CSECU-DSG at SemEval-2026 Task 6: Imbalance-Aware Transformers for Unmasking Political Question Evasions

Subha Shesgin, Sumaiya Nazneen, and Abu Nowshed Chy
Department of Computer Science and Engineering
University of Chittagong, Chattogram-4331, Bangladesh
{subha.cse.cu, nsumaiya.cu}@gmail.com and nowshed@cu.ac.bd

Abstract

Clarity-Level Classification predicts the degree of clarity of a response to a query. It is essential to the advancement of many NLP applications, including conversational AI, customer support automation, and instructional technology. However, it is challenging to assess the clarity of an answer due to unclear wording, incomplete answers, and the contextual dependency in Q/A. In this paper, we present our approach to the shared task on clarity classification introduced in SemEval 2026 Task 6. We formulate the problem as both a regression and multi-class classification task over question–answer pairs and propose a transformer-based architecture to model their contextual interactions. Our system leverages a fine-tuned DeBERTa-v3-base backbone to capture nuanced semantic relationships between questions and answers. To mitigate class imbalance, we incorporate class-weighted loss functions and apply data-level oversampling techniques. Experimental results demonstrate that our approach achieves competitive performance in the shared task, highlighting the effectiveness of transformer-based contextual modeling combined with imbalance-aware training strategies.

1 Introduction

Strategic ambiguity is common in political discourse, especially in high-stakes situations like public press conferences, presidential debates, and interviews. Politicians often use evasive communication techniques that leave listeners wondering whether a question has been sufficiently answered. This phenomenon, also known as *evasion* or *equivocation*, happens when responses purposefully omit giving direct or unambiguous answers (Bull, 2003). Speakers can handle delicate subjects, keep their messaging flexible, and affect public opinion without making overt claims thanks to this ambiguity.

There are a number of difficulties in automatically identifying and categorizing response clarity. It can be challenging to define a label when responses contain both evasive and informative content. Because clarity depends so heavily on context, models must take into account both the question and its response. Additionally, there is a significant imbalance in the dataset provided for SemEval-2026 Subtask-1: clear replies and clear non-replies are underrepresented, while ambiguous or evasive responses predominate. This imbalance raises the possibility of biased predictions by making model evaluation and training more difficult.

In this work, we propose a transformer-based method for clarity-level classification for a given question-answer pair as defined in SemEval-2026 Task 6 Subtask 1 (Thomas et al., 2026). The task builds upon the CLARITY dataset and taxonomy introduced in prior work on response clarity classification (Thomas et al., 2024). To address the dataset imbalance, we use class-balancing techniques, including class-weighted loss and oversampling. Later, we fine-tune a pre-trained DeBERTa-v3-base model on question-answer pairs. Our approach seeks to provide precise, scalable, and automated identification of evasive or unclear responses in political discourse by capturing the contextual relationship between questions and answers.

2 Related Work

Our approach to the CLARITY task combines transformer-based language models with targeted training strategies for three-class response-clarity classification, building upon research in pre-trained architectures, class-imbalanced text classification, and question answering.

Recent classification stems largely from leveraging pre-trained transformers. BERT introduced bidirectional context encoding, establishing strong baselines for question answering and classifica-

*The first two authors have equal contributions.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1094–1099

tion (Devlin et al., 2019). DeBERTa improved upon this by disentangling content and position information (He et al., 2021), while DeBERTa-v3 further optimized pre-training with ELECTRA-style training and gradient-disentangled embedding sharing for better downstream performance (He et al., 2023). These advances directly inform our model initialization phase, where we adopt DeBERTa-v3 as our backbone.

Class imbalance presents a key challenge in real-world classification tasks, including response clarity prediction. Prior work in lexical complexity prediction at SemEval-2021 Task 1 addressed skewed distributions through ensemble methods (Shardlow et al., 2021), feature engineering (Mosquera, 2021), multi-task learning (Taya et al., 2021), and combinations of deep learning with hand-crafted features (Zaharia et al., 2021). Some systems combined multiple transformer models to improve prediction robustness (Aziz et al., 2021), while others employed assembly models with novel phonological measures (Islam et al., 2021). These strategies inform our class balancing approach, which combines oversampling before data splitting with weighted loss during training.

The CLARITY task extends prior work on question answering by focusing specifically on response clarity. This connects to answerability detection in datasets like SQuAD 2.0, which introduced unanswerable questions to reading comprehension (Rajpurkar et al., 2018). More directly, the "I Never Said That" dataset formalized response clarity classification in QA pairs, providing both a taxonomy of unclear responses and baseline transformer adaptations (Thomas et al., 2024). Our data preprocessing—combining question and answer texts before encoding—follows approaches validated in this prior work. Recent shared tasks have explored related challenges in structured contexts; SemEval-2025 Task 8 addressed QA over tabular data using LLM-driven code generation (Site et al., 2025), chain-of-thought prompting (Mokhtar et al., 2025), and SQL-based pipelines (Giobergia, 2025; Tyagi et al., 2025; Antropova et al., 2025; Gao et al., 2025; Osés Grijalba et al., 2025). While focused on structured data, these approaches demonstrate the broader trend of combining pre-trained models with task-specific pipelines.

Collectively, these research streams inform our workflow: we adopt a DeBERTa-v3 backbone validated in prior classification tasks, incorporate oversampling and weighted loss to address class imbalance, and frame response clarity as a three-class clarity classification problem building on the "*I Never Said That*" taxonomy.

3 Proposed Framework

In this section, we describe our political question clarity classification framework as shown in Figure 1. Our goal is to predict the clarity label of an answer given the corresponding question.

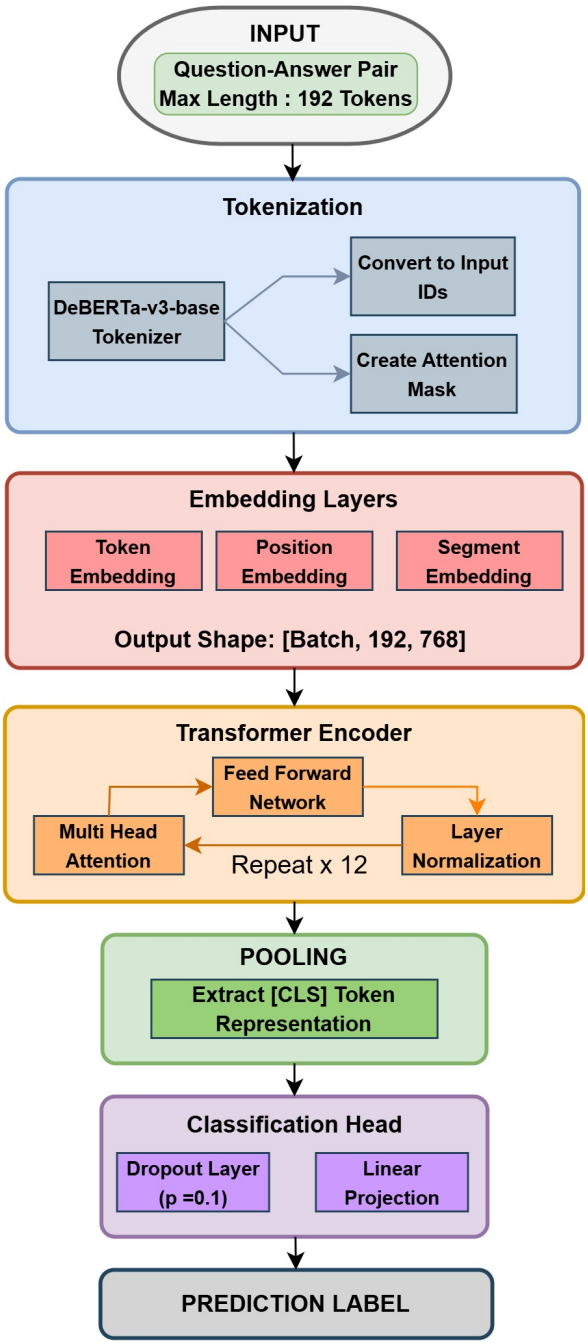

Figure 1: Overview of our proposed framework

In our framework, we use a sentence-pair classification approach with transformer models

to perform clarity classification, in which input question-answer pairs are concatenated into a single sequence. After performing question-answer pair classification through the DeBERTa-v3-base model, we estimate the final clarity label. Our pairwise learning strategy jointly encodes the question and the answer to detect directness versus ambiguity, essential in political discourse, where evasive responses are common.

3.1 Fine-tuned Transformer Model

We fine-tune the transformer model to perform sentence-pair classification for Question Answer Clarity Classification (QACC) using DeBERTa-v3-base. We describe the details approach in subsequent sections.

3.1.1 Input Representation

We practice with question-answer pairs to gain a deeper understanding of their contextual relationship, which helps us determine how clear the answer is to the question. We use the Huggingface transformers library (Wolf et al., 2020) for paired training, in which the input question and response are disconnected with the `[SEP]` token and form a single sequence. We make use of the DeBERTa-v3-base pre-trained transformer model (He et al., 2023). The model's initial token for QACC task training is the unique `[CLS]` token, which appears at the start of each sequence and is also responsible for the final-layer classification logits. For each sequence, we separate every pair with `[SEP]` token (as presented in Figure 1) where the question belongs to `text_a` and the answer belongs to `text_b`. We fine-tune the architecture using the pre-trained DeBERTa-v3-base model to assess clarity.

In order to maximize computational efficiency, we additionally shorten the input sequences, restricting questions to 80 characters and replies to 120 characters. Preliminary tests revealed that lengthier sequences only slightly improved performance at a greater computational cost.

3.1.2 DeBERTa-v3-base

By utilizing an improved mask decoder and disentangled attention mechanisms, DeBERTa (Decoding-enhanced BERT with disentangled attention) (He et al., 2021) outperforms the BERT and RoBERTa models. The third iteration of the DeBERTa model, known as DeBERTa-v3-base, adds even more enhancements to pre-training effectiveness and downstream task performance. On a variety of NLP tasks, including text classification, sentence-pair regression/classification, and natural language understanding, it achieves state-of-the-art results. We utilize the DeBERTa tokenizer and model to classify question-answer pairs in which the query and response form a single sequence.

In case of CUDA unavailability or memory constraints, we implement a fallback mechanism to BERT-base-uncased (Devlin et al., 2019) to ensure system robustness across different hardware configurations. DeBERTa-v3-base was successfully loaded and used throughout our testing; no reported run resulted in the BERT fallback.

3.2 Class Balancing Techniques

We use two complementary methods - oversampling and the class-weighted loss function to rectify the class imbalance in the training data.

3.2.1 Oversampling

Class imbalance in the initial training data may have skewed the model in favor of the majority classes. To counteract this, we oversample minority groups to achieve a balanced distribution of 1500 samples per class. While the majority of classes are kept in their original format, classes with fewer than 1500 samples are replicated until the desired number is obtained.

3.2.2 Class-Weighted Loss

By adding class weights to the loss function during training, we further mitigate class imbalance. The following formula determines the class weights:

$$w_c = \frac{N}{k \times n_c} \tag{1}$$

where N is the total number of samples, k is the number of classes (3), and n_c is the number of samples in class c. The weighted cross-entropy loss is then computed as:

$$\mathcal{L} = -\sum_{c=1}^{k} w_c \cdot y_c \cdot \log(\hat{y}_c) \tag{2}$$

where y_c is the ground truth and $\hat{y}_c$ is the predicted probability for class c.

4 EXPERIMENTS AND EVALUATION

4.1 Dataset Description

The organizers of SemEval-2026 Task 6: CLARITY introduced a benchmark dataset on response

clarity classification, focusing on political interviews (Thomas et al., 2026). The QEvasion dataset is constructed from U.S. presidential news conferences, spanning four presidents: George W. Bush, Barack Obama, Donald J. Trump, and Joseph R. Biden. Each instance consists of an interview question and the corresponding answer, annotated for clarity using a two-level hierarchical taxonomy.

The CLARITY task is divided into two subtasks: Task 1 focuses on three-class classification of response clarity, while Task 2 focuses on fine-grained nine-class evasion classification. Our participation is limited to Task 1. The top-level taxonomy for Task 1 includes three labels: *Clear Reply*, where the answer directly addresses the question; *Ambivalent Reply*, where the answer partially addresses the question but remains ambiguous; and *Clear Non-Reply*, where the answer avoids addressing the question entirely.

The complete dataset contains approximately 3700 annotated instances, with 3448 in the training set and approximately 300 in the test set. The dataset is stored in Parquet format and accessible via the Hugging Face datasets library. Table 1 presents the distribution of instances across the dataset splits for Task 1.

4.2 Experimental Settings

We now describe the set of parameters we used to design our proposed response-clarity classification model for Task 1. In our system, we utilize the DeBERTa-v3-base model (He et al., 2023) from the Huggingface Transformers library (Wolf et al., 2020) with fine-tuning. DeBERTa-v3 improves upon previous transformer models through disentangled attention mechanisms and ELECTRA-style pre-training with gradient-disentangled embedding sharing. We implement our system using PyTorch and the Huggingface Transformers library.

Data Preprocessing: For each instance, we combine the interview question and answer into a single input sequence using the format: "Question: {question} Answer: {interview_answer}". To manage computational constraints, we truncate questions to 80 characters and answers to 120 characters, resulting in a maximum sequence length of 192 tokens. This character-level truncation is implemented as a preprocessing step before tokenization; the tokenizer then encodes the resultant strings with a maximum sequence length of 192 tokens. Preliminary tests showed little effect on classification performance for this dataset, even though this method may occasionally truncate responses mid-sentence.We map the three clarity labels to integer values: Ambivalent (0), Clear Non-Reply (1), and Clear Reply (2).

Class Balancing: In order to mitigate class imbalance, we oversample the minority classes prior to splitting, aiming for 1,500 samples per class through repetition, which yields roughly 4,500 balanced samples. This balanced dataset was then stratified into training (80%), validation (10%), and test (10%) sets. We recognize that using oversampling before splitting may result in duplicate samples showing up across splits, which could cause naively inflated validation estimates. We therefore conduct the final evaluation using a different held-out test set, created from the original non-oversampled data, to obtain a more accurate measure of real-world performance (see Final Evaluation below).

Model Configuration: We fine-tune the model using the following hyperparameters: number of epochs = 10, learning rate = 2e-5, batch size = 8 (CPU) or 4 (GPU), maximum sequence length = 192 tokens, and weight decay = 0.01. We employ the AdamW optimizer with a linear learning rate scheduler, including warmup for the first 10% of training steps. To further address class imbalance, we compute class weights based on the original training distribution and use them with the class-weighted loss formula.

Training Procedure: We train our system on the balanced training data for 10 epochs, evaluating on the validation set after each epoch. We monitor the macro-averaged F1-score on the validation set to select the best model checkpoint, saving the model state that achieves the highest validation F1.

Final Evaluation: After selecting the best model based on validation performance, we evaluate on a test set that preserves the original class distribution. We create this test set by splitting the original (non-oversampled) data using an 85/15 train-test split with stratification to maintain the original class distribution. This ensures that our final results reflect real-world performance on naturally occurring data.

4.3 Evaluation Measures

To evaluate the performance of participants' response clarity classification systems for Task 1, the CLARITY task organizers defined macro-averaged F1-score as the primary evaluation measure. This metric is particularly appropriate for the task, given

the dataset's inherent class imbalance, as it treats all three clarity categories equally, regardless of their frequency. Following the task guidelines, systems are ranked based on macro F1-score.

4.4 Results and Analysis

Our DeBERTa-v3-based system achieved a macro F1-score of 0.67 on the test set, ranking 31st among participating systems for Task 1. Table 1 presents the performance of the top-4 ranked participating systems, along with our result.

Here, we see that our proposed method obtained moderate performance compared to the top-performing systems. The gap of 0.22 macro F1 between our system and the top-ranked system (TeleAI) indicates several areas for improvement.

System	Macro F1	Rank
csecudsg	0.67	31
Top performing systems based on macro F1		
TeleAI	0.89	1
AsymVerify	0.85	2
CSE-UOI	0.85	3
Rasende Rakete	0.83	4

Table 1: Comparative results with top-4 performing participants for Task 1.

5 CONCLUSION AND FUTURE DIRECTIONS

In this work, we described our method for the question-answer clarity classification challenge. To address the issue, we used the DeBERTa-v3-base transformer model within a unified architecture to classify question-answer pairs. We used paired learning to estimate the clarity label by leveraging the contextual relationships between question-answer pairs. To address imbalanced training data, we implemented class balancing via oversampling and class-weighted loss. Our proposed classification framework outperformed the baseline and achieved competitive scores.

In the future, we intend to use cutting-edge neural techniques in conjunction with a variety of manually constructed linguistic variables to extract relationships between question-answer pairs and estimate clarity. We also intend to investigate ensemble approaches that combine several transformer models. Complementary signals for the classification of clarity may also be obtained through multi-task learning with related tasks such as attitude analysis and response relevance. Applying our framework to other relevant domains of interest, as well as in low-resource language settings (e.g., Bangla), can be another avenue.

References

Ekaterina Antropova, Egor Kratkov, Roman Derunets, and 1 others. 2025. Tabaqa at semeval-2025 task 8: Column augmented generation for qa over tabular data. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*.

Abdul Aziz, MD. Akram Hossain, and Abu Nowshed Chy. 2021. CSECU-DSG at SemEval-2021 task 1: Fusion of transformer models for lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 627–631, Online. Association for Computational Linguistics.

Peter Bull. 2003. *Political communication and the media: How politicians communicate with the public*. Sage Publications, London, UK.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yuze Gao, Bin Chen, and Jian Su. 2025. I2r-nlp at semeval-2025 task 8: Question answering on tabular data. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*.

Flavio Giobergia. 2025. Minds at semeval-2025 task 8: Question answering over tabular data via llm-generated sql queries. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. *Preprint*, arXiv:2006.03654.

Aadil Islam, Weicheng Ma, and Soroush Vosoughi. 2021. BigGreen at SemEval-2021 task 1: Lexical complexity prediction with assembly models. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 667–677, Online. Association for Computational Linguistics.

Omar Mokhtar, Minah Ghanem, and Nagwa El-Makky. 2025. Alexnlp-mo at semeval-2025 task 8: A chain of thought framework for question-answering over tabular data. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*.

Alejandro Mosquera. 2021. Alejandro mosquera at SemEval-2021 task 1: Exploring sentence and word features for lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 554–559, Online. Association for Computational Linguistics.

Jorge Osés Grijalba, L. Alfonso Ureñ-López, Eugenio Martínez Cámara, and Jose Camacho-Collados. 2025. Semeval-2025 task 8: Question answering over tabular data. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*.

Pranav Rajpurkar, Jian Jia, and Percy Liang. 2018. Squad 2.0: Reading comprehension with unanswerable questions. In *NAACL*.

Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. SemEval-2021 task 1: Lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16, Online. Association for Computational Linguistics.

Atakan Site, Emre Hakan Erdemir, and Gülşen Eryiğit. 2025. Itunlp at semeval-2025 task 8: A zero-shot approach using llm-driven code generation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*.

Yuki Taya, Lis Kanashiro Pereira, Fei Cheng, and Ichiro Kobayashi. 2021. OCHADAI-KYOTO at SemEval-2021 task 1: Enhancing model generalization and robustness for lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 17–23, Online. Association for Computational Linguistics.

Konstantinos Thomas, Giorgos Filandrianos, Maria Lymperaiou, Chrysoula Zerva, and Giorgos Stamou. 2024. "I never said that": A dataset, taxonomy and baselines on response clarity classification. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5204–5233, Miami, Florida, USA. Association for Computational Linguistics.

Konstantinos Thomas, Giorgos Filandrianos, Maria Lymperaiou, Chrysoula Zerva, and Giorgos Stamou. 2026. Semeval-2026 task 6: Clarity – unmasking political question evasions. *Preprint*, arXiv:2603.14027.

Rishit Tyagi, Mohit Gupta, and Rahul Bouri. 2025. Aestar at semeval-2025 task 8: Agentic llms for question answering over tabular data. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

George-Eduard Zaharia, Dumitru-Clementin Cercel, and Mihai Dascalu. 2021. UPB at SemEval-2021 task 1: Combining deep learning and hand-crafted features for lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 609–616, Online. Association for Computational Linguistics.