VerbaNexAI at SemEval-2026 Task 6: Automatic Detection of Political Evasion through Hierarchical Classification with RoBERTa Large

Jeison D. Jimenez, Deyson Gómez Sánchez, Jairo E. Serrano, Juan C. Martinez-Santos, Edwin Puertas
Universidad Tecnológica de Bolívar, Cartagena, Colombia
(jalvear,deygomez,jserrano,jcmartinezs,epuerta)@utb.edu.co

Abstract

This paper describes VerbaNex AI's participation in SemEval-2026 Task 6: CLARITY, a shared task on automatic detection of question evasion in political interview transcripts. The task requires classifying question-answer pairs into three clarity levels (Task 1) and identifying nine evasion techniques (Task 2). We propose and evaluate two independent systems based on RoBERTa-Large. The first is a standard sequence classifier that treats each question-answer pair as a single input sequence, leveraging RoBERTa's native two-segment encoding to model the relationship between the two texts jointly. The second is a dual-encoder architecture that processes the question and answer independently and computes geometric interaction features to model the semantic misalignment between them explicitly. Both systems are trained on Task 2 labels and derive Task 1 predictions via the hierarchical mapping proposed by the task organizers. Our best result was achieved by the standard sequence classifier, reaching Rank 10 on Task 2 and Rank 25 on Task 1.

1 Introduction

When politicians are asked direct questions in interviews or press conferences, they rarely give straight answers. (Bull, 2003) found that politicians provided clear responses to only 39–46% of interview questions, compared to 70–89% for non-politicians. Despite how widespread this behavior is, automatically detecting and classifying it remains an open challenge in Natural Language Processing, partly due to the complexity of capturing pragmatic intent and the need to reason over long contexts (Thomas et al., 2024).

SemEval-2026 Task 6 addresses this problem through the CLARITY (Thomas et al., 2026), which asks systems to analyze question-answer pairs from US presidential interviews and determine the evasiveness of each response. The task operates at two levels: deciding whether a response is a Clear Reply, an Ambivalent Reply, or a Clear Non-Reply (Task 1), and identifying which of nine specific evasion techniques the respondent used (Task 2). Both tasks draw on the QEvasion dataset, and two-level taxonomy introduced by (Thomas et al., 2024), which builds on established typologies from political science and linguistics (Bull and Mayer, 1993; Rasiah, 2010).

Evasion Label	Train	Test
Explicit	1052	91
Dodging	706	54
Implicit	488	60
General	386	52
Deflection	381	21
Declining to answer	145	12
Claims ignorance	119	8
Clarification	92	4
Partial/half-answer	79	6

Table 1: Class distribution of the QEvasion dataset.

To tackle this task, we developed and compared two systems built on RoBERTa-Large, each differing in how they encode the question-answer relationship, as described in the following sections. Our best result came from a standard sequence classifier, demonstrating that careful fine-tuning with class-weighted loss and label smoothing provides a strong, competitive baseline. The code is publicly available at https://github.com/VerbaNexAI/SemEval2026/tree/main/Task_6-Clarity.

2 Background

Related work has approached the problem of political evasion from complementary perspectives. (Ferracane et al., 2021) crowdsourced annotations on political interview answers to classify responses as answer, shift, or non-answer, also capturing perceived speaker intent. Work on question answer-

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1429–1435

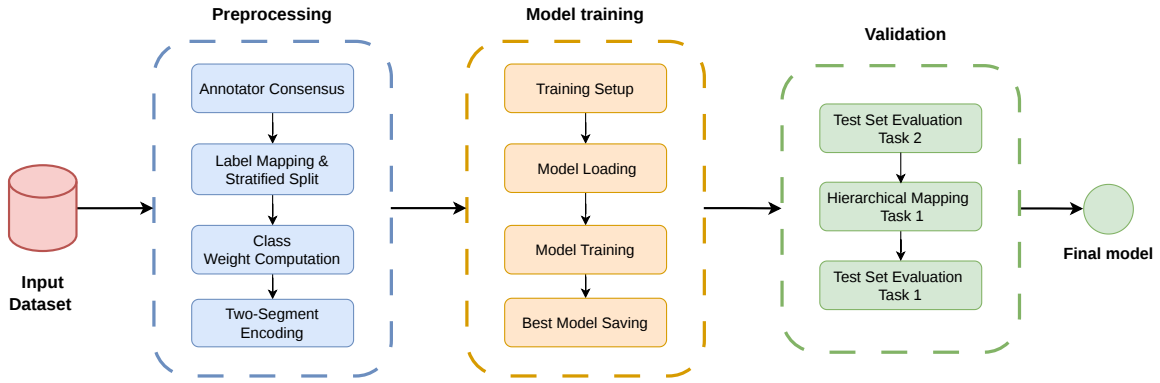

Figure 1: System Pipeline Model 1.

ability in open-domain QA (Rajpurkar et al., 2018) shares structural similarities with the CLARITY challenge but focuses on whether a passage contains an answer rather than whether a speaker chose to provide one. (Thomas et al., 2024) proposed framing response clarity as an NLP classification task, experimenting with encoder models such as RoBERTa, DeBERTa, and XLNet, as well as large language models including Llama-2 and Falcon via instruction-tuning with LoRA (Hu et al., 2021), achieving their best result with an instruction-tuned Llama-70B at a macro F1 of 0.682.

Despite these advances, fine-grained evasion classification remains an open challenge. Chain-of-thought prompting (Kojima et al., 2023) improved evasion-based classification but offered no consistent gains for direct clarity prediction (Thomas et al., 2024). More recently, (Prahallad et al., 2026) evaluated prompt-based strategies on the CLARITY dataset, finding that while structured prompting reliably improves high-level clarity prediction, fine-grained evasion categories remain unstable and difficult to classify even for state-of-the-art models. These limitations motivate exploring fine-tuning strategies and explicit interaction modeling as complementary approaches to the task.

3 Systems Overview

Both systems proposed in this work share RoBERTa-Large (Liu et al., 2019) as their backbone encoder, selected for its robust contextual representations and proven effectiveness on sentence-pair classification tasks (Martinez et al., 2023; Almanza et al., 2025). To address the pronounced class imbalance described in Section 2, both systems employ a class-weighted cross-entropy loss where the weight assigned to each class c is computed as:

$$w_c = \frac{n_{\text{samples}}}{n_{\text{classes}} \times n_{\text{samples}_c}} \quad (1)$$

where n_{samples} is the total number of training instances, n_{classes} is the number of classes, and n_{samples_c} is the number of instances belonging to class c. The resulting weights are clipped to the range $[0.5, 5.0]$ to avoid extreme penalization of the rarest classes (Sánchez et al., 2025a,b). Both systems also apply label smoothing ($\epsilon = 0.1$) and use early stopping with macro F1 as the reference metric. The two systems differ in how they encode the question-answer relationship, as described in the following subsections.

3.1 System 1: Standard Sequence Classifier

The first system fine-tunes RoBERTa-Large for sequence classification. As illustrated in Figure 1, the pipeline comprises three stages: preprocessing, model training, and validation.

In the preprocessing stage, annotator consensus is resolved on the test set via majority vote across three independent annotations, with the first annotator serving as a tiebreaker. The training set is split into 80% for training and 20% for validation using stratified sampling to preserve the class distribution in both subsets. Class weights are computed exclusively on the training subset to prevent data leakage from the validation set into the loss function. Each question-answer pair is then encoded using RoBERTa's native two-segment format: `<s> question </s></s> answer </s>`, which allows the model's self-attention mechanism to model the relationship between both texts across all layers jointly.

The model is configured with a single linear classification head that projects from the hidden size to

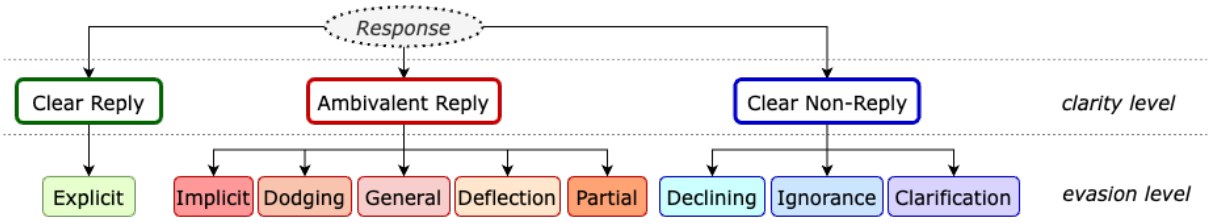

Figure 2: Two-level response clarity taxonomy proposed by (Thomas et al., 2024).

the nine evasion classes, with dropout applied both within the transformer layers and in the classification head. The best checkpoint is selected based on the validation macro F1 and saved for evaluation. The saved model is then evaluated on the held-out test set for Task 2, and Task 1 predictions are derived via the hierarchical mapping defined by the task organizers.

3.2 System 2: Dual Encoder with Interaction Features

The second system introduces a dual-encoder architecture that explicitly models the semantic misalignment between the question and the answer, the core signal for evasion detection. As illustrated in Figure 3, the preprocessing stage is identical to System 1, with the exception that each question and answer is encoded independently using a shared RoBERTa-Large encoder rather than as a single concatenated sequence.

For each input pair, the shared encoder processes the question and the answer in separate forward passes, extracting the hidden state of the <s> token at position 0 as the representation embedding for each text, yielding question embedding $\mathbf{Q}$ and answer embedding $\mathbf{A}$, each of 1024 dimensions. A set of geometric interaction features is then computed from both embeddings to capture different aspects of the question-answer relationship:

- $\mathbf{Q} - \mathbf{A}$: element-wise difference, capturing aspects of the question absent from the answer
- $\mathbf{Q} \odot \mathbf{A}$: element-wise product, capturing local semantic alignment
- $\mathbf{A}_\perp$: the component of $\mathbf{A}$ orthogonal to $\mathbf{Q}$, representing what the answer conveys that is unrelated to the question
- cos_sim$(\mathbf{Q}, \mathbf{A})$: cosine similarity, a global alignment scalar
- $\|\mathbf{Q} - \mathbf{A}\|_2$: L2 distance, a global divergence scalar
- alignment_ratio: fraction of $\mathbf{A}$'s norm aligned with $\mathbf{Q}$
- relative_magnitude: ratio $\|\mathbf{A}\|/\|\mathbf{Q}\|$, indicating proportional response length

The five vector features and four scalars are concatenated into a single representation of $5 \times 1024 + 4 = 5124$ dimensions, which is passed to a two-layer MLP classifier with BatchNorm, ReLU activation, and dropout, projecting to the nine evasion classes.

While both systems share the same backbone, training objective, and hyperparameter configuration, they differ fundamentally in how they model the question-answer relationship. Table 2 summarizes these architectural differences.

4 Experimental Setup

This section describes the dataset used for training and evaluation, followed by the hardware environment and hyperparameter configuration shared by both systems.

4.1 Data

The QEvasion dataset, introduced by (Thomas et al., 2024), consists of question-answer pairs extracted from US presidential interview transcripts spanning 2006 to 2023. The dataset contains 3,448 training instances and 308 test instances, annotated according to the two-level hierarchical taxonomy shown in Figure 2. At the first level, responses are classified into three clarity categories: Clear Reply, Ambivalent Reply, and Clear Non-Reply. At the second level, each category maps to a set of fine-grained evasion techniques, yielding nine classes in total. The annotation was carried out by three non-expert annotators and validated by a political science expert, with inter-annotator agreement of k=0.644 for clarity labels and k=0.48 for evasion labels. Table 1 shows the class distribution across both splits, revealing a pronounced imbalance: Explicit responses account for approximately

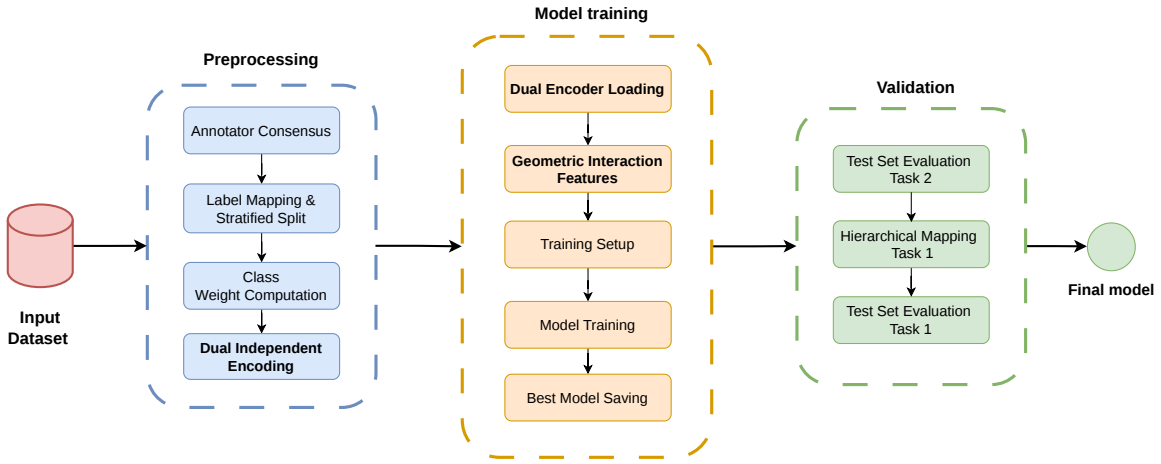

Figure 3: System Pipeline Model 2.

Component	System 1 (Seq. Classifier)	System 2 (Dual Encoder)
Encoding	Joint: <s> Q </s></s> A </s>	Independent: <s> Q </s> + <s> A </s>
Q-A interaction	Implicit cross-segment self-attention	Explicit geometric features
Feature dim	1,024	5,124
Classifier	Linear (1,024 → 9) + Dropout	MLP: BN → ReLU → Dropout → Linear
Max length	512 tokens (joint Q+A)	512 tokens per segment

Table 2: Architectural comparison of the two proposed systems.

30% of instances, while minority classes such as Partial/half-answer, Clarification, and Claims ignorance each represent fewer than 4%.

4.2 Model config

We trained both systems using the AdamW optimizer with a linear learning rate schedule. We chose a learning rate of 2e-5 following the recommendations of (Liu et al., 2019) for fine-tuning RoBERTa on downstream classification tasks. A warmup ratio of 0.1 was applied to stabilize training in the early epochs before the full learning rate takes effect. We use a weight decay of 0.01 to prevent overfitting on the relatively small training set. We applied gradient clipping with a maximum norm of 1.0 to avoid unstable updates during fine-tuning.

An effective batch size of 32 was achieved through gradient accumulation. A label smoothing of 0.1 was applied to reduce overconfident predictions for ambiguous evasion categories. We used early stopping with a patience of 7 epochs to select the best checkpoint based on validation macro F1, preventing overfitting while allowing sufficient training time for minority classes to be learned. Table 3 summarizes the full configuration.

We conducted all experiments on a single workstation equipped with an Intel Core i7-12700KF CPU, an NVIDIA RTX 4060 Ti GPU with 16 GB of VRAM, and 32 GB of system RAM. Training was performed with mixed-precision (fp16) in PyTorch to optimize memory usage during fine-tuning of RoBERTa-Large. Both systems were trained end-to-end on the same hardware to ensure comparability of training times and resource consumption.

Hyperparameter	Value
Effective batch size	32
Learning rate	2e-5
Weight decay	0.01
Max epochs	25
Dropout	0.1
Label smoothing	0.1
Class weight clip	[0.5, 5.0]
Early stopping patience	7

Table 3: Shared hyperparameter configuration for both systems.

5 Results

We evaluated both systems on the QEvasion test set across Task 1 and Task 2. All metrics reported are F1-score in its macro and weighted variants. System 1 consistently outperforms System 2 across both tasks, achieving a macro F1 of 0.3760 on

	Task 1		Task 2	
System	**Macro**	**Weight**	**Macro**	**Weight**
System 1	**0.6299**	**0.6747**	**0.3760**	**0.3580**
System 2	0.5118	0.5661	0.3022	0.2735

Table 4: F1 results on the QEvasion test set for Task 1 and Task 2.

Task 2 and 0.6299 on Task 1, compared to 0.3022 and 0.5118 for System 2, as shown in Table 4. These results suggest that jointly encoding the question-answer pair through RoBERTa's native two-segment attention is more effective for this task than computing explicit geometric interaction features on independently encoded representations.

A per-class analysis of System 1 on Task 2 reveals considerable variation across evasion categories, as shown in Table 5. The model performs best on Clarification (F1=0.889) and Explicit (F1=0.519), the two most linguistically distinctive categories (Thomas et al., 2024), achieving near-perfect classification (8/9 correct) and the highest absolute correct predictions (71/84) respectively (Figure 5). Performance degrades systematically for categories under the *Ambivalent Reply* node of the taxonomy (Thomas et al., 2024): General (F1=0.328) and Implicit (F1=0.255) share the property of engaging with the question indirectly or underspecifically, making them difficult to distinguish from each other and from Explicit, a difficulty corroborated by low inter-annotator agreement reported for these pairs ($\kappa = 0.58$ for General vs. Explicit) (Thomas et al., 2024). Similarly, Dodging (F1=0.270) and Deflection (F1=0.147) occupy adjacent positions in the *Clear Non-Reply* branch, differing only in whether the speaker acknowledges the question before shifting focus; this subtle distinction leads Deflection to receive zero correct predictions, with its instances largely absorbed by Dodging (Figure 5). Finally, Partial/half-answer (F1=0.000) suffers from low training support and high semantic proximity to Explicit, consistent with $\kappa = 0.68$ reported for this pair (Thomas et al., 2024).

Table 6 presents the per-class breakdown for System 2 on Task 2. Compared to System 1 (Table 5), System 2 shows a more balanced distribution across the majority classes—Explicit (F1=0.413), Implicit (F1=0.288), and Claims ignorance (F1=0.400)—but consistently underperforms System 1 in all categories. The most pronounced

Class	**P**	**R**	**F1**
Explicit	0.592	0.462	0.519
Implicit	0.353	0.200	0.255
Dodging	0.343	0.222	0.270
General	0.276	0.404	0.328
Deflection	0.106	0.238	0.147
Partial/half-answer	0.000	0.000	0.000
Declining to answer	0.308	0.667	0.421
Claims ignorance	0.500	0.625	0.556
Clarification	0.800	1.000	0.889

Table 5: Per-class results for System 1 on Task 2 (QEvasion test set).

Class	**P**	**R**	**F1**
Explicit	0.409	0.418	0.413
Implicit	0.314	0.267	0.288
Dodging	0.233	0.130	0.167
General	0.200	0.135	0.161
Deflection	0.047	0.095	0.062
Partial/half-answer	0.125	0.167	0.143
Declining to answer	0.200	0.500	0.286
Claims ignorance	0.333	0.500	0.400
Clarification	0.667	1.000	0.800

Table 6: Per-class results for System 2 on Task 2 (QEvasion test set).

gaps appear in Clarification (0.889 vs. 0.800), Claims ignorance (0.556 vs. 0.400), and Explicit (0.519 vs. 0.413). Notably, System 2 achieves a non-zero F1 on Partial/half-answer (F1=0.143), a category in which System 1 also struggles, suggesting that the geometric interaction features in System 2 capture some signal in this low-support category despite its lower overall macro F1. These results reinforce the advantage of joint encoding in System 1, while highlighting that explicit interaction modeling may offer complementary strengths for underrepresented evasion types.

Regarding the official competition evaluation, System 1 was assessed on a separate held-out dataset distinct from QEvasion. The system achieved a macro F1 of 0.72 on Task 1, ranking 25th, and a macro F1 of 0.52 on Task 2, ranking 10th among all participating teams, as reported in Table 7.

An error analysis based on the confusion matrices from the official evaluation set provides further insight into the model's behavior. As shown in Figure 4, the main source of error in Task 1 is the confusion between Ambivalent Reply and Clear

Task	Macro F1	Rank
Task 1 (3 classes)	0.72	25
Task 2 (9 classes)	0.52	10

Table 7: Official competition results for System 1.

Reply, with 35 Ambivalent instances misclassified as Clear Reply. This pattern suggests that the model tends to interpret responses that contain partial information as direct answers, failing to capture the subtle pragmatic cues that distinguish evasion from genuine engagement.

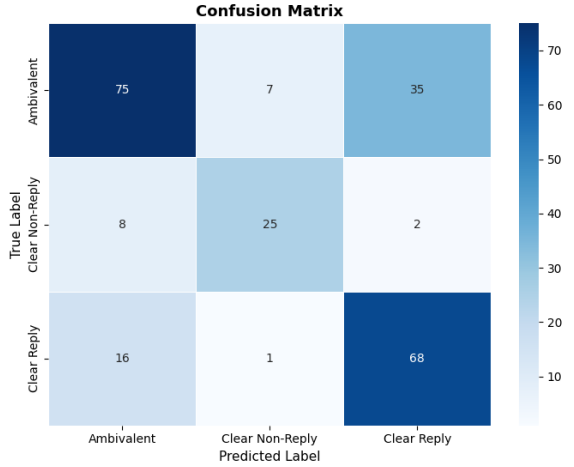

Figure 4: Confusion matrix for Task 1 on the official evaluation set.

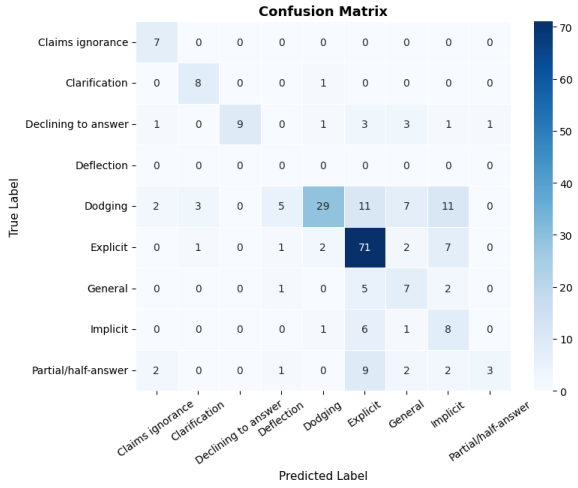

Figure 5: Confusion matrix for Task 2 on the official evaluation set.

For Task 2, the confusion matrix in Figure 5 reveals that misclassifications are concentrated within the Ambivalent Reply cluster, where Dodging, Implicit, General, and Deflection share overlapping linguistic surfaces. Deflection is the most problematic category, with zero correct predictions, as Dodging and Explicit largely absorb its instances. A strong attraction toward Explicit is also observed across multiple classes, including Partial/half-answer, where 9 out of 19 instances are misclassified as Explicit. In contrast, Claims ignorance and Clarification achieve near-perfect classification, reflecting their more distinctive linguistic markers.

6 Conclusion

Both systems demonstrated the viability of RoBERTa-Large as a backbone for question evasion detection in political interview transcripts within the SemEval-2026 Task 6: CLARITY framework. The first system, based on standard sequence classification with native two-segment encoding, consistently outperformed the second, which introduced a dual-encoder architecture with explicit geometric interaction features computed from independently encoded question and answer representations. The standard sequence classifier achieved macro-F1 scores of 0.72 on Task 1 and 0.52 on Task 2 in the official competition, ranking 25th on Task 1 and 10th on Task 2. Notably, on Task 1 (direct clarity), our system outperformed the fine-tuned Llama-70B baseline—the strongest among the provided reference models—despite being a considerably smaller architecture, demonstrating that appropriate domain-specific fine-tuning may be sufficient to dispense with larger-scale language models on this task.

The superiority of the standard sequence classifier over the dual encoder stems from the nature of RoBERTa's self-attention mechanism. When the question and answer are jointly encoded in the two-segment format, cross-segment attention spans all transformer layers, allowing the model to capture the pragmatic relationship between the two segments in a deeply integrated manner. In contrast, the dual encoder computes interaction features from a single summary token representation of each segment independently, which may constitute an insufficient approximation of the subtle pragmatic complexity inherent to evasive political discourse.

Error analysis reveals that the main difficulty for both tasks lies in the Ambivalent Reply group, where categories such as Dodging, Implicit, General, and Deflection share overlapping linguistic

surfaces that are difficult to distinguish even for fine-tuned transformer models. The Deflection category proved particularly intractable, with zero correct predictions in the official evaluation, underscoring the need for more targeted strategies for minority classes.

Future work will explore ensemble approaches combining both systems, as well as more specific data augmentation strategies focused on the most confusing evasion categories within the Ambivalent Reply group. Additionally, we propose investigating asymmetric attention mechanisms for the joint encoding of question-answer pairs, so that the model can assign differentiated weights to each segment during processing, reflecting the inherent pragmatic asymmetry of evasive political discourse, in which the interviewee's response contains the most determinative linguistic signals for evasion classification.

7 Acknowledgments

The authors would like to acknowledge the support provided by the master's degree scholarship program in engineering at the Universidad Tecnológica de Bolívar (UTB) in Cartagena, Colombia.

References

Danileth Almanza, Juan Martinez-Santos, and Edwin Puertas. 2025. VerbaNexAI at SemEval-2025 task 11 track a: A RoBERTa-based approach for the classification of emotions in text. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1192–1197.

Peter Bull. 2003. *The Microanalysis of Political Communication: Claptrap and Ambiguity*. Routledge, London.

Peter Bull and Kate Mayer. 1993. How not to answer questions in political interviews. *Political Psychology*, 14:651–666.

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.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. *Preprint*, arXiv:2205.11916.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *Preprint*, arXiv:1907.11692.

Elizabeth Martinez, Juan Cuadrado, Juan Carlos Martinez-Santos, Daniel Peña, and Edwin Puertas. 2023. Automated depression detection in text data: Leveraging lexical features, phonesthemes embedding, and RoBERTa transformer model. In *Proceedings of the Iberian Languages Evaluation Forum 2023 (IberLEF 2023)*.

Lavanya Prahallad, Sai Utkarsh Choudarypally, Pragna Prahallad, and Pranathi Prahallad. 2026. Prompt-based clarity evaluation and topic detection in political question answering. *Preprint*, arXiv:2601.08176.

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.

Parameswary Rasiah. 2010. A framework for the systematic analysis of evasion in parliamentary discourse. *Journal of Pragmatics*, 42:664–680.

Deyson Gómez Sánchez, J Jimenez, M Ramírez, and J Martinez. 2025a. RoBERT-IA: Human-AI collaborative text classification. *Working Notes of CLEF*.

Deyson Gómez Sánchez, Jeison D Jimenez, Elizabeth Ruíz Padilla, Jairo E Serrano, Juan C Martinez-Santos, and Edwin Puertas. 2025b. Twitbaiter: Model of clickbait detection in spanish. In *Proceedings of the Iberian Languages Evaluation Forum 2025 (IberLEF 2025)*.

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.