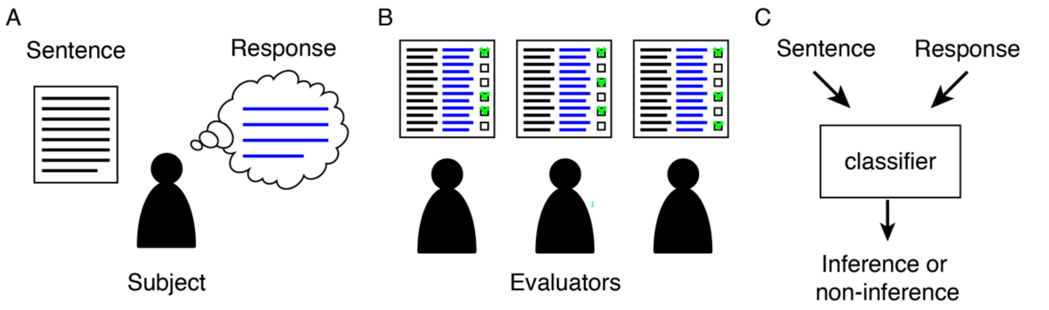

Figure 1. Using the think-aloud protocol, the subject's response to each sentence is collected (**A**). Human evaluators then assess the sentence–response pairs to determine whether an inference was drawn (**B**). The proposed method is used to classify a sentence–response pair as inferential or not, without having to involve human experts (**C**).

The remainder of this paper is organized as follows. Section 2 describes the data collection, human expert evaluation, and the procedure of fine-tuning pre-trained language models and their evaluation. In Section 3, we assess the accuracy of the inference classification using the proposed method, and we analyze the error patterns. Then, in Section 4, we discuss the results and their implications. In Section 5, we draw general conclusions with future research directions.

2. Materials and Methods

2.1. Data Collection and Evaluation by Human Experts

A dataset of 720 sentence–response pairs was collected as follows. The stimulus text in Korean consisted of 10 sentences taken from an elementary school reading textbook. The average number of words per sentence was 9.2. The think-aloud protocol was administered to 72 third- and fourth-grade elementary school students in public elementary schools in South Korea. The participant's verbal response to each sentence was recorded and transcribed.

Among the total 720 sentence–response pairs, 58 pairs were removed because the reader failed to produce any meaningful response. The remaining 662 sentence–response pairs were used for further analysis. Three evaluators individually assessed each sentence–response pair according to the nine inference types defined in a previous study [21] (summarized in Table 1). Each evaluator first identified the type of inference and then determined whether an inference was made (labeled as 1) or not (labeled as 0). The evaluators' decisions agreed on 642 sentence–response pairs (97%). For the mismatched 20 pairs (3%), evaluators discussed until they reached the same conclusion. Among the 662 sentence–response pairs, 438 pairs were labeled as inference (1), and 224 were labeled as no inference (0).

Table 2 shows typical examples of sentence–response pairs in which inferences were made. The evaluators agreed that, in the first example in Table 2, the inference was made because the reader tried to make a prediction about what would come next based on what had come before. In the second example in Table 2, the reader tried to connect the background knowledge (The plants seem to be inactive in general.) when explaining the given sentence. In the third example in Table 2, the reader tried to associate the information in the given sentence with their background knowledge (The Venus firetrap is a famous example of a plant that eats insects.). In the fourth example in Table 2, the reader's response was a simple paraphrase of the given sentence, and the sentence and response were almost identical in meaning. In the last example in Table 2, there was no meaningful response from the reader.

Table 1. Inference types and definitions.

Inference	Inference Type	Definition
Inference	elaboration	explanation about the contents of the current sentence using background knowledge
	prediction	anticipation of what will occur next in the text
	association	concept from background knowledge, brought to mind by the text
	bridging	connecting contents of the current sentence with local/near or global/distant text information
	metacognitive response	reflection of understanding or agreement with the text
	evaluative comment	opinion about the text
	affective response	emotion related to the contents of the text
Non-inference	paraphrase	putting the current sentence or part of the current sentence into their own words, or restating the text verbatim
	meaningless response	no response or meaningless response

Table 2. Examples of sentence–response pairs in which inferences were made.

Sentence	Response	Inference Type
Plants prevent insects from eating them in several ways.	I think plants discourage insects from eating them in some ways, such as thorns or bad smell.	prediction
Plants like roses produce thorns to keep insects away.	A rose cannot do anything because it's a plant, but it has a lot of thorns. When insects come, they get stung and die. So, the insects cannot eat the plant.	elaboration
Unusually, there are plants that eat insects.	I think it refers to some plants that eats insects, like the Venus flytrap!	association
Plants prevent insects from eating them in several ways.	It means that plants have different ways of preventing insects from trying to eat them.	paraphrase
Plants like roses produce thorns to keep insects away.	Hmm... I do not know.	meaningless response

2.2. Pre-Trained Language Models

In this study, three widely used Transformer-based pre-trained language models [27] were selected to investigate their effectiveness for classifying inference. Due to the relatively small sample size (662) when compared to the huge parameter space of Transformer-based models, we decided to use BERT [24] and its variants in order to avoid overfitting. These models, namely the BERT-base [28], RoBERTa-base [29], and RoBERTa-large [29], have become fundamental in the field of natural language understanding due to their success in capturing complex patterns and representations, achieved through extensive pre-training on large corpora. The selection of these models allows for a comprehensive analysis of their performance and adaptability in fine-tuning scenarios for inference classification. Each model brings its own set of characteristics, such as model size, masking, and training objectives, which are summarized in Table 3.

Table 3. Comparison of pre-trained models' architecture and training objectives.

Model	BERT-Base	RoBERTa-Base	RoBERTa-Large
Number of layers	12	12	24
Number of parameters	110 million	110 million	335 million
Masking	static	dynamic	dynamic
Next sentence prediction	included	removed	removed

Korean versions of the three base models were pre-trained using the KLUE (Korean Language Understanding Evaluation) benchmark dataset [30]. The benchmark contains eight Korean natural language understanding tasks, including topic classification, semantic textual similarity, natural language inference, named entity recognition, relation extraction, dependency parsing, machine reading comprehension, and dialogue state tracking. The corpora used for this benchmark include news headlines, Wikipedia, Wikinews, policy news, The Korea Economics Daily News, and Acrofan News for formal texts, and ParaKQC, Airbnb reviews, and the NAVER Sentiment Movie Corpus for colloquial texts. The tokenizer for the dataset is a morpheme-based sub-word tokenizer [30], which first divides an input text into morphemes using a morphological analyzer, and then tokenizes them using the byte pair encoding (BPE) technique [31]. The pre-trained models and the tokenizer are publicly available [32–34].

2.3. Transfer Learning and Evaluation Using k-Fold Cross-Validation

The inference classifiers were trained by fine-tuning the three pre-trained language models as follows. Each pair of input sentences and corresponding responses was concatenated with the separator symbol ([SEP]) between them. The combined text was then provided as the input. For each base model, the last layer (classification head) was replaced with a dense layer with the output size set to one, and its weights were initialized with random values. As a result, the model will produce a single number for each input, and this output is used as the logit of the target class. The loss is defined by the binary cross entropy between the output of the model and true class label as follows:

$$\begin{aligned} \mathrm{Loss} &= -(\mathrm{y}\log(\sigma(\mathrm{z})) + (1-\mathrm{y})\log(\sigma(\mathrm{z}))), \\ \sigma(\mathrm{z}) &= \tfrac{1}{1+e^{-z}}, \end{aligned}$$

where y is the true class label (0 or 1), z is the model output, and σ is the sigmoid function that transforms the input logit into a probability. This loss is minimized using the Adam optimizer [35].

Hyperparameters for the training process were optimized as follows: As suggested by the authors of the pre-training models [26], the batch size, learning rate, warm-up ratio, and weight decay were varied independently, and the highest F1 score for each model was reported. First, the batch size was set to either 4, 8, 16, and 32. Larger batch sizes provide a more accurate estimate of the gradient, resulting in more stable training. However, they require more memory and processing power. Second, the learning rate was set to either 10^{-5}, 2×10^{-5}, 3×10^{-5}, and 5×10^{-5}. Too low a learning rate can lead to a long convergence time or to being fixed at a local minimum, while too high a learning rate can cause the training to be unstable and diverge. Third, the warm-up ratio was set to either 0, 0.1, 0.2, and 0.6. During the warm-up iterations, the learning rate gradually increased from zero to the target learning rate. This technique helps to stabilize the fine-tuning in the early iterations. Fourth, the weight decay was set to either 0, 0.01, 0.02, 0.04, and 0.08. The weight decay regularizes the model and prevents overfitting by penalizing large weights. The four hyperparameters varied independently, resulting in a total of 320 configurations.

For each base model, five-fold cross-validation was performed for each training configuration in order to rigorously assess the classification accuracy and generalization ability of the model. Specifically, the dataset was randomly divided into five distinct subsets, so that the proportions of the inference subtypes were equal. Four of these subsets were used for training and the remaining one was used for validation in order to calculate the F1 score with corresponding precision and recall scores. This cycle was repeated five times, with each subset serving as a validation set once. As a result, five F1 scores were collected for each training configuration. The configuration corresponding to the highest average F1 score was found for each base model. The five F1 scores of the best model were compared for models that were fine-tuned from different base models, using the paired *t*-test.

3. Results

3.1. Accuracy of Inference Classification

Table 4 shows the hyperparameters that resulted in the best models. Different batch sizes yielded the highest F1 scores for different base models. In contrast, the same learning rate (10^{-5}) and warm-up ratio (0) corresponded to the highest F1 scores. Here, the best warm-up ratio of zero means that the warm-up was not necessary. A non-zero weight decay was useful only for the largest model (RoBERTa-large). This suggests that the regularization via weight decay was effective only for the largest model.

Table 4. Hyperparameters for the best models.

Base Model	Batch Size	Learning Rate	Warm-Up Ratio	Weight Decay
BERT-base	16	10^{-5}	0	0
RoBERTa-base	8	10^{-5}	0	0
RoBERTa-large	32	10^{-5}	0	0.02

Table 5 shows the classification accuracies of the best models fine-tuned from the three pre-trained models. Both the precision and recall scores of the RoBERTa-base model were higher than those of the BERT-base model. As a result, the F1 score of the RoBERTa-base model was higher than that of the BERT-base model. The precision score of the RoBERTa-large model was lower than that of the RoBERTa-base model, but the recall scores of the two models were the same. As a result, the F1 score of the RoBERTa-large model was lower than that of the RoBERTa-base model.

Table 5. Accuracies of the best models.

Base Model	Precision	Recall	F1
BERT-base	0.89	0.86	0.87
RoBERTa-base	0.91	0.94	0.92
RoBERTa-large	0.87	0.94	0.90

Figure 2 shows the F1 scores of the inference classification by fine-tuning the three pre-trained language models. First, training the inference classifier from the BERT-base model resulted in an average F1 score of 0.87, with a standard error of the mean (SEM) of 0.01. Second, training the inference classifier from the RoBERTa-base model resulted in a higher average F1 score of 0.92, with an SEM of 0.01. The F1 score of the fine-tuned RoBERTa-base model was significantly higher than that of the fine-tuned BERT-base model (paired *t*-test, $p < 0.05$). Third, training the classifier from the RoBERTa-large model resulted in a lower average F1 score of 0.90, with a higher SEM of 0.02. However, the difference in the F1 scores of the fine-tuned RoBERTa-base and RoBERTa-large models were not statistically significantly (paired *t*-test, $p > 0.05$).

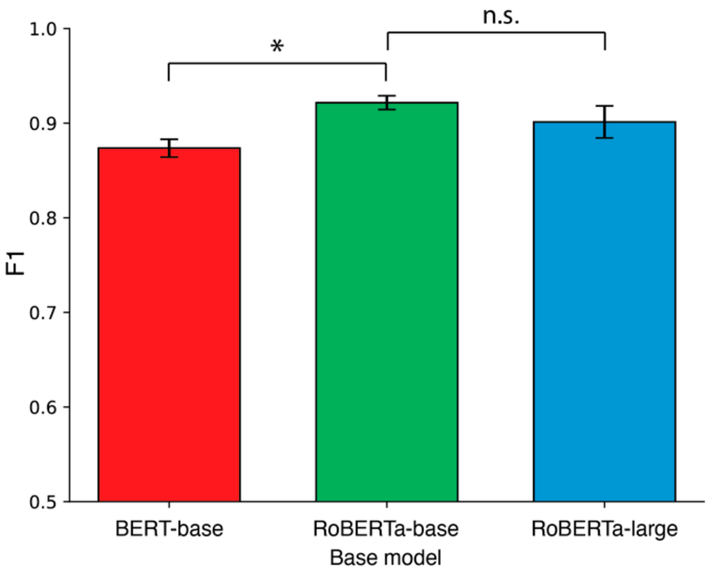

Figure 2. F1 scores of inference classification using different base models. Each error bar represents the standard error of the mean (SEM) measured from five-fold cross-validation. The asterisk corresponds to a statistically significant difference (paired *t*-test, $p < 0.05$), and n.s. indicates that the difference is not statistically significant (paired *t*-test, $p > 0.05$).

3.2. Error Analysis

Figure 3 shows the proportions of the inference subtypes in the errors. The proportions of the inference subtypes in the entire dataset are shown as gray bars for reference. The proportions of the inference subtypes in inaccurate predictions are shown in red, green, and blue for the BERT-base, RoBERTa-base, and RoBERTA-large models, respectively.

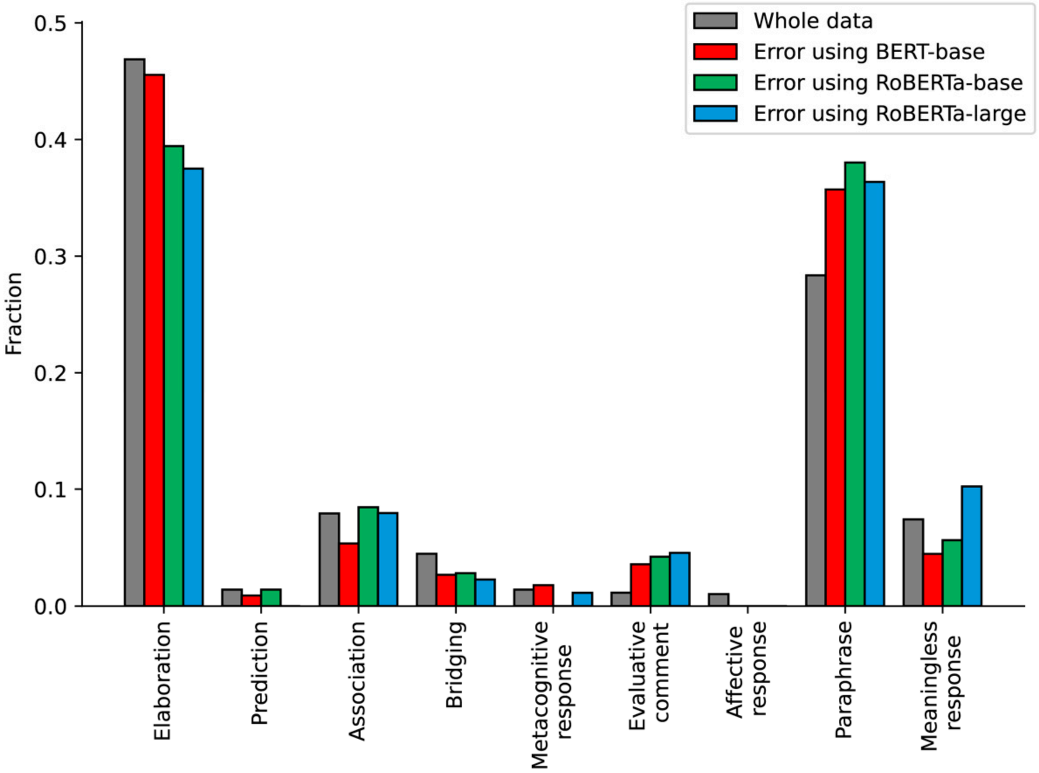

Figure 3. The proportions of the inference subtypes in the entire dataset (gray) and in the errors made by fine-tuned models using different base models (red, green, and blue).

A comparison of the proportions of the inference subtypes in the errors with the proportion in the entire dataset shows the relative difficulty of classifying the inference subtypes. The proportion of elaboration and bridging subtypes in the errors decreased as larger language models were used. This suggests that larger language models classify these types of inferences with a higher accuracy. In contrast, the proportions of evaluative comments and paraphrases increase as larger language models are used. This suggests that these inference subtypes are more susceptible to overfitting. Furthermore, the largest language model (RoBERTa-large) made more errors in classifying meaningless responses than the other two language models (BERT-base and RoBERTa-base). This is another indication of overfitting.

4. Discussion

4.1. Effects of Pre-Trained Language Models on the Classification Accuracy

The accuracy of inference classification is significantly influenced by the choice of the base pre-trained language model. The comparison of the F1 scores based on the BERT-base and RoBERTa-base models shows the effect of the training strategies used during the pre-training phase of a language model on the classification accuracies of the fine-tuned models. The more advanced training strategies used for the RoBERTa-base model resulted in significantly higher F1 scores for the inference classification than that of the BERT-base model of the same size. This is consistent with previous findings which indicate that models trained with RoBERTa-base models outperform models trained with BERT-base models for various downstream tasks [22].

Furthermore, the comparison of the F1 scores based on the RoBERTa-base and RoBERTa-large models suggests that the larger model is not necessarily better for classifying inference in the current dataset. The RoBERTa-base and RoBERTa-large models share the same Transformer architecture and training objective (masked language model with dynamic masking), but the main difference between the two models is the number of layers (12 vs. 24) and the model sizes (110 million vs. 355 million parameters). Despite the greater representational capacity, the RoBERTa-large model did not perform significantly better than the smaller model with the same architecture. The average F1 score became even lower, and the SEM of the F1 scores became larger. The error analysis using the inference subtypes shows that the largest language model (RoBERTa-large) made more errors in classifying evaluative comments, paraphrases, and meaningless responses.

This lower accuracy of the larger model is probably due to overfitting. In Table 4, the optimal weight decay value was zero for RoBERTa-base and non-zero for RoBERTa-large. This suggests that regularization with weight decay worked for RoBERTa-large, yielding a higher F1 score than RoBERTa-large without weight decay. Without weight decay, the F1 score of RoBERTa-large would be even lower than that of RoBERTa-base. More data are needed to train the larger model, and a well-trained small model (RoBERTa-base) would be the preferred choice for inference classification with a relatively small dataset.

4.2. Insights from the Error Analysis for Automating Inference Classification

Error analysis using the inference subtypes provides further insights into automated inference classification as follows:

Elaboration was the most common subtype of inference in the dataset, and the error rates of elaboration decreased for more complex (RoBERTa-base) and larger (RoBERTa-large) models. Given enough elaboration samples for training, scaling up the model could further improve the classification accuracy of elaboration.

Paraphrase, the second most common inference subtype in the dataset, shows a different pattern. The error rates when analyzing paraphrases using all the three models were higher than the proportion in the dataset, and there was no clear order among the models. This suggests that fine-tuning the pretrained models is not effective for paraphrase classification.

The accuracies for classifying elaboration and inference seem to be in a trade-off relationship. Because paraphrase is classified as non-inferring, sentence pairs which are too close are classified as negative samples. This is the opposite of the typical applications that measure the similarity between sentences and look for meaningfully related sentences. To correctly classify paraphrases as non-inferring, the classifier should reject similar sentence pairs. However, this would reduce the accuracy of classifying elaborations. Solving this problem would require new model architectures or training strategies, which are topics for future research.

For the classification of evaluative comments and meaningless responses, the error rates increased for more complex (RoBERTa-base) and larger (RoBERTa-large) models. The low accuracy in classifying evaluative comments would be due to a lack of samples. In contrast, there were more samples of meaningless responses, but the error rates increased much more for the largest model (RoBERTa-large). These different patterns in error rates between the different subtypes suggest that the degree of overfitting varies for different inference subtypes.

5. Conclusions

In this study, we investigated the feasibility of using language models to classify the inferences of sentence–response pairs. The proposed method achieved high F1 scores by fine-tuning a Transformer-based pre-trained language models. Specifically, the highest F1 score, 0.92, was achieved by fine-tuning RoBERTa-base, which was higher than that of a model with the same size fine-tuned from BERT-base. A larger language model (RoBERTa-large) did not increase the classification accuracy. This suggests that choosing pre-trained language models of high quality and appropriate size is important.

The proposed method would allow the automated quantification of reader responses to a given text and improve the effectiveness of the think-aloud protocol. In the think-aloud protocol, reader responses are open ended and provide rich information which requires a trained expert to evaluate. This study opens the possibility of simulating trained experts in evaluating inferences, which is one of the key qualities of an effective reader.

Further work could be conducted in three ways. First, expanding the genre of the stimulus text would be a natural next step. Different genres have unique structures, styles, and conventions that would affect reading behaviors. Therefore, confirming the feasibility of our method for different genres would test the generality of our method. Second, classifying inference subtypes would be an interesting future work. This requires a larger dataset, so that enough samples are collected for each inference subtype. Third, another future research direction is to adapt the proposed method to educational settings. For example, the automated inference classification of user responses could enrich interactive learning by tailoring content to meet individual learning needs and preferences. We are eager to further explore applications and improve learning outcomes.

Funding: This research received no external funding.

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of Chonnam National University (1040198-170920-HR-074-02) on 13 October 2017.

Informed Consent Statement: A comprehensive written consent form was provided to all participants and their parents. These forms outlined the objectives of the research, the extent of participation, potential benefits and risks, assurances of confidentiality, and the right of the participants to withdraw from the study at their discretion.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Strzalkowski, T. *Natural Language Information Retrieval*; Springer: Dordrecht, The Netherlands, 1999.
2. Yunshi, L.; He, G.; Jiang, J.; Jiang, J.; Zhao, W.X.; Wen, J.-R. A Survey on Complex Knowledge Base Question Answering: Methods, Challenges and Solutions. In Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence Survey Track, Montreal, QC, Canada, 19–27 August 2021; International Joint Conferences on Artificial Intelligence Organization: Santa Clara, CA, USA, 2021; pp. 4483–4491.
3. Jochen, H.; Heitmann, M.; Siebert, C.; Schamp, C. More than a feeling: Accuracy and application of sentiment analysis. *Int. J. Res. Mark.* **2023**, *40*, 75–87.
4. Koyel, C.; Bhattacharyya, S.; Bag, R. A survey of sentiment analysis from social media data. *IEEE Trans. Comput. Soc. Syst.* **2020**, *7*, 450–464.
5. Ilias, C.; Jana, A.; Hartung, D.; Bommarito, M.; Androutsopoulos, I.; Katz, D.; Aletras, N. LexGLUE: A Benchmark Dataset for Legal Language Understanding in English. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland, 22–27 May 2022; Volume 1, pp. 4310–4330.
6. Zhuang, L.; Huang, D.; Huang, K.; Li, Z.; Zhao, J. Finbert: A pre-trained financial language representation model for financial text mining. In Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, Yokohama, Japan, 7–15 January 2021; pp. 4513–4519.
7. Dadi, R.; Sanampudi, S.K. Automated essay scoring systems: A systematic literature review. *Artif. Intell. Rev.* **2022**, *55*, 2495–2527.
8. Snowling, M.J.; Hulme, C. *The Science of Reading: A Handbook*; Blackwell Publishing: Oxford, UK, 2005.
9. Duke, N.K.; Cartwright, K.B. The science of reading progresses: Communicating advances beyond the simple view of reading. *Read. Res. Q.* **2021**, *56*, S25–S44. [CrossRef]
10. Riazi, A.M. *The Routledge Encyclopedia of Research Methods in Applied Linguistics*; Routledge: New York, NY, USA, 2016.
11. Gallagher, K. *Readicide: How Schools Are Killing Reading and What You Can Do about It*; Routledge: New York, NY, USA, 2009.
12. Petersen, S.E.; Ostendorf, M. A machine learning approach to reading level assessment. *Comput. Speech Lang.* **2009**, *23*, 89–106. [CrossRef]
13. Kincaid, J.P., Jr.; Fishburne, R.P.; Rodgers, R.L.; Chisson, B.S. *Derivation of New Readability Formulas for Navy Enlisted Personnel*; Research Branch Report 8-75; US Naval Air Station: Memphis, TN, USA, 1975.
14. Boonthum-Denecke, C.; McCarthy, P.; Lamkin, T.; Jackson, G.T.; Magliano, J.P.; McNamara, D.S. Automatic natural language processing and the detection of reading skills and reading comprehension. In Proceedings of the Twenty-Fourth International FLAIRS Conference, Palm Beach, FL, USA, 18–20 May 2011.
15. Landauer, T.; McNamara, D.; Dennis, S.; Kintsch, W. *Handbook of Latent Semantic Analysis*; Psychology Press: Mahwah, NJ, USA, 2013.
16. Allen, L.K.; Snow, E.L.; McNamara, D.S. Are you reading my mind? Modeling students' reading comprehension skills with Natural Language Processing techniques. In Proceedings of the Fifth International Conference on Learning Analytics and Knowledge, Poughkeepsie, NY, USA, 16–20 March 2015.
17. McNamara, D.S.; Graesser, A.C.; McCarthy, P.; Cai, Z. *Automated Evaluation of Text and Discourse with Coh-Metrix*; Cambridge University Press: Cambridge, UK, 2014.
18. MacGinitie, W.H. *Gates-MacGinitie Reading Tests*; Houghton Mifflin: Boston, MA, USA, 1989.
19. Yuill, N.; Oakhill, J. *Children's Problems in Text Comprehension: An Experimental Investigation*; Cambridge University Press: New York, NY, USA, 2010.
20. Cain, K.; Oakhill, J.V. Inference making and its relation to comprehension failure in young children. *Read. Writ. Interdiscip. J.* **1999**, *11*, 489–503. [CrossRef]
21. Bowyer-Crane, C.; Snowling, M.J. Assessing children's inference generation: What do tests of reading comprehension measure? *Br. J. Educ. Psychol.* **2005**, *75*, 189–201. [CrossRef] [PubMed]
22. Ericsson, K.A.; Simon, H.A. *Protocol Analysis: Verbal Reports as Data*; MIT Press: London, UK, 1993.
23. McMaster, K.L.; van den Broek, P.; Espin, C.A.; White, M.J.; Rapp, D.N.; Kendeou, P.; Bohn-Gettler, C.M.; Carlson, S. Making the right connections: Differential effects of reading intervention for subgroups of comprehenders. *Learn. Individ. Differ.* **2012**, *22*, 100–111. [CrossRef]
24. Carlson, S.E.; Seipel, B.; McMaster, K. Development of a new reading comprehension assessment: Identifying comprehension differences among readers. *Learn. Individ. Differ.* **2014**, *32*, 40–53. [CrossRef]
25. Bowles, M.A. *The Think-Aloud Controversy in Second Language Research*; Routledge: New York, NY, USA, 2010.
26. Bonan, M.; Ross, H.; Sulem, E.; Veyseh, A.P.B.; Nguyen, T.H.; Sainz, O.; Agirre, E.; Heintz, I.; Roth, D. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Comput. Surv.* **2023**, *56*, 1–40.
27. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
28. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
29. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
30. Park, S.; Moon, J.; Kim, S.; Cho, W.I.; Han, J.; Park, J.; Song, C.; Kim, J.; Song, Y.; Oh, T.; et al. KLUE: Korean language understanding evaluation. *arXiv* **2021**, arXiv:2105.09680.

31. Sennrich, R.; Haddow, B.; Birch, A. Neural machine translation of rare words with subword units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; Association for Computational Linguistics: Kerrville, TX, USA, 2016; Volume 1, pp. 1715–1725.
32. BERT-Base Model Pretrained on KLUE. Available online: https://huggingface.co/klue/bert-base (accessed on 1 March 2024).
33. RoBERTa-Base Model Pretrained on KLUE. Available online: https://huggingface.co/klue/roberta-base (accessed on 1 March 2024).
34. RoBERTa-Large Model Pretrained on KLUE. Available online: https://huggingface.co/klue/roberta-large (accessed on 1 March 2024).
35. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.