UMUTeam at SemEval-2026 Task 6: Soft-Voting Transformer Ensembles for Detecting and Classifying Response Ambiguity in Political Discourse

Tomás Bernal-Beltrán, Ronghao Pan, Jorge Gómez-Navalón, José Antonio García-Díaz, Rafael Valencia-García
Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100 Murcia, Spain
{ tomas.bernalb, ronghao.pan, jorge.gomezn, joseantonio.garcia8, valencia}@um.es

Abstract

Political discourse frequently involves strategically ambiguous responses, particularly in high-stakes settings such as presidential debates and interviews. Detecting whether a politician has directly answered a question, provided an ambiguous reply or issued a clear non-reply remains a challenging task due to the pragmatic and rhetorical nature of political language. This paper describes our participation in the SemEval 2026 CLARITY shared task on response ambiguity detection and classification in English. We focused exclusively on Task 1 (Clarity-level Classification) and proposed a weighted soft-voting ensemble that combines four fine-tuned encoder-only transformer models: RoBERTa-large, BERT-large-cased, DistilBERT-cased and ModernBERT-large. Each model was optimized through grid search and their predicted class probability distributions were aggregated using a weighted linear combination. On the official test set, our system achieved a macro-F1 score of 0.71, ranking 26th out of 41 participating teams. Even with the performance gap compared to top-ranked systems, our results demonstrate that a lightweight set of moderately sized encoder models can provide stable and competitive performance without relying on external data or large-scale architectures.

1 Introduction

Political discourse is inherently rich in ambiguity, often serving strategic communicative purposes. In high-stakes settings such as presidential debates or interviews, politicians frequently employ evasive language that leaves audiences with multiple interpretations of whether the requested information was conveyed. This phenomenon, known as equivocation or evasion, has been extensively documented in political communication research, which shows that politicians provide direct replies significantly less often than non-politicians in comparable contexts (Bull and Mayer, 1993; Bull, 2012). Such findings highlight the strategic nature of political communication in public settings and motivate the need for automated methods that can assess response clarity at scale.

Automatically detecting ambiguity in political discourse remains challenging. Political language relies on rhetorical strategies, connotations and contextual dependencies that are difficult for surface-level models to capture (Paci et al., 2025). Since questions may address multiple issues simultaneously, accurate clarity assessment also requires reliable contextual grounding (Thomas et al., 2024). From an NLP perspective, ambiguity manifests at lexical, syntactic and pragmatic levels, and current language models still struggle with implicit meaning and speaker intention in political contexts (Fortuny and Payrató, 2024; Paci et al., 2025).

The CLARITY shared task on response ambiguity detection and classification in English (Thomas et al., 2026), organized as part of SemEval 2026, aims to advance computational methods for identifying and categorizing ambiguous responses in political discourse. The task focuses on QA pairs extracted from presidential debates and interviews, challenging systems to assess the clarity of a politician's reply in relation to the question posed. The task is divided into two "subtasks": (1) **Clarity-level Classification**, which requires systems to classify each answer as Clear Reply, Ambiguous or Clear Non-Reply; and (2) **Evasion-level Classification**, which extends this formulation by requiring fine-grained categorization of ambiguous responses into one of nine predefined evasion techniques derived from political communication theory.

In this work, we focus exclusively on the Clarity-level Classification subtask. Our approach is based on a weighted soft-voting ensemble that combines multiple fine-tuned encoder-only transformer models. Each model is trained to perform ternary classification over the clarity labels, and their predicted probability distributions are aggregated through

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 475–482

a weighted linear combination. The final prediction corresponds to the class with the highest accumulated weighted probability. This ensemble-based strategy aims to leverage architectural diversity and complementary prediction patterns to improve robustness in detecting ambiguous political responses.

On the official test set, our system achieved a macro-F1 score of 0.71, ranking 26th out of 41 participating teams. Our quantitative and qualitative analyses indicate that the system performs reliably on clear replies and clear non-replies, while it encounters greater difficulty when distinguishing genuinely ambiguous answers from subtly evasive but contextually grounded responses. These findings highlight both the effectiveness and the limitations of sequence-level modeling for capturing nuanced pragmatic phenomena in political discourse.

The remainder of this paper is organized as follows. Section 2 reviews related work on ambiguity detection, equivocation in political discourse and recent advances in transformer-based text classification. Section 3 presents the architecture of our weighted soft-voting ensemble and details the modeling approach adopted for the Clarity-level Classification subtask. Section 4 describes the dataset, the training configuration, hyperparameter selection and the evaluation performed. Section 5 reports and analyzes the official test set results, including quantitative performance and qualitative observations. Finally, Section 6 summarizes our main findings and outlines directions for future research.

2 Background Information

Political QA exchanges, particularly in televised interviews and debates, frequently exhibit strategic ambiguity, that is, responses that appear relevant while preserving multiple plausible interpretations or avoiding explicit commitment. In political communication research, this phenomenon is described as equivocation or evasion, arising from communicative dilemmas in which direct answers may be politically costly. Early theoretical accounts formalized equivocation as a situational strategy rather than a speaker-specific trait (Bavelas et al., 1988), while interview-based studies documented systematic avoidance patterns and interactionally well-formed non-answers (Bull and Mayer, 1993; Bull, 2012). Pragmatic perspectives further emphasize that clarity depends not only on literal content but also on connotation and interpretation (Drăm-nescu, 2016). These foundations motivate computational approaches that operationalize clarity as a measurable annotation target (Thomas et al., 2024).

Prior to large-scale NLP modeling, political interview research proposed operational schemes to distinguish answers from non-answers. Conversation-analytic and pragmatic frameworks described how politicians can preserve topical relevance while strategically withholding requested information (Harris, 1991). Bull and colleagues introduced systematic annotation distinctions between replies and non-replies, showing that evasiveness can be consistently identified in transcript data (Bull, 1994; Bull and Mayer, 1993), and subsequent microanalytic work consolidated recurring avoidance strategies within institutional interview dynamics (Bull, 2012). These structured taxonomies laid the groundwork for later computational formulations.

Modern clarity classification systems typically build upon pretrained transformer encoders, whose contextual representations can be fine-tuned for supervised text classification (Devlin et al., 2019; Liu et al., 2019). Improvements in encoder architectures, such as DeBERTa (He et al., 2020), and parameter-efficient adaptation methods, including adapters (Houlsby et al., 2019) and low-rank updates (LoRA) (Hu et al., 2022), have facilitated experimentation with larger models under computational constraints. In parallel, multi-task learning provides a principled way to share representations across related objectives (e.g., clarity and evasion), often improving generalization when labels are sparse or noisy (Ruder, 2017).

Ensembling remains a widely adopted strategy for improving predictive performance and robustness by combining diverse models (Dietterich, 2000; Alpaydin, 2007). Beyond majority voting, modern NLP systems frequently aggregate probabilities (soft voting) and may use higher-level combination schemes such as stacking, where a meta-learner is trained to fuse base predictions (Wolpert, 1992). Such approaches are particularly relevant in ambiguity detection settings, where different models may capture complementary pragmatic or semantic cues.

Recent work on the CLARITY dataset primarily employs transformer-based classifiers fine-tuned on QA pairs, reflecting the strength of pretrained encoders for supervised natural language understanding (Thomas et al., 2024). In addition, prompt-

based large language model approaches, including chain-of-thought and few-shot learning strategies, have been explored, revealing that prompt design can meaningfully affect clarity and evasion prediction quality (Prahallad et al., 2026). Complementary research suggests incorporating discourse and pragmatic signals, such as hedging or uncertainty detection (Farkas et al., 2010), persuasion-technique modeling (Dimitrov et al., 2021) and emotional analysis (Cochrane et al., 2022), which may contribute to improved ambiguity detection.

3 System overview

Figure 1 illustrates the overall system architecture for the clarity level classification subtask. At the core of the system lies a weighted soft-voting ensemble composed of multiple fine-tuned transformer-based models. Each model independently produces a probability distribution over the three clarity labels. These probabilities are aggregated through a weighted linear combination, where each model contributes according to a predefined weight reflecting its individual performance. The final prediction corresponds to the label with the highest aggregated weighted probability.

Given a question and its corresponding answer extracted from presidential debates or interviews, the system performs ternary text classification to determine whether the politician's response constitutes a Clear Reply ("Clear Reply", label 0), a Clear Non-Reply ("Clear Non-Reply", label 1) or an Ambiguous answer ("Ambivalent", label 2).

To construct the ensemble, six Transformer-based encoder-only models were evaluated: RoBERTa-large (Liu et al., 2019), BERT-large-cased (Devlin et al., 2019), DistilBERT-cased (Sanh et al., 2019), EuroBERT-610M (Boizard et al., 2025), NeoBERT (Breton et al., 2025) and ModernBERT-large (Warner et al., 2025). Each model was fine-tuned for the ternary clarity classification task using a supervised learning setup with a task-specific classification head.

Hyperparameter selection was performed through grid search independently for each model. The explored search space included the following values: learning rate $\in [1 \times 10^{-5}, 2 \times 10^{-5}, 3 \times 10^{-5}]$, per-device batch size $\in$ [8, 16], number of training epochs $\in$ [3, 5, 10] and weight decay $\in$ [0.0, 0.01]. The final configuration for each model was selected based on its performance on the development set.

After obtaining the optimized individual models, we conducted a systematic search to identify the best-performing ensemble configuration. Specifically, we evaluated all possible model combinations ranging from pairs to the full six-model ensemble. For each combination, we explored a weighted soft-voting scheme where model weights were drawn from the grid [0.0, 0.5, 1.0, 1.5, 2.0] and subsequently normalized to sum to one.

Given a set of M models and their corresponding normalized weights $w_1, \ldots, w_M$, the final prediction for a given instance is computed by aggregating the weighted class probabilities:

$$\hat{y} = \arg\max_{c \in \mathcal{C}} \sum_{m=1}^{M} w_m \cdot P_m(c)$$

where $P_m(c)$ denotes the probability assigned by model m to class c and w_m is its normalized weight.

Each ensemble configuration was evaluated using macro-averaged F1-score on the validation set. The final system corresponds to the combination and weight assignment achieving the highest macro-F1 score.

We used the HuggingFace Transformers library to perform fine-tuning for the clarity level classification subtask, formulating it as a ternary text classification problem. Each input instance was assigned a single label according to the clarity level of the answer: 0 for Clear Reply, 1 for Clear Non-Reply and 2 for Ambiguous.

For the encoder-only models, a task-specific classification head was added on top of the encoder. This head consists of a single linear layer applied to the contextualized representation of the `[CLS]` token, producing logits for the three target classes. These logits were optimized using cross-entropy loss.

Fine-tuning was carried out using the `Trainer` API. Standard preprocessing steps were applied prior to training, including tokenization of the input texts, truncation or padding to a fixed maximum sequence length, and dynamic batching for efficient GPU utilization.

The tokenizer automatically handled the insertion of special tokens and padding according to the requirements of each model architecture. Since the task was formulated as sequence-level classification, no additional label masking or token-level supervision was required.

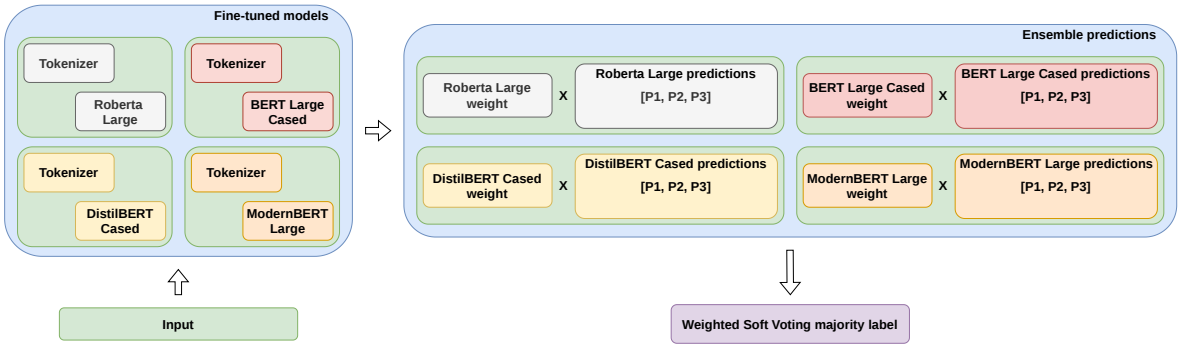

Figure 1: System architecture for the clarity level classification subtask.

4 Experimental setup

For the clarity level classification subtask, our experiments were based exclusively on the training and development sets provided by the task organizers. Table 1 shows the number of examples and class distribution for each split.

Table 1: Clarity level classification subtask dataset distribution. Each cell reports the number of examples and the percentage they represent.

	Training	Development
Total examples	3448	308
Clear Reply	1052 (30.5%)	79 (25.6%)
Clear Non-Reply	356 (10.3%)	23 (7.5%)
Ambiguous	2040 (59.2%)	206 (66.9%)

To build the final ensemble for this subtask, we selected the four models that achieved the highest macro-F1 scores on the development set: RoBERTa-large, BERT-large-cased, DistilBERT-cased and ModernBERT-large. The ensemble weights were determined through the systematic search procedure described in Section 3, selecting the configuration that maximized macro-F1 on the development split.

The optimal weighted soft-voting configuration assigned the following raw weights (prior to normalization) to each model: RoBERTa-large (1.5), BERT-large-cased (0.5), DistilBERT-cased (2.0) and ModernBERT-large (0.5). These weights were subsequently normalized to sum to one and used to aggregate the class probability distributions produced by the individual models.

Regarding hyperparameters, the following configurations correspond to the best-performing settings identified during the grid search procedure described in Section 3, selected based on macro-F1 score on the development set. RoBERTa-large was fine-tuned for 5 epochs with a learning rate of 1×10^{-5}, no weight decay and a batch size of 8. BERT-large-cased was fine-tuned for 3 epochs with a learning rate of 2×10^{-5}, no weight decay and a batch size of 8. DistilBERT-cased was fine-tuned for 3 epochs with a learning rate of 3×10^{-5}, a weight decay of 0.01 and a batch size of 8. Finally, ModernBERT-large was fine-tuned for 5 epochs with a learning rate of 3×10^{-5}, no weight decay and a batch size of 8.

In addition to the hyperparameter configuration, we provide further implementation details regarding input representation and training configuration for reproducibility. Each instance was constructed by concatenating the question and the corresponding answer into a single sequence using the format: `Question: <question> Answer: <answer>`. The resulting sequence was tokenized using the corresponding tokenizer of each model, with truncation and padding applied to a maximum sequence length of 128 tokens.

All encoder-based models were fully fine-tuned end-to-end using the HuggingFace Transformers Trainer API, updating all parameters during training. No layers were frozen, and the classification head was trained jointly with the encoder using cross-entropy loss.

5 Results

In this section, we present and analyze the results obtained by our top-performing submission on the official test set. During the development phase, we conducted multiple experiments, exploring different fine-tuning configurations, alternative model combinations for the ensemble and various weighting strategies for aggregating model predictions in the clarity level classification subtask. The results discussed here correspond to the best-performing

configuration submitted.

To better understand the contribution of the proposed ensemble approach, we compare its performance on the development set against the individual models that compose the final ensemble, as well as an unweighted soft-voting variant of the ensemble. Table 2 summarizes the macro-F1 scores obtained by each model and the different ensembling strategies.

Table 2: Comparison between individual models and ensemble strategies on the development set.

Model	Macro-F1 score
RoBERTa-large	0.60
BERT-large-cased	0.56
DistilBERT-cased	0.55
ModernBERT-large	0.43
Unweighted ensemble	0.63
Weighted ensemble	**0.66**

As shown in Table 2, the weighted ensemble outperforms both the individual models and the unweighted soft-voting variant. In particular, it achieves an improvement of +0.06 macro-F1 points over the best-performing individual model (RoBERTa-large) and +0.03 points over the unweighted ensemble. These results highlight the benefit of combining complementary prediction patterns across heterogeneous architectures, while assigning differentiated importance to each model further improves performance.

Table 3 reports the official results of the clarity level classification subtask, ranked by macro-averaged F1-score (the official evaluation metric). Our system (*tbernal*) ranked 26th overall, achieving an F1-score of 0.71.

As shown in Table 3, the top-performing systems achieved macro-F1 scores above 0.85, with the best submission reaching 0.89. Our approach obtained a macro-F1 score of 0.71, ranking 26th out of 41 participating systems. This positions our model within the central performance band of the competition, clearly outperforming lower-ranked systems while remaining below the leading approaches.

The performance gap between our system and the top-ranked submissions (approximately 0.18 macro-F1 points) suggests that additional gains could potentially be achieved through more sophisticated modeling strategies, such as larger-scale architectures, external data augmentation or more advanced ensembling techniques. Nevertheless, the proposed weighted soft-voting ensemble demonstrates stable and competitive performance using only task-provided data and standard fine-tuning procedures. Importantly, our approach relies exclusively on encoder-only models of moderate size and does not incorporate external resources, making it a comparatively lightweight and computationally efficient solution.

Table 3: Clarity level classification subtask results, the table shows the F1-score (evaluation metric) obtained.

Place	Name	F1-score
1	TeleAI	0.89
2	AsymVerify	0.85
3	CSE-UOI	0.85
4	Rasende Rakete	0.83
5	Evaluators	0.83
	...	
26	**tbernal**	**0.71**
27	rafsan	0.68
28	mikebeth	0.68
	...	
33	B&B	0.64
34	AI@UMS	0.62
35	uir_cis	0.61
	...	
39	Happy frogs	0.42
40	lakksh	0.31
41	yx-ym	0.28

It is important to consider the dataset characteristics when interpreting these results. The class distribution is notably skewed toward the Ambiguous category (approximately 60% of the training data), which increases the difficulty of optimizing macro-F1 score, as improvements in minority classes have a stronger influence on the final score. In this work, we deliberately preserve the original class distribution and do not apply explicit balancing techniques (e.g., resampling or class weighting), in order to reflect the natural data conditions defined by the task and avoid introducing additional biases during training.

In this context, the ensemble approach contributes to performance stability by combining complementary prediction patterns from heterogeneous architectures. The relatively higher weight assigned to DistilBERT-cased in the optimal configuration further indicates that architectural diversity, rather than model size alone, plays a meaningful role in capturing ambiguity-related cues.

To provide a more detailed understanding of the system behavior, we further analyze performance at the class level. Table 4 reports the per-class precision, recall and F1-score obtained by the final weighted ensemble on the development set.

Table 4: Per-class performance of the final weighted ensemble on the development set.

Class	Precision	Recall	F1-score
Clear Reply	0.64	0.53	0.58
Clear Non-Reply	0.69	0.48	0.56
Ambiguous	0.79	0.87	0.83

As shown in Table 4, performance varies significantly across classes, largely influenced by the dataset distribution. The Ambiguous class, which represents the majority of the data, achieves the highest performance, with an F1-score of 0.83, supported by both high precision (0.79) and recall (0.87).

In contrast, the Clear Reply and Clear Non-Reply classes exhibit lower performance, with F1-scores of 0.58 and 0.56, respectively. In particular, both classes show reduced recall (0.53 for Clear Reply and 0.48 for Clear Non-Reply), indicating that the ensemble tends to misclassify these instances, often predicting them as Ambiguous. This behavior is consistent with the class imbalance present in the dataset, where the dominance of the Ambiguous class biases the model towards this category.

Overall, these results highlight the inherent difficulty of distinguishing subtle pragmatic differences between response types, especially in underrepresented classes. They also suggest that improving minority class recognition remains a key challenge for future work.

6 Conclusion and Further Work

In this work, we presented our system for the SemEval 2026 CLARITY shared task, focusing on the clarity level classification subtask. Our approach is based on a weighted soft-voting ensemble that combines four fine-tuned encoder-only transformer models. The final system achieved a macro-F1 score of 0.71 on the official test set, ranking 26th out of 41 participating teams.

The results demonstrate that combining heterogeneous encoder-based architectures through probability-level aggregation provides a stable and competitive solution for determining the clarity level of politicians' responses in English-language presidential debates and interviews. Despite relying exclusively on task-provided data and standard fine-tuning procedures, the proposed ensemble achieved solid mid-tier performance while maintaining a relatively lightweight and computationally efficient design.

The performance gap with respect to the top-ranked systems suggests that clarity classification remains a challenging problem, particularly in cases involving Ambiguous responses. These instances often exhibit subtle linguistic cues, implicit evasion strategies or partial answers that blur the boundary between clear replies and non-replies. This highlights the intrinsic difficulty of modeling ambiguity in political discourse using sequence-level classification alone.

For future work, several directions may further improve performance. First, incorporating larger pretrained architectures or parameter-efficient adaptation techniques could enhance representational capacity without significantly increasing computational cost (Bernal-Beltrán et al., 2025; Xu et al., 2026). Second, exploring more advanced ensembling strategies, such as probability calibration, stacking or meta-learning approaches, may allow for more effective aggregation of complementary model behaviors (Malebary and Abulfaraj, 2024; Ruder, 2017).

Finally, integrating discourse-level or pragmatic features, such as emotional and rhetorical analysis, potentially through multi-task learning or auxiliary objectives, could provide deeper insight into ambiguity patterns beyond surface-level lexical signals (Song et al., 2025; Pan et al., 2025). Enhancing the system's ability to differentiate between genuinely clear responses and strategically ambiguous ones is particularly relevant, since political discourse often relies on persuasive strategies, hedging, and emotionally charged language that may contribute to perceived ambiguity.

Acknowledgments

This work is part of the research project LaTe4PoliticES (PID2022-138099OB-I00) funded by MICIU/AEI/10.13039/501100011033 and the European Regional Development Fund (ERDF/EU - FEDER/UE)-a way of making Europe. Mr. Tomás Bernal-Beltrán is supported by University of Murcia through the predoctoral programme.

References

Ethem Alpaydin. 2007. Combining pattern classifiers: Methods and algorithms (kuncheva, li; 2004)[book review]. *IEEE Transactions on Neural Networks*, 18(3):964–964.

Janet Beavin Bavelas, Alex Black, Lisa Bryson, and Jennifer Mullett. 1988. Political equivocation: A situational explanation. *Journal of Language and Social psychology*, 7(2):137–145.

Tomás Bernal-Beltrán, Ronghao Pan, José Antonio García-Díaz, and Rafael Valencia-García. 2025. Umuteam at ta1c 2025: Leveraging large language models for identifying and spoiling clickbait in spanish.

Nicolas Boizard, Hippolyte Gisserot-Boukhlef, Duarte M Alves, André Martins, Ayoub Hammal, Caio Corro, Céline Hudelot, Emmanuel Malherbe, Etienne Malaboeuf, Fanny Jourdan, et al. 2025. Eurobert: Scaling multilingual encoders for european languages. *arXiv preprint arXiv:2503.05500*.

Lola Le Breton, Quentin Fournier, Mariam El Mezouar, John X Morris, and Sarath Chandar. 2025. Neobert: A next-generation bert. *arXiv preprint arXiv:2502.19587*.

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

Peter Bull. 2012. The microanalysis of political discourse. *Philologia Hispalensis*.

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

Christopher Cochrane, Ludovic Rheault, Jean-François Godbout, Tanya Whyte, Michael W-C Wong, and Sophie Borwein. 2022. The automatic analysis of emotion in political speech based on transcripts. *Political Communication*, 39(1):98–121.

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.

Thomas G Dietterich. 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Semeval-2021 task 6: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021)*, pages 70–98.

Bianca Drămnescu. 2016. Pragmatic approaches in the analysis of the political discourse. *Philosophy, communication, media sciences*, 4(4):45–51.

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The conll-2010 shared task: learning to detect hedges and their scope in natural language text. In *Proceedings of the fourteenth conference on computational natural language learning–Shared task*, pages 1–12.

Jordi Fortuny and Lluís Payrató. 2024. Ambiguity in linguistics 1. *Studia Linguistica*, 78(1):1–7.

Sandra Harris. 1991. Evasive action: How politicians respond to questions in political interviews. *Broadcast talk*, 7699.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

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. *arXiv preprint arXiv:1907.11692*.

Sharaf J Malebary and Anas W Abulfaraj. 2024. A stacking ensemble based on lexicon and machine learning methods for the sentiment analysis of tweets. *Mathematics*, 12(21):3405.

Walter Paci, Alessandro Panunzi, and Sandro Pezzelle. 2025. They want to pretend not to understand: The limits of current llms in interpreting implicit content of political discourse. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15569–15593.

Ronghao Pan, Jose Antonio Garcia-Diaz, and Rafael Valencia-Garcia. 2025. Spanish mtlhatecorpus 2023: Multi-task learning for hate speech detection to identify speech type, target, target group and intensity. *Computer Standards & Interfaces*, 94:103990.

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

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Changhao Song, Yazhou Zhang, Hui Gao, Ben Yao, and Peng Zhang. 2025. Large language models for subjective language understanding: A survey. *arXiv preprint arXiv:2508.07959*.

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.

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

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547.

David H Wolpert. 1992. Stacked generalization. *Neural networks*, 5(2):241–259.

Lingling Xu, Haoran Xie, S Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2026. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.