aset_clarity at SemEval-2026 Task 6: An Imbalance-Aware RoBERTa Cross-Encoder for Political Response Clarity Classification

Maria-Antonia-Emanuela Pascu[1], Dan Dodun-des-Perrieres[2], Daniela Gifu[2,3],
[1]Faculty of Computer Science, Alexandru Ioan Cuza University of Iaşi, Romania
[2]Institute of Computer Science, Romanian Academy – Iaşi Branch, Romania
[3]Academy of Romanian Scientists, Romania
{maria.pascu, dan.dodun}@info.uaic.ro
daniela.gifu@iit.academiaromana-is.ro

Abstract

We address response-clarity classification in political interviews as defined in SemEval-2026 Task 6: CLARITY - Unmasking Political Question Evasions, Task 1, where systems must label each question–answer pair as *Clear Reply*, *Ambivalent*, or *Clear Non-Reply*. We present a reproducible end-to-end pipeline built around a single-stream RoBERTa-large cross-encoder fine-tuned for three-way classification using deterministic text normalization, concatenated QA inputs, and imbalance-aware training (minority oversampling and class-weighted loss). To improve robustness, we train a 5-fold stratified ensemble and combine models via soft-voting. Our official shared-task submission obtains 0.76 macro-F1 on the official leaderboard, ranking 16 out of 41 participating systems. Finally, we deploy the classifier in a lightweight web application supporting both direct text input and audio-based analysis through automatic transcription, enabling interactive inspection of predicted clarity categories.

1 Introduction

Political interviews frequently involve strategic ambiguity, partial replies, or overt evasions, making it difficult to determine whether a speaker has meaningfully addressed a question. Task 6: CLARITY - Unmasking Political Question Evasions formalizes this challenge by requiring systems to classify each question–answer pair as *Clear Reply*, *Ambivalent*, or *Clear Non-Reply* (Thomas et al., 2024), (Thomas et al., 2026). The task highlights a persistent difficulty: the pragmatic boundary between Clear Reply and Ambivalent (ambiguous/partially evasive), where answers may acknowledge a topic while reframing or withholding essential information. This motivates a legitimate research question: *How can a computational model reliably distinguish between clear replies, partial replies, and evasions in political interviews, given the subtle pragmatic cues that define the Ambivalent category?*

To address this question, we develop a single-stream RoBERTa-large cross-encoder trained with imbalance-aware learning and stratified 5-fold ensembling, and we deploy it in an end-to-end application that mirrors the shared-task inference pipeline for both text-based and audio-derived inputs. Our contributions are threefold:

1. A reproducible RoBERTa-large cross-encoder using only question and answer text fields, avoiding metadata shortcuts.
2. An imbalance-aware training and ensembling strategy that stabilizes performance on borderline Clear–Ambivalent cases.
3. A deployable end-to-end application enabling interactive clarity analysis for text and audio. Under the official shared-task evaluation, our system achieves 0.76 macro-F1.

The report also provides a public repository link with the released code. [1]

2 Background

Research on political interviews has long documented how speakers manage accountability through equivocation, reframing, and selective omission. Foundational studies describe how politicians avoid answering questions directly (Bull, 1994, 2003; Clayman, 2001; Bull and Mayer, 1993), while more recent computational work formalizes clarity and evasion taxonomies (Thomas et al., 2024) and motivates the Task 6: CLARITY - Unmasking Political Question Evasions (Thomas et al., 2026). Complementary perspectives come from discourse analysis and political communication. Romanian and European scholarship has shown that symbolic violence, strategic framing, and partial answering are central mechanisms in

[1] https://github.com/dand101/ASET_2025_CLARITY

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1068–1074

political discourse, offering theoretical grounding for computational clarity modeling. Relevant contributions include analyses of symbolic violence and electoral discourse (Gifu, 2010); (Gifu, 2011), integrative perspectives on political communication (Gifu, 2013), opinion and factivity in political speech (Delmonte et al., 2013), propaganda detection (Ermurachi and Gîfu, 2020), and diachronic semantic variation relevant to political meaning shifts (Gifu, 2016). From a computational standpoint, transformer-based encoders such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) have become standard for modeling contextual semantics, including tasks involving pragmatic inference and discourse relations. Their ability to jointly encode question–answer pairs makes them well-suited for clarity classification, where subtle pragmatic cues often determine the correct label.

Together, these strands of research motivate systems that combine contextual semantic modeling, pragmatic sensitivity, and interpretability, aligning directly with the goals of this shared task.

3 Dataset and Methods

This section outlines the dataset used in our experiments and the methodological framework underlying our system.

3.1 Dataset

We use the official Task 1 dataset of the SemEval 2026 Task 6: CLARITY - Unmasking Political Question Evasions, which contains English question–answer (QA) pairs extracted from political interviews, presidential press conferences, and news conference transcripts (Thomas et al., 2026), (Thomas et al., 2024; AILS NTUA, 2026). Each instance is annotated with one of three labels:

- *Clear Reply*
- *Ambivalent*
- *Clear Non-Reply*

This dataset is imbalanced, with substantially fewer *Clear Non-Reply* examples (see Figure 1). The imbalance, motivates macro-F1 evaluation and imbalance-aware training (oversampling and class-weighted loss).

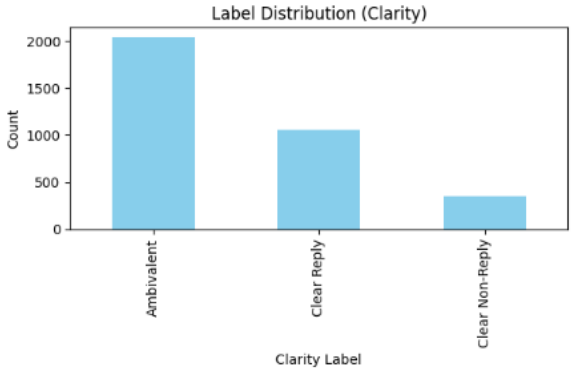


Figure 1: Class distribution in the CLARITY Task 1 dataset (public release).

Following the shared task definition, we use only the textual fields `interview_question` and `interview_answer`, deliberately excluding metadata (speaker, date, URL) to avoid correlational shortcuts and to ensure that the model learns the pragmatic relation between question and answer.

Preprocessing is lightweight and deterministic: Unicode normalization, whitespace collapsing, trimming, and optional filtering of empty, extremely short, or "inaudible" segments. No additional cleaning is applied, as clarity cues often rely on subtle lexical and pragmatic signals.

3.2 Input Construction and Preprocessing

For each instance, we construct a single stream QA sequence by concatenating the question and answer using a fixed template. This deterministic construction ensures that the model receives a consistent representation of the interactional structure. We tokenize using the pretrained RoBERTa tokenizer, applying truncation/padding to a maximum length of 512 tokens. We do not use token type embeddings or metadata features, keeping the modeling assumptions minimal and fully aligned with the shared task constraints.

3.3 System Overview

Our system employs a RoBERTa-large cross-encoder (Liu et al., 2019; Devlin et al., 2019) fine-tuned for three-way clarity classification. Each question–answer pair is concatenated into a single input sequence and tokenized with the pretrained RoBERTa tokenizer, using a maximum length of 512 tokens. The resulting input IDs and attention masks are provided to the encoder. The overall architecture is shown in Figure 2.

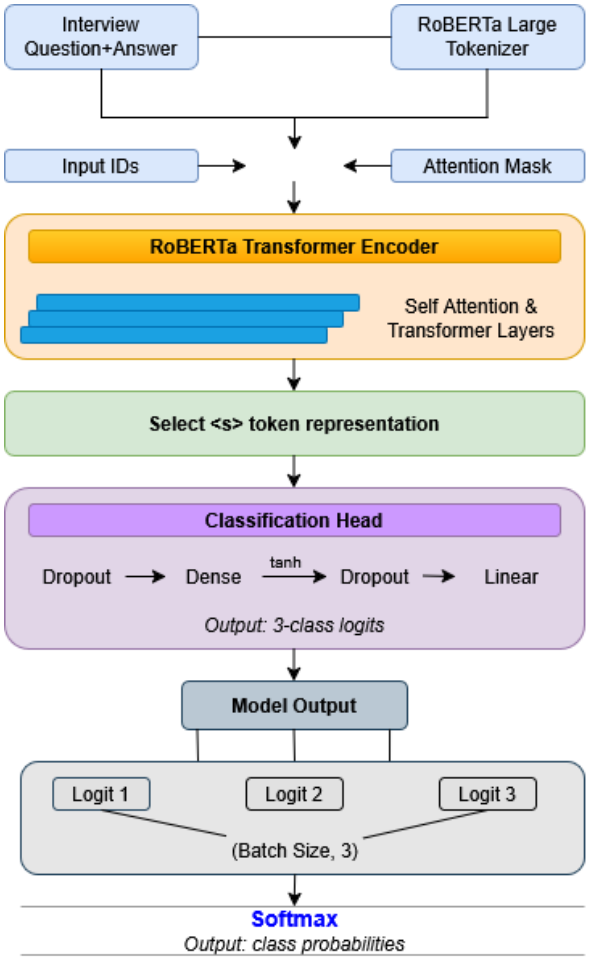


Figure 2: Overview of the proposed system.

RoBERTa-large comprises 24 Transformer layers with hidden size $d = 1024$. For a batch of size B and sequence length L, the encoder outputs contextualized token representations: $\mathbf{X} \in \mathbb{R}^{B \times L \times d}$.

For sequence-level prediction, we use the final-layer representation of the first token (<s>) as the sequence representation. Let $\mathbf{s}^{(i)} \in \mathbb{R}^d$ denote the <s> representation for the i-th input; stacking these vectors yields: $\mathbf{S} \in \mathbb{R}^{B \times d}$. This $\mathbf{S}$ serves as the fixed-dimensional encoding of each question–answer pair for classification.

The representation $\mathbf{S}$ is passed through the standard RoBERTa classification head, consisting of dropout, a dense projection ($d \to d$) with `tanh` activation, a second dropout layer, and a final linear projection to three output logits. The model produces $\mathbf{z} \in \mathbb{R}^{B \times 3}$, corresponding to the labels *Clear Reply*, *Ambivalent*, and *Clear Non-Reply*.

Training minimizes class-weighted cross-entropy computed directly on the logits. During inference, softmax is applied to obtain class probabilities.

3.4 Handling Class Imbalance

The dataset is highly imbalanced, with Ambivalent as the majority class and substantially fewer Clear Reply and Clear Non-Reply instances. We therefore applied oversampling only to the two minority classes, Clear Reply and Clear Non-Reply, while no Ambivalent examples were added.

We used two augmentation strategies. First, in preliminary experiments, we applied a lightweight rule-based answer-level augmentation method. This method did not generate new QA pairs, but produced label-preserving variants of existing minority-class answers through small lexical and discourse-level edits, such as hedge replacement, discourse-marker insertion or replacement, and optional that-clause toggling.

Second, for the final system, we used GPT-4.1 to generate synthetic question–answer pairs for the two minority classes. We generated Clear Reply and Clear Non-Reply examples. GPT-4.1 was prompted with real training examples and instructed to preserve the target label semantics without copying the original wording. The prompt specified the target label, the expected answer behavior, and the output format: question, answer, label. Clear Reply examples were required to answer the question directly, whereas Clear Non-Reply examples were required to avoid answering it while remaining topically plausible.

All augmented examples were added only to the training split of each fold, never to validation or test data. We manually inspected the GPT-4.1 examples and removed duplicates, off-topic cases, label-inconsistent samples, and examples that were too similar to the original prompts. Class weights were computed after augmentation, using the effective training distribution.

3.5 Training and Ensembling

We fine tune RoBERTa large using: 3 epochs; batch size 8; `max_len`=512; mixed precision (bfloat16); linear warmup, and early stopping based on development macro F1.

To improve robustness, we train a 5 fold stratified ensemble, producing five independently fine tuned models. At inference time, we apply soft-voting, averaging per class-softmax probabilities across folds. Hyperparameters are selected via Bayesian optimization (learning rate, weight decay, warmup ratio, scheduler, label smoothing, augmentation probability).

Training optimizes class-weighted cross-entropy. If the model produces logits $z_i \in \mathbb{R}^3$ for instance i, with gold label y_i and class weight w_{y_i}, the loss is:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} w_{y_i} \log \mathrm{softmax}(z_i)_{y_i}. \quad (1)$$

3.6 Deployment Pipeline

For completeness, we deploy the trained ensemble in a lightweight web application that mirrors the shared-task inference pipeline. The application supports two input modes: direct text input and audio-based analysis. In the text mode, the user provides an interview question and answer, which are normalized, concatenated using the same fixed template as during training, and passed to the RoBERTa ensemble.

In the audio mode, the uploaded audio file is first transcribed automatically using an ASR module based on Whisper. The resulting transcript is then split into question and answer fields by the user, or corrected manually if needed, before being processed by the same classifier. This design keeps the prediction stage identical for text and audio inputs, while allowing users to inspect and correct transcription errors before classification.

The source code for the tool is publicly available through the released repository, together with the training and inference code, although we do not provide a permanently hosted public demo. We note that the hosted web interface is intended as a lightweight demonstration rather than a production system. In particular, ASR errors may affect named entities, negation, or short answers, which can increase ambiguity and lead to incorrect clarity predictions.

4 Experimental Setup

All experiments follow the official SemEval-2026 CLARITY Task 1 protocol, using only the dataset splits released by the organizers and adhering strictly to the shared-task evaluation rules. No external labeled data is introduced at any stage. Model selection and reporting are based on macro-averaged F1, the primary metric of the task.

4.1 Hyperparameter Search

Hyperparameters are optimized through two Bayesian sweeps tracked in Weights and Biases, using macro-F1 as the objective. The search space includes learning rate, weight decay, warmup ratio, scheduler type, label smoothing, and online augmentation probability. The selected configuration—learning rate 7.42×10^{-6}, weight decay 0.05, warmup ratio 0.05, linear scheduler, and no label smoothing or online augmentation—is reused unchanged for all cross-validation folds.

4.2 Cross-Validation and Ensembling

To improve robustness and reduce variance, we employ 5-fold `Stratified` cross-validation. Each fold is trained independently, with augmentation and balancing applied only to the corresponding training split to avoid leakage. At inference time, we aggregate predictions via soft-voting by averaging per-class softmax probabilities across folds.

4.3 Evaluation Sets and Metrics

Evaluation follows the shared-task protocol: macro-averaged F1 is the primary metric, ensuring equal weight for all three classes despite label imbalance. We report both the official leaderboard score (0.76 macro-F1) and reproducible results on the public Task 1 test split (308 instances), including per-class metrics and diagnostic analyses such as confusion matrices and ROC/PR curves.

4.4 Implementation

The system is implemented using Hugging Face Transformers for model training and inference, scikit-learn for cross-validation and metrics, and Weights and Biases for experiment tracking. All components are fully reproducible and consistent with the shared-task constraints.

5 Results

This section reports the performance of our system on the SemEval-2026 CLARITY Task 1 benchmarks. We present the official leaderboard score, reproducible results on the public test split, and a brief analysis of the main error patterns.

5.1 Main Performance

Our final 5-fold soft-voting ensemble achieves 0.76 macro-F1 on the official shared-task leaderboard. On the public Task 1 test split (308 instances), the system obtains 0.710 macro-F1 and 0.749 weighted F1, with per-class scores reported in Table 2. Per-class precision, recall, and F1 scores are reported in Table 3, with the highest performance on Ambivalent (F1=0.815) and the lowest on Clear (F1=0.582), reflecting the difficulty of borderline partial-answer cases.

System / attempt	Macro-F1
RoBERTa cross-encoder, plain fine-tuning	0.58
RoBERTa bi-encoder, absolute-difference matching	0.61
Our final system (public test split)	**0.710**

Table 1: Summary of early baselines/attempts and our final public test performance.

Result	Macro-F1
Official shared-task evaluation	
Leaderboard score (official submission)	0.76
Public split (reproducible)	
Public Task 1 test split (308 instances)	0.710

Table 2: Macro-F1 on the official shared-task evaluation (leaderboard) and on the public Task 1 test split. These scores are computed on different evaluation sets.

Label	P	R	F1	Support
Clear	–	–	0.582	79
Ambivalent	0.806	0.825	0.815	206
Clear Non-Reply	0.833	0.652	0.732	23
Macro avg	–	–	0.710	308
Weighted avg	–	–	0.749	308

Table 3: Per-class performance on the CLARITY Task 1 public test split (308 instances).

Compared to the baselines in Table 1, our final system improves over two earlier RoBERTa-based configurations. The plain classifier uses the same single-stream input format as our final model, i.e., the interview question and answer are concatenated and encoded jointly by RoBERTa, followed by the standard sequence classification head. However, this baseline is trained directly on the original training data, without minority-class augmentation, class-weighted loss, or cross-validation ensembling. Its lower macro-F1 score (0.58) suggests that simple fine-tuning is insufficient under the strong class imbalance of the task.

The second baseline is a RoBERTa bi-encoder. In this setup, the question and answer are encoded separately using shared RoBERTa weights, producing two independent sequence representations from the final <s> token. These representations are combined using their absolute difference and passed to a linear classification layer for three-way prediction. This model reaches 0.61 macro-F1, slightly improving over the plain classifier, but remains below the cross-encoder ensemble, indicating that early token-level interaction between the question and answer is beneficial for response-clarity classification.

5.2 Ablation Study

We evaluate the contribution of the three main components of the final RoBERTa cross-encoder: minority oversampling, class-weighted loss, and 5-fold stratified ensembling. All variants use the same concatenated question–answer input, preprocessing, and hyperparameters, and are evaluated on the public Task 1 test split.

Model configuration	Macro-F1
Plain RoBERTa classifier	0.58
+ rule-based answer-level augmentation	0.65
+ class-weighted loss	0.62
+ rule-based augmentation and class-weighted loss	0.68
+ GPT-4.1 synthetic augmentation and 5-fold ensemble	0.710

Table 4: Ablation of the proposed components.

The results show that the rule-based answer-level augmentation provides the largest individual improvement over the plain RoBERTa classifier, increasing macro-F1 from 0.58 to 0.65. This augmentation applies label-preserving lexical and discourse-level edits to minority-class answers, such as hedge replacement, discourse-marker insertion, and optional that-clause toggling. Class-weighted loss provides a smaller improvement, while the final system further benefits from GPT-4.1 synthetic augmentation and 5-fold stratified ensembling.

5.3 Error Characteristics

The following error analysis is performed on the public Task 1 test split, not on the hidden leaderboard test set. The hidden test set was used only for the official leaderboard score, since gold labels were not available for diagnostic analyses.

Most misclassifications occur at the *Clear* vs. *Ambivalent* (F1 = 0.815) boundary. The normalized confusion matrix in Figure 5 shows that 42% of *clear* instances are predicted as *ambiguous*, while a *clear non-reply*, despite limited support, is more easily separable. The macro one-vs-rest ROC curves in Figure 3 indicate substantial overlap for *Ambivalent* (AUC=0.756), whereas *Clear Non-Reply* is the most distinct class (AUC=0.920). The corresponding precision-recall curves in Figure 4 show the highest average precision for Ambivalent (AP=0.840), followed by *Clear Non-Reply*

(AP=0.694), while *Clear* remains the most challenging class due to its pragmatic proximity to partial replies.

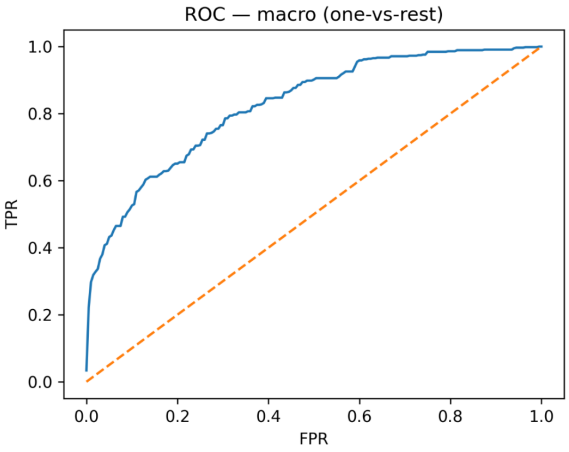
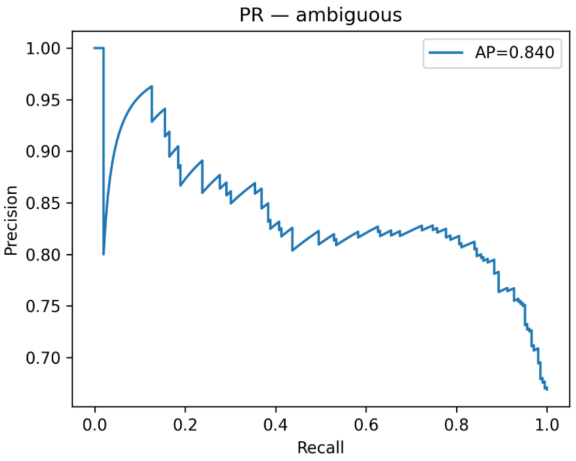
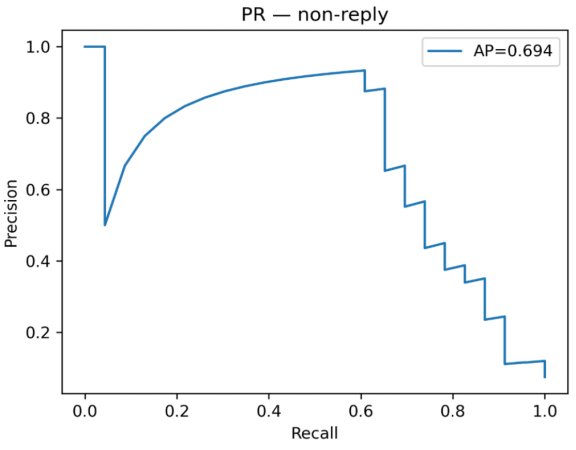


Figure 3: Macro one-vs-rest ROC curve on the public test split.



(a) *Ambivalent*



(b) *Clear Non-Reply*

Figure 4: Precision–recall curves on the public test split.

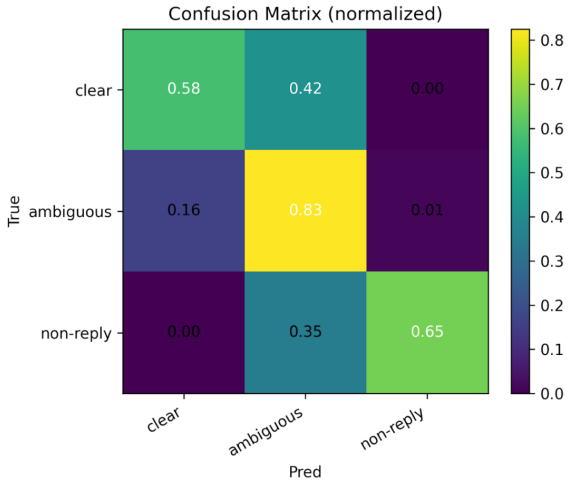


Figure 5: Normalized confusion matrix on the public test split.

Qualitative inspection reveals that errors typically involve answers that acknowledge the topic but avoid committing to requested details, aligning with the pragmatic ambiguity targeted by the task.

6 Conclusion

This study addressed a central question for the SemEval - 2026 Task 6: CLARITY - Unmasking Political Question Evasions, Task 1,: *to what extent can current NLP models reliably assess the clarity of political answers?* All these experiments show that a RoBERTa-large cross-encoder with deterministic preprocessing, imbalance-aware training, and 5-fold ensembling surpasses reported baselines. Ranking competitively on the official leaderboard (0.76 macro-F1), it establishes a reproducible baseline under shared-task constraints.

At the same time, this analysis reveals that the clarity assessment is far from being solved. Errors persist at the *Clear/Ambivalent* boundary; *Clear Non-Reply* generalization is limited by sparsity; and ASR noise affects performance in realistic deployments. This aligns with prior: pragmatic underspecification and evasive intent often require modeling beyond surface semantics.

Overall, our contribution demonstrates that carefully engineered fine-tuning pipelines can push the performance frontier for the CLARITY Task 1, but further progress requires modeling pragmatic, discourse-level cues and robustness to noisy or incomplete input. We hope that the system, analyses, and deployment insights presented here will support future research toward more reliable and context-aware clarity assessment in political and other high-stakes domains.

Acknowledgments

This work was carried out partially within the project "Tools for Processing Online Texts Specific to Cultural and Scientific Diplomacy", funded by the Academy of Romanian Scientists.

References

AILS NTUA. 2026. QEvasion: CLARITY task 1 public dataset. Hugging Face Datasets. Accessed 2026-01-28.

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. 2003. The analysis of equivocation in political interviews. In Glynis M. Breakwell, editor, *Doing Social Psychology Research*, pages 205–228. Wiley-Blackwell, Oxford, UK.

Peter Bull and Kate Mayer. 1993. How not to answer questions in political interviews. *Political Psychology*, 14(4):651–666.

Steven E. Clayman. 2001. Answers and evasions. *Language in Society*, 30(3):403–442.

Rodolfo Delmonte, Rocco Tripodi, and Daniela Gîfu. 2013. Opinion and factivity analysis of italian political discourse. In *Proceedings of the 4th edition of the Italian Information Retrieval Workshop (IIR 2013)*, pages 88–99, Pisa, Italy. CEUR-WS on-line proceedings series.

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 (NAACL-HLT)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Vlad Ermurachi and Daniela Gîfu. 2020. Uaic1860 at semeval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation, (SemEval-2020)*, pages 1835–1840, Barcelona, Spain. Association for Computational Linguistics.

Daniela Gifu. 2010. The written press discourse and symbolic violence. presidential election analysis. PhD, Alexandru Ioan Cuza Universtiy of Iasi, Romania.

Daniela Gifu. 2011. In *Violenta simbolica in discursul electoral*, Cluj-Napoca, Romania. Casa Cartii de Stiinta.

Daniela Gifu. 2013. In *Temeliile Turnului Babel*, Bucuresti, Romania. Editura Academiei Romane.

Daniela Gifu. 2016. Lexical semantics in text processing. contrastive diachronic studies on romanian language. PhD, Alexandru Ioan Cuza Universtiy of Iasi, Romania.

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.

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.