HIJLI_JU at SemEval-2024 Task 7: Enhancing Quantitative Question Answering Using Fine-tuned BERT Models

Partha Sarathi Sengupta
Computer Science and Engineering
Jadavpur University, Kolkata
jitendriyo@gmail.com

Sandip Sarkar
Computer Science and Application
Hijli College, Kharagpur
sandipsarkar.ju@gmail.com

Dipankar Das
Computer Science and Engineering, Jadavpur University, Kolkata
dipankar.dipnil2005@gmail.com

Abstract

In data and numerical analysis, Quantitative Question Answering (QQA) becomes a crucial instrument that provides deep insights for analyzing large datasets and helps make well-informed decisions in industries such as finance, healthcare, and business. This paper explores the "HIJLI_JU" team's involvement in NumEval Task 1 within SemEval 2024, with a particular emphasis on quantitative comprehension. Specifically, our method addresses numerical complexities by fine-tuning a BERT model for sophisticated multiple-choice question answering, leveraging the Hugging Face ecosystem. The effectiveness of our QQA model is assessed using a variety of metrics, with an emphasis on the f1_score() of the scikit-learn library. Thorough analysis of the macro-F1, micro-F1, weighted-F1, average, and binary-F1 scores yields detailed insights into the model's performance in a range of question formats.

1 Introduction

Quantitative Question Answering (QQA) is a crucial tool in the large field of data and numerical analysis because it uses sophisticated computer methods to extract and interpret significant information from large datasets. Imagine it as a strong force that has particular sway over important industries like finance, healthcare, and business, where it plays a crucial role in forecasting trends and assisting in the making of well-informed decisions. QQA acts as a driving force behind wise decisions by skillfully converting apparently complicated data into useful knowledge and clearing a way through the complexities of numerical data.

Our team, "HIJLI_JU," participated in Task 1 (Quantitative Understanding) of NumEval within SemEval 2024[1], thereby actively engaging in the competitive landscape. An annual series of international natural language processing (NLP) competitions called SemEval (Semantic Evaluation) evaluates the state-of-the-art in a range of tasks pertaining to semantic analysis and understanding. These challenges serve as a forum for practitioners and scholars from both academia and industry to investigate and expand the field of computational linguistics. SemEval is well known for its broad range of tasks, which address a variety of difficulties in natural language processing.

The presented QQA task in the context of the SemEval 2024 NumEval competition provides a platform for researchers and developers to showcase advancements in quantitative understanding (num). SemEval, an annual challenge, encompasses diverse language tasks, including sentiment analysis and word meanings, contributing to the ongoing progress in systems designed for language understanding and processing. Table 1 shows the example of the SemEval-2024 Task dataset.

This paper is organized as follows: in Section 2, a survey of related literature is presented, and in Section 3, a detailed description of the dataset is provided. Section 4 explores the details of our suggested model while Section 5 explains the experimental setup. Experiments using our model are shown in Section 6, and observations are discussed in Section 7. Bringing everything together, we wrap up the paper in Section 8 and offer some suggestions for future directions for study.

2 Related Work

The HIJLI_JU team participated in the IJCNLP-2017 Task 5 on Multi-choice Question Answering, focusing on vector representations and machine learning for classification (Sarkar et al., 2017). Their model, designed exclusively for English language questions. The methodology involves representing questions and answers in vector space, computing cosine similarity, and employing a classification approach to identify the correct answer

[1] https://sites.google.com/view/numeval/tasks

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 279–284

Task	Question	Answer
QP	FED'S DUDLEY REPEATS EXPECTS GDP GROWTH TO PICK UP IN 2014, FROM [Masked] PCT POST-RECESSION AVERAGE	1
QNLI	S1: Nifty traded above 7500, Trading Calls Today, S2: Nifty above 7400	Entailment
QQA	Elliot weighs 180 pounds whereas Leon weighs 120 pounds. Who has a bigger gravity pull? Option1: Elliot Option2: Leon	Option 1

Table 1: Task Questions and Answers

option.

Sandip presented a novel approach to enhance science-based Multiple Choice Question Answering (MCQA) systems by leveraging distributed semantic similarity and a classification approach. Three models (Model 1, Model 2, and Model 3) were developed to address differences in dataset formats, specifically focusing on IJCNLP Task 5 and SciQ datasets (Sarkar et al., 2020).

Zuccon looks into using fancy word techniques from computer language models for finding information better. They use special methods to understand words and put them into a translation model. The results show that this approach improves how well information is found, and it's flexible – it works well even if the word understanding is done differently or comes from a different set of information (Zuccon et al., 2015).

Researchers in the field of Quantitative Question Answering (QQA) have been investigating ways to enhance computer systems' capacity to respond to numerical questions. They've tried a range of tactics, such as deep learning and sophisticated machine learning. To improve response accuracy, certain studies might incorporate external data.

For QQA, standardized tests (datasets) covering a range of numerical questions in disciplines like science and finance are being developed. They're also developing equitable methods to evaluate the efficacy of these Q&A platforms.

The brittleness of existing AI systems, including large-scale language models, in arithmetic reasoning within natural language understanding is addressed by the proposed multi-task benchmark NUMGLUE (Mishra et al., 2022a). In order to prepare AI systems for increasingly difficult mathematical tasks, the benchmark attempts to promote the development of systems that are able to reason robustly about arithmetic in language.

EQUATE is a framework assessing quantitative reasoning in textual entailment for natural language understanding systems (Ravichander et al., 2019). State-of-the-art models don't consistently outperform a basic baseline, highlighting a potential gap in implicit quantity reasoning. This framework aims to spur the development of models focusing on quantitative reasoning in language understanding.

Chen and his colleague investigates whether neural network models can acquire numeracy skills, focusing on predicting numeral magnitudes in text (Chen et al., 2019). Introducing the Numeracy-600K benchmark dataset, the study explores various models. Additionally, they highlights a practical application scenario by demonstrating the task's utility in detecting exaggerated information.

Chen also addresses innumeracy issues in pretrained language models, focusing on the fundamental task of teaching language models to understand numerals in text (Chen et al., 2023). It suggests a method that combines a comparing-number task with number notation exploration, modification, and pre-finetuning. Their research shows enhanced performance in three benchmark datasets for tasks related to quantitative analysis, especially for RoBERTa.

question	Option1	Option2	answer	type
Jame's mother has a photo of Jane standing at a height of 14 inches, whereas a mountain appears to have height of 26 cm. It looks that way because?	the mountain was farther away	Jane was farther away	Option 2	Type_3
Tina is racing her two dogs. Her greyhound weighs 40 kgs, and her rottweiler weighs 35 kgs. The dog that gets faster more quickly is the?	rottweiler	greyhound	Option 1	Type_3
A toddler is rolling a ball for more than 1 mins on the grass and rolls it on to the sand where it stops after 43 seconds. The sand stopped the ball because it has _____ than the grass.?	more friction	less friction	Option 1	Type_3
The fish glided with a speed of 4 mph through the water and 1 mph through the jello because the _____ is smoother.?	jello	water	Option 2	Type_3

Table 2: Example of SemEval-2024 Task 7 Dataset

3 Dataset

Semeval-2024 Utilizing current benchmark datasets for three different task types—Quantitative Prediction (QP), Quantitative Natural Language Inference (QNLI), and Quantitative Question Answering (QQA)—is the one of the task of NumEval Task 1: Quantitative Understanding (Chen et al., 2023; Mishra et al., 2022b). Managing numbers, forecasting numerical values, deciphering logical connections in numerical sentences, and responding to inquiries requiring numerical data are all part of these tasks. The goal is to assess and improve the performance of models in handling these quantitative tasks.

The provided data format appears to represent a set of questions along with options, correct answers, and additional attributes [2], [3]. Table 2 shows the different fields of Task 1 of NumEval Dataset. On the other hand, Table 3 gives the description of the statics of the dataset. Here's a description of the key components in the data format:

- **"question":** The primary question text is contained in this field.
- **"Option1" and "Option2":** There are two options available to answer the question in these fields.
- **"answer":** Indicates which option is the correct answer.
- **"type":** Specifies the type of question.
- **"question_sci_10E":** Represents the same question as "question" but with numerical values expressed in scientific notation (e.g., 1.4000000000×10^1 inches).
- **"question_char":** Represents the same question with numerical values written as characters (e.g., "1 4 inches").
- **"question_sci_10E_char":** Combines scientific notation and characters for numerical values in the question.
- **"question_mask":** Presents the question with placeholders like "[Num]" indicating where numerical values are expected to be filled.

In summary, this data format is designed to offer various question formats, including options and the

[2] https://drive.google.com/drive/folders/10uQI2BZrtzaUejtdqNU9Sp1h0H9zhLUE?usp=sharing

[3] https://sites.google.com/view/numeval/data

right response, as well as various ways to represent numerical values (scientific notation, characters, masked placeholders, etc.). It appears to be intended as a test of numerical information interpretation and comprehension in various formats.

4 System Description

We explored the realms of natural language processing, immersing ourselves in Hugging Face's dynamic ecosystem known for its transformative libraries such as transformers and datasets. At the heart of our machine learning endeavor was the fine-tuning of a BERT model for nuanced multiple-choice question answering, including numerical complexities. Hugging Face's BertTokenizer and TFBertForMultipleChoice powered our training, and the fine-tuned model effortlessly transitioned into competent inference on the test dataset [4].

We started our data preparation process by adding "context" and "label" to JSON files. Next, we transformed the data into Dataset objects that were kept in a DatasetDict. We managed a preprocess_function, used BertTokenizer, and used the Datasets map method with 'batched=True' with caution to optimize our operations. Performance was improved by enabling dynamic sentence padding during collation using the Data Collator For Multiple Choice modification. Figure 1 shows the system description of HIJLI_JU for the participation in SemEval-2024 Task 7.

BERT, or Transformers' Bidirectional Encoder Representations, which had been pre-trained using masked language modeling and next sentence predictions on a substantial amount of unlabeled text data. Our approach was based on its bidirectional capabilities, which enabled it to simultaneously capture semantic subtleties from both sides.

5 Experimental Setup

We set up the parameters for training with a batch size of 16 over ten epochs, a starting learning rate of 0.00001, and no warm-up phases. Our datasets' dimensions, comprising 564 examples for training, 81 for development, and 162 for testing, demonstrated accuracy. The purpose of this rigorous training setup was to guarantee our QQA model's generalizability and robustness.

Using Jupyter Notebooks, Google Colab is a cloud-based platform that offers an interactive and collaborative environment for Python coding. Well-known for providing free access to GPU and TPU resources, Colab has grown to be a well-liked option in the data science and machine learning domains. Its smooth integration with Google Drive makes sharing and collaborative editing simple, which improves the effectiveness of team projects. Because of its intuitive interface and free availability of robust computational resources, Google Colab is an indispensable resource for individuals and groups working on a variety of computational tasks.

6 Results

In our quest to evaluate the efficacy of our Quantitative Question Answering (QQA) model, we employed a comprehensive set of metrics and examined its performance across various question formats. The `scikit-learn` library's `f1_score()` function served as our tool for this evaluation, offering insights into the model's proficiency in different contexts.

$$F1 = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

Here, TP represents the number of true positives, FP represents the number of false positives, and FN represents the number of false negatives.

The F1 Score is a fundamental metric in machine learning, providing a balanced evaluation of classification models by combining precision and recall. This versatile metric has several variants, each suited to different scenarios. In this essay, we delve into macro-F1, micro-F1, weighted-F1, average, and binary-F1, exploring their applications and significance.

6.1 Macro-F1 Score

The macro-F1 Score calculates the F1 Score for each class independently and then computes the unweighted average. This approach treats all classes equally, making it valuable when assessing a model's performance across diverse classes without bias towards larger ones.

[4] https://huggingface.co/docs/transformers/en/tasks/multiple_choice

Files	Size
QQA_train.json	564
QQA_dev.json	81
QQA_test.json	162

Table 3: Statistics of SemEval-2024 Task 7 Dataset

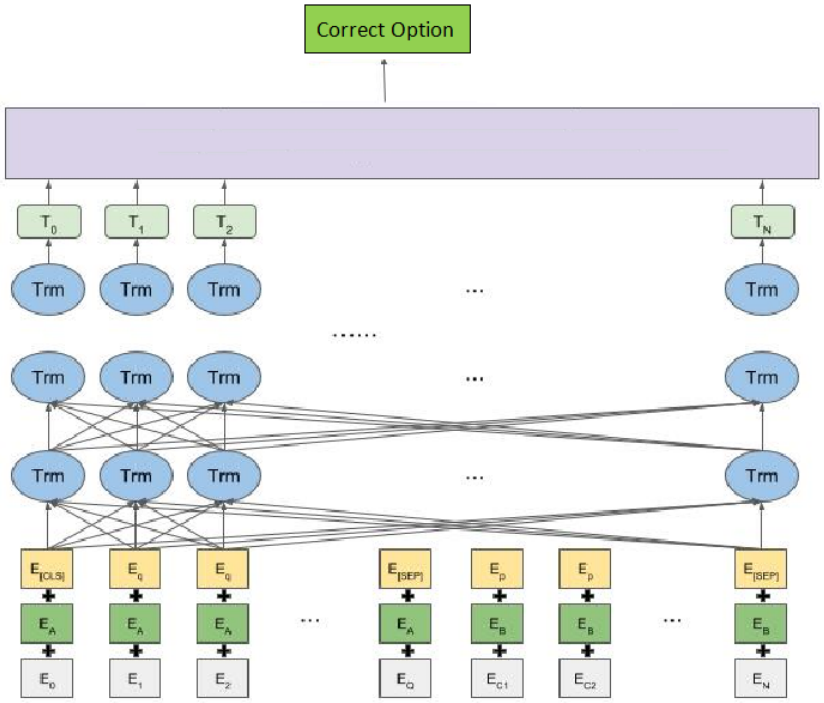

Figure 1: BERT Model

6.2 Micro-F1 Score

In contrast, the micro-F1 Score aggregates the contributions of all classes into a single F1 Score. Particularly useful for imbalanced datasets, it considers the varying sizes of different classes, providing an overall evaluation that accounts for class imbalances.

6.3 Weighted-F1 Score

The weighted-F1 Score extends the macro-F1 approach by considering class sizes. It calculates F1 Scores for each class and then computes a weighted average based on the number of instances in each class. This adjustment ensures that larger classes contribute proportionally more to the overall score.

6.4 Average F1 Score

The term "average F1 Score" is a general descriptor that encompasses various approaches to aggregating F1 Scores across multiple classes. It may refer to micro-F1, macro-F1, or other weighted or unweighted averages, depending on the context.

6.5 Binary F1 Score

The binary F1 Score is the traditional F1 Score applied to a binary classification problem with two classes – positive and negative.

7 Observations

The observed results highlight the nuanced performance of the Quantitative Question Answering (QQA) model across different question formats. Notably, questions presented in the character format consistently outperform other representations, demonstrating its robustness in handling diverse classes independently, particularly in imbalanced datasets. The Macro-F1, Micro-F1, and Weighted-F1 scores consistently identify the question_char format as the most effective in achieving a balanced evaluation. This format excels not only in independently handling varied classes but also in proportionally contributing to overall performance based on class instances. The Average F1 scores further affirm the versatility of the question_char format, emphasizing its capacity for a well-rounded

Field used	Macro-F1	Micro-F1	Weighted F1	average=None	Binary F1
question	0.50344	0.50617	0.50345	array([0.46667, 0.54023])	0.54023
question_char	0.53058	0.53704	0.53058	array([0.47552, 0.58564])	0.58564
question_sci_10E	0.44026	0.44444	0.44026	array([0.48864, 0.39189])	0.39189
question_sci_10E_char	0.51489	0.51852	0.51489	array([0.47297, 0.55682])	0.55682

Table 4: Result of HIJLI_JU on SemEval-2024 Task 7

evaluation across multiple classes.

8 Conclusion and Future Work

In conclusion, we found that Quantitative Question Answering (QQA) is like a helpful tool for understanding numbers better. It's useful in important areas like business, healthcare, and finance, helping with predicting trends and making smart decisions. QQA is like a guide that empowers organizations and people to understand and use tricky data. The research we did shows how important QQA is for understanding numbers better.

Looking ahead, there are exciting possibilities for more research on QQA. We could explore new tasks and find ways to use QQA in specific areas like healthcare. Working with experts in different fields could help make QQA more useful in different situations. Also, we can improve how we measure the success of QQA and make it better by using the latest technology and techniques in language understanding. This ongoing exploration will keep pushing QQA to new places and make it even more important in understanding both language and numbers.

References

Semeval-2024 task 7: Numeral-aware language understanding and generation.

Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2019. Numeracy-600K: Learning numeracy for detecting exaggerated information in market comments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6307–6313, Florence, Italy. Association for Computational Linguistics.

Chung-Chi Chen, Hiroya Takamura, Ichiro Kobayashi, and Yusuke Miyao. 2023. Improving numeracy by input reframing and quantitative pre-finetuning task. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 69–77, Dubrovnik, Croatia. Association for Computational Linguistics.

Swaroop Mishra, Arindam Mitra, Neeraj Varshney, Bhavdeep Sachdeva, Peter Clark, Chitta Baral, and Ashwin Kalyan. 2022a. Numglue: A suite of fundamental yet challenging mathematical reasoning tasks.

Swaroop Mishra, Arindam Mitra, Neeraj Varshney, Bhavdeep Sachdeva, Peter Clark, Chitta Baral, and Ashwin Kalyan. 2022b. NumGLUE: A suite of fundamental yet challenging mathematical reasoning tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3505–3523, Dublin, Ireland. Association for Computational Linguistics.

Abhilasha Ravichander, Aakanksha Naik, Carolyn Rose, and Eduard Hovy. 2019. EQUATE: A benchmark evaluation framework for quantitative reasoning in natural language inference. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 349–361, Hong Kong, China. Association for Computational Linguistics.

Sandip Sarkar, Dipankar Das, and Partha Pakray. 2017. JU NITM at IJCNLP-2017 task 5: A classification approach for answer selection in multi-choice question answering system. In *Proceedings of the IJCNLP 2017, Shared Tasks*, pages 213–216, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Sandip Sarkar, Dipankar Das, Partha Pakray, and David Eduardo Pinto Avendaño. 2020. Developing MCQA framework for basic science subjects using distributed similarity model and classification based approaches. *Int. J. Asian Lang. Process.*, 30(3):2050015:1–2050015:18.

Guido Zuccon, Bevan Koopman, Peter Bruza, and Leif Azzopardi. 2015. Integrating and Evaluating Neural Word Embeddings in Information Retrieval. In *Proceedings of the 20th Australasian Document Computing Symposium*, ADCS '15, pages 12:1–12:8, New York, NY, USA. ACM.