ThinkVision at SemEval-2026 Task 6: A Transformer-Based Ensemble System for Clarity Detection

Purohit Gourav Ghanshyam, Praveen Swami, Shriyans S Sahoo, Jenish Bhati, Supriya Nadiger, Sunil Saumya
Indian Institute of Information Technology Dharwad, Karnataka, India
24bcs112@iiitdwd.ac.in, 24bcs108@iiitdwd.ac.in
24bcs142@iiitdwd.ac.in, 24bds027@iiitdwd.ac.in
supriya.nadiger@iiitdwd.ac.in, sunil.saumya@iiitdwd.ac.in

Abstract

We study the problem of assessing the clarity of political question–answer pairs, where the goal is to determine whether a response directly addresses the question, avoids it, or remains ambiguous. This task is particularly challenging in political discourse, where evasiveness can be subtle and context-dependent. To tackle this problem, we propose an ensemble-based approach built on the transformer-based model *DeBERTa-v3-base*, fine-tuned on concatenated question–answer inputs. Special attention is given to class imbalance during training to ensure robust performance across all categories. To better capture uncertainty in borderline cases, we train multiple models with different random seeds and employ Monte Carlo Dropout at inference time. Final predictions are obtained by averaging logits across ensemble models and stochastic forward passes, yielding a more stable decision boundary. Our system achieves a Macro-F1 score of 0.76 on the evaluation dataset. Error analysis reveals that responses that partially engage with the topic while failing to provide a direct answer remain the most challenging, highlighting the inherent difficulty of detecting nuanced evasiveness in political communication.

1 Introduction

Political discourse frequently involves strategic ambiguity, where public figures respond to questions in ways that appear relevant while avoiding a direct answer. Detecting such evasiveness is important for media analysis, accountability studies, and computational social science. However, identifying whether a response genuinely addresses a question or subtly avoids it is a challenging task, requiring modeling semantic alignment, discourse structure, and pragmatic intent.

In this work, we study clarity classification in political question–answer pairs, where the goal is to determine whether a response directly answers the question, explicitly avoids it, or remains ambiguous. The difficulty lies in capturing subtle forms of partial answering and implicit evasiveness, which often depend on nuanced semantic and contextual cues rather than explicit signals.

To address this problem, we propose a transformer-based ensemble framework built upon pre-trained language models. Our approach incorporates strategies for handling class imbalance and leverages stochastic inference through Monte Carlo Dropout to better manage uncertainty in ambiguous cases. Predictions from multiple model instances are aggregated to improve robustness and stability.

Our system achieves a Macro-F1 score of 0.76 on the official SemEval-2026 Task 6 evaluation set, ranking 17th among all participating teams.

Contributions. The main contributions of this work are:

- A transformer-based ensemble framework for clarity detection in political question–answer pairs.
- An uncertainty-aware inference strategy using Monte Carlo Dropout to improve prediction stability on ambiguous instances.
- An empirical analysis highlighting the challenges of modeling partially evasive political responses.

2 Problem Statement

This work addresses the problem of clarity classification in political question–answer (QA) pairs, as introduced in SemEval-2026 Task 6 (CLARITY)(Thomas et al., 2026), Subtask 1. The task builds upon prior research on evasive communication and equivocation in political discourse (Bull, 1994; Thomas et al., 2024). Given a political interview question and its corresponding response, the

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1973–1978
July 3-4, 2026 ©2026 Association for Computational Linguistics

objective is to determine how directly the answer addresses the question.

Formally, given a question q and an answer a, the task is to predict one of three labels:

- **Clear Reply** - the answer directly and substantively addresses the question.
- **Clear Non-Reply** - the answer explicitly avoids or refuses to address the question.
- **Ambivalent** - the answer partially addresses the question or contains indirect, evasive, or incomplete content.

Among these categories, the *Ambivalent* class presents the greatest challenge. Such responses often engage with the topic of the question while omitting a direct or explicit answer. For example:

Question: Do you support imposing sanctions on country X?

Answer: We are closely monitoring the situation and working with our allies to ensure stability.

Although the response is topically related, it does not clearly state a position, making classification non-trivial. Unlike standard text classification tasks that operate on a single input sequence, clarity detection requires modeling semantic and pragmatic interactions between two texts. The task is therefore closely related to discourse understanding and natural language inference, as it demands capturing implicit relations and subtle pragmatic cues.

We participate only in Subtask 1 and restrict ourselves to the data provided by the organizers, without using external resources.

2.1 Dataset Description

We use the official dataset released for SemEval-2026 Task 6 (CLARITY), Subtask 1. The dataset consists of political interview QA pairs collected across multiple decades, speakers, and contexts, annotated with the three clarity labels described above.

Table 1 presents the class distribution across the training, test, and evaluation splits.

As shown in Table 1, the dataset exhibits substantial class imbalance, with the *Ambivalent* category forming the majority class across all splits. This imbalance increases the difficulty of learning robust decision boundaries, particularly for the *Clear Non-Reply* class, which is comparatively underrepresented.

The preprocessing requirements are minimal and include basic whitespace normalization and sequence truncation to fit model input constraints. The training split corresponds to the QEvasion dataset training portion (Thomas et al., 2024), while the evaluation split matches the official CLARITY evaluation file provided by the organizers.

3 Related Work

Evasive and equivocal responses have long been studied in political communication research. Early work in discourse and speech act theory examined how public figures strategically avoid answering questions while maintaining topical relevance (Bull, 1994). More recent efforts have formalized clarity and equivocation in political interviews, leading to annotated datasets and shared evaluation frameworks (Thomas et al., 2024). These studies highlight the complexity of identifying partial or indirect responses, particularly when evasiveness is conveyed through pragmatic rather than explicit linguistic cues. Prior work on question answering has explored detecting unanswerable or non-responsive answers, such as SQuAD 2.0 (Rajpurkar et al., 2018), which introduced answerability prediction.

In natural language processing, related problems have been approached through natural language inference (NLI), stance detection, discourse modeling, and response intent modeling(Ferracane et al., 2021). These tasks similarly require capturing relationships between paired texts and modeling implicit semantic alignment. Clarity detection differs from standard text classification in that it demands reasoning over question–answer interactions, often requiring the system to detect subtle shifts in focus or incomplete responses.

Transformer-based architectures have achieved strong performance across a wide range of paired-text tasks, including question answering, dialogue modeling, and NLI. *DeBERTa-v3-base* (He et al., 2021) uses disentangled attention mechanisms and ELECTRA-style pre-training for improved contextual representations. Additionally, ensemble learning and uncertainty-aware inference methods, such as Monte Carlo Dropout (Gal and Ghahramani, 2016), have been shown to enhance robustness by reducing predictive variance and providing better calibrated outputs.

Our work builds upon these advances by com-

Split	Total	Clear Reply	Clear Non-Reply	Ambivalent
Training	3,450	1,052 (30.5%)	356 (10.3%)	2,042 (59.2%)
Test	308	79 (25.6%)	23 (7.5%)	206 (66.9%)
Evaluation	237	85 (35.9%)	35 (14.8%)	117 (49.3%)

Table 1: Class distribution of the CLARITY dataset for SemEval-2026 Task 6 Subtask 1

bining transformer-based modeling with ensemble diversity and stochastic inference to address clarity detection in political question–answer pairs.

4 System Overview

Our system is a transformer-based ensemble designed to classify the clarity of political responses in question–answer (QA) pairs. The overall pipeline consists of input construction, contextual encoding with a pre-trained transformer, classification, ensemble training, uncertainty-aware inference, and prediction aggregation (Figure 1).

4.1 Input Representation

Each instance consists of a question and its corresponding answer. We concatenate the two texts using a structured template:

```
Question: <question text>
+ Answer: <answer text>
```

This explicit formatting provides structural cues that help the model distinguish between the question and response segments. Inputs are tokenized using the DeBERTa-v3 tokenizer with a maximum sequence length of 320 tokens. Longer sequences are truncated, and shorter sequences are padded to enable efficient batching.

4.2 Base Encoder

We adopt *DeBERTa-v3-base* as the backbone encoder due to its strong performance on paired-text understanding tasks and its disentangled attention mechanism, which enables more effective modeling of relationships between question and answer tokens. The model produces contextualized token representations for the concatenated QA sequence.

Rather than using the representation of the `[CLS]` token, we apply attention-mask-weighted mean pooling over the final hidden states. This choice is motivated by the need to capture distributed evidence across longer responses, where relevant information may not be localized to a single token. In preliminary experiments, mean pooling provided more stable performance than CLS-based representations, particularly for longer and more complex answers.

$$h = \frac{\sum_i m_i x_i}{\sum_i m_i}$$

where x_i denotes the embedding of token i and m_i is the corresponding attention mask value. This pooling strategy aggregates information across the full sequence and is particularly effective for longer answers containing distributed evidence.

4.3 Classification Layer

The pooled representation is passed through a two-layer feedforward network consisting of a linear projection, Layer Normalization, GELU activation, and dropout ($p = 0.2$). A final linear layer produces logits over the three clarity classes.

4.4 Handling Class Imbalance

The dataset exhibits substantial class imbalance, with the *Ambivalent* class being dominant. To mitigate bias toward majority classes, we employ weighted cross-entropy loss. Class weights are defined as:

$$w_c = \frac{N}{C \cdot n_c}$$

where N is the total number of training instances, C is the number of classes, and n_c is the number of samples in class c.

4.5 Ensemble Training

To improve robustness, we train multiple instances of the same architecture with different random initialization seeds $\{42, 123, 456, 789, 2024\}$. Model diversity arises from stochastic optimization dynamics, leading to complementary decision boundaries. Each model is trained independently, and predictions are combined at inference time.

4.6 Uncertainty-Aware Inference

To better handle ambiguous cases, we employ Monte Carlo Dropout during inference. For each model, we perform two forward passes: (1) standard deterministic inference with dropout disabled,

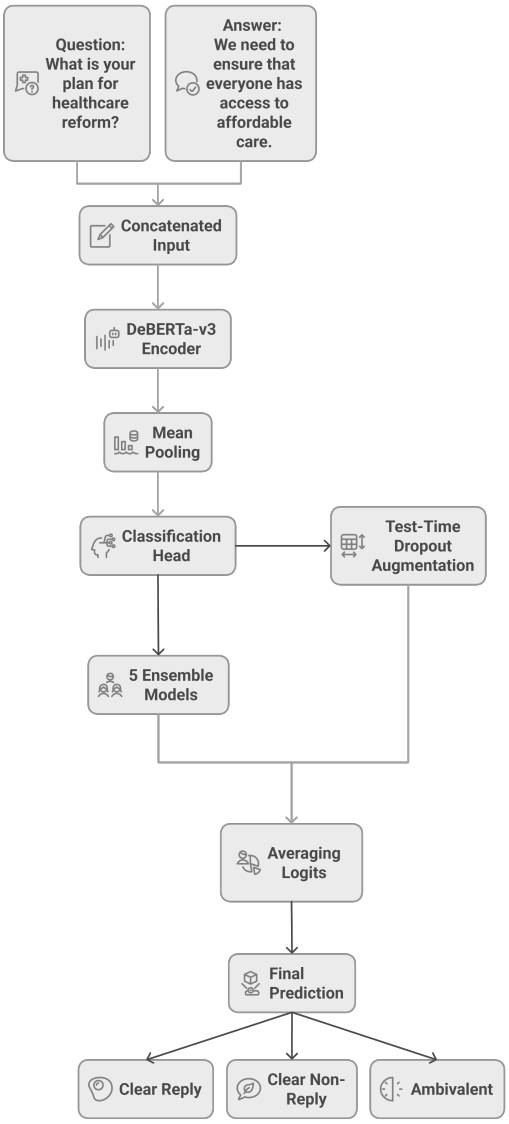

Figure 1: Overview of the ThinkVision system architecture

and (2) stochastic inference with dropout enabled. This introduces controlled variability that helps capture predictive uncertainty in borderline examples.

4.7 Prediction Aggregation

Final predictions are obtained by averaging logits across all ensemble models and inference passes:

$$\hat{y} = \arg\max_{c} \left(\frac{1}{K} \sum_{k=1}^{K} z_k \right)$$

where z_k denotes the logits from the k-th model or stochastic pass, and K is the total number of predictions aggregated. The predicted label corresponds to the class with the highest averaged logit.

5 Experimental Setup

All models are trained on the official training split provided by the SemEval-2026 Task 6 organizers. Inputs are tokenized using the DeBERTa-v3 tokenizer with a maximum sequence length of 320 tokens.

We fine-tune the models for four epochs using the AdamW optimizer with a learning rate of 4×10^{-5}, weight decay of 0.01, and a warmup ratio of 0.1. The batch size is set to 20. No external datasets or additional task-specific pre-trained models are used.

To construct the ensemble, five models are trained independently with different random seeds $\{42, 123, 456, 789, 2024\}$.

5.1 Evaluation Protocol

System performance is evaluated using Macro-F1 score, which is the official evaluation metric of the shared task. Macro-F1 equally weights all classes and is therefore appropriate for assessing performance under class imbalance.

5.2 Implementation Details

Our implementation is based on the Hugging Face `transformers` library (Wolf et al., 2020) with PyTorch as the backend framework. Experiments are conducted on NVIDIA Tesla T4 GPUs on the Kaggle platform. Model checkpoints are saved using the `safetensors` format.

During inference, batch processing is employed to improve computational efficiency. All experiments are implemented in Python using publicly available libraries, ensuring reproducibility of our system.

6 Results

Our final submitted system achieves a Macro-F1 score of 0.76 on the official evaluation set, ranking 17th among participating teams. The results demonstrate consistent improvements as additional modeling components are incorporated.

Table 2 presents the performance of progressively enhanced system configurations. A single DeBERTa-v3-base model provides a strong baseline. Introducing imbalance-aware training improves performance further. The largest gains are observed when incorporating ensemble learning, with performance improving from 0.61 to 0.739. This demonstrates that model diversity significantly enhances robustness. Further improvements are

achieved by incorporating Monte Carlo Dropout, increasing performance to 0.76. This indicates that uncertainty-aware inference provides additional benefits, particularly for handling ambiguous instances.

Table 3 presents the per-class performance of the final system. The model achieves the highest performance on the *Clear Non-Reply* class (F1 = 0.82), which is characterized by explicit evasive cues. In contrast, the *Ambivalent* class remains the most challenging (F1 = 0.73), reflecting the difficulty of identifying partially addressed responses. The *Clear Reply* class achieves moderate performance (F1 = 0.72), with errors often arising from semantically relevant but incomplete answers.

Analysis of the confusion matrix shows that most errors occur between the *Ambivalent* and *Clear Reply* classes. A substantial number of *Ambivalent* instances are misclassified as *Clear Reply*, indicating that the model tends to rely on semantic relevance rather than fully capturing whether the response directly answers the question. In contrast, the *Clear Non-Reply* class is more distinctly identified, with minimal confusion with other classes. These findings highlight that effective clarity detection requires modeling deeper discourse intent beyond surface-level semantic similarity.

6.1 Ablation Study

To assess the contribution of individual components, we analyze the performance of progressively enhanced system configurations shown in Table 2. A single DeBERTa-v3-base model achieves a Macro-F1 score of 0.55, which improves to 0.61 with imbalance-aware training, highlighting the importance of addressing class imbalance.

Incorporating ensemble learning results in a substantial performance gain, increasing Macro-F1 to 0.739. This confirms that diversity from multiple model initializations leads to more robust predictions.

Finally, incorporating Monte Carlo Dropout further improves performance to 0.76. While the improvement is smaller compared to ensembling, it consistently enhances performance on ambiguous instances by reducing overconfident predictions near decision boundaries.

7 Error Analysis

To better understand system limitations, we conduct a qualitative analysis of misclassified instances. Several recurring error patterns highlight the difficulty of modeling subtle evasiveness in political discourse.

Ambivalent vs. Clear Reply Confusion. The most frequent errors involve predicting *Clear Reply* for instances labeled as *Ambivalent*. Many such responses contain partial answers mixed with general or tangential information. Although topically aligned with the question, they often omit a direct or complete response. The model tends to rely on overall semantic relevance rather than evaluating whether the central intent of the question has been fully addressed.

Implicit Evasion and Multi-Part Questions. Implicit evasive strategies further contribute to misclassification. Responses that shift focus, reframe the issue, or use procedural language (e.g., "we are considering this issue carefully") rarely include explicit refusal cues, making them difficult to distinguish from genuine replies. Similarly, multi-part questions pose challenges: when an answer addresses only one component, the system often predicts *Clear Reply* despite incomplete coverage. For example:

Question: Will you approve the proposed reform bill?

Answer: We are reviewing all aspects of the proposal carefully and will make a decision at the appropriate time.

Although policy-relevant terminology is present, the response avoids a direct commitment. This suggests that semantic overlap alone is insufficient for accurate clarity detection.

Overall, the findings indicate that effective clarity classification requires deeper modeling of discourse structure and communicative intent beyond surface-level similarity.

8 Conclusion and Limitations

We presented a transformer-based ensemble framework for clarity detection in political question–answer pairs for SemEval-2026 Task 6, Subtask 1. Our approach combines pre-trained language models, class imbalance handling, multi-seed ensembling, and Monte Carlo Dropout for uncertainty-aware inference. The final system achieves a Macro-F1 score of 0.76 on the official evaluation set. Our code is publicly available at: https://github.com/gourav18dart/CLARITY-SEMEVAL2026-System.

Configuration	Macro-F1
DeBERTa-v3-base (single model)	0.55
Enhanced model (imbalance-aware training)	0.61
Ensemble (5 models)	0.739
Final ensemble (5 models + MC Dropout)	**0.76**

Table 2: Performance of progressively enhanced system configurations. Intermediate models are evaluated on the test set, while the final ensemble models correspond to the official evaluation submission

Class	Precision	Recall	F1
Clear Reply	0.77	0.68	0.72
Clear Non-Reply	0.81	0.83	0.82
Ambivalent	0.68	0.78	0.73

Table 3: Per-class performance of the final ensemble model on the evaluation set

Error analysis shows that implicitly evasive and partially answered responses remain challenging, particularly when topical relevance is present without explicit commitment. This highlights the need for deeper discourse-level modeling beyond semantic similarity.

Despite competitive performance, the approach has limitations. It relies on supervised fine-tuning on a task-specific dataset and does not explicitly model discourse structure or rhetorical strategies. Additionally, ensemble inference increases computational cost. Future work may explore discourse-aware architectures and more efficient uncertainty modeling techniques.

References

Peter Bull. 1994. On identifying questions, replies, and non-replies in political interviews. *Journal of Language and Social Psychology*, 13(2):115–131.

Elisa Ferracane, Greg Durrett, Junyi Jessy Li, and Katrin Erk. 2021. Did they answer? subjective acts and intents in conversational discourse. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1626–1644, Online. Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML)*.

Pengcheng He, Jianfeng Gao, Anoop Kumar, Mingjian Huang, Xiaodong He, Mingyang Zhou, Lei Ji, Yu Cao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. 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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clément Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 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, Online. Association for Computational Linguistics.