FER at SemEval-2026 Task 6: Analysis of Different Approaches to Unmasking Political Question Evasions

Matija Akrap, Andrija Bilić, Luka Čuturilo, Fran Račić, Roko Šimpraga
Faculty of Electrical Engineering and Computing
University of Zagreb
{first.last}@fer.hr

Abstract

We tackle classifying evasive political answers within the context of SemEval-2026 Task 6 and compare three modeling strategies: a flat baseline, a hierarchical cascade, and a multitask learning approach. Our experiments demonstrate that a hierarchical RoBERTa-base model achieves the best performance, particularly by leveraging the distinctiveness of the class *Clear Non-Reply*. Conversely, we find that standard multitask learning frequently produces structurally invalid label combinations in a significant fraction of predictions. Our demonstrations show that applying a constrained inference mask eliminates these errors entirely while improving F_1 performance, whereas a fully joint training approach underperforms due to data sparsity. Finally, we employ dataset cartography to compare training dynamics between the hierarchical and multitask approach.

1 Introduction

Politicians often give evasive answers to sensitive or controversial questions during interviews. Such evasions can obscure accountability and mask true intentions, making their automatic detection an important problem for political science and natural language processing.

This work focuses on detecting and categorizing evasions of political questions as defined in SemEval-2026 Task 6 (Thomas et al., 2026). We follow the taxonomy and dataset introduced by Thomas et al. (2024). The task is structured as a two-level classification problem. At the first level, each answer is assigned a *clarity label*, indicating whether the response is a *Clear Reply*, a *Clear Non-Reply*, or an *Ambivalent* answer. At the second level, answers are further annotated with fine-grained evasion categories, corresponding to the specific techniques used by the speaker. These fine-grained labels are hierarchically dependent on the clarity level; for instance, *dodging* is only applicable to ambivalent replies. The entire two-level taxonomy tree of political answer types is presented in Figure 2.

We focus on the approaches described in Figure 1. Our code can be accessed on GitHub.[1]

[1] https://github.com/ma55530/FER-SemEval2026-T6

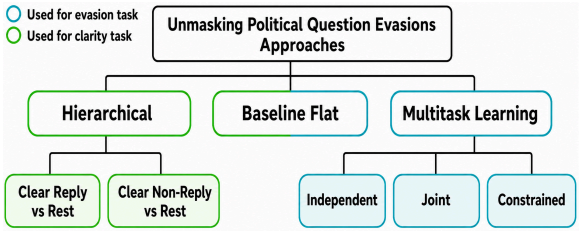


Figure 1: An overview of the approaches used for the *Clarity* and *Evasion* tasks.

2 Related Work

Political evasion detection has evolved from early social science typologies (Bull, 2003) into a well-defined computational task. Whereas prior work focused on speaker intent and subjective interpretations of equivocation, Thomas et al. (2024) formalize the problem as response clarity classification, introducing a two-level taxonomy and dataset of question–answer pairs from political interviews, and evaluating both fine-tuned models and large language model-based approaches. Related work has explored multimodal approaches to political discourse, such as argumentative fallacy detection in debates, where incorporating audio signals improves classification of fallacious arguments (Mancini et al., 2024).

Hierarchical text classification (HTC) leverages taxonomic structures to improve performance over flat baselines, typically through local cascades or global models that integrate label dependencies (Silla and Freitas, 2011; Dampa, 2025). Multitask learning (MTL) further enhances generalization by sharing representations across related tasks, a strategy applied in joint sentiment and tone analysis (Barić et al., 2023). However, joint training in low-data regimes can suffer from data sparsity and task interference (Hanneke and Xu, 2026), leading to logically incompatible label predictions. Recent work addresses these issues by explicitly modeling label dependencies or applying constrained inference to enforce hierarchical consistency (Wulamu et al., 2025).

3 Task Description

The hierarchical label structure (Fig. 2) poses several challenges, including strong class imbalance, semantic overlap between evasion categories, and ambiguity in borderline cases (Thomas et al., 2024).

Given the nature of the problem, the evaluation

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 30–36

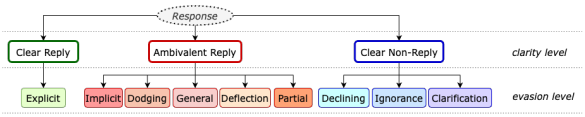


Figure 2: A two-level taxonomy of political answer types (Thomas et al., 2024): top-level clarity labels (*Clear Reply*, *Ambivalent*, *Clear Non-Reply*) and the corresponding fine-grained evasion categories.

Label	Count	%
Clarity Labels		
Clear Non-Reply	356	10.3%
Clear Reply	1,052	30.5%
Ambivalent	2,040	59.2%
Top Evasion Labels		
Explicit	1,052	30.5%
Dodging	706	20.5%
Implicit	488	14.2%

Table 1: Distribution of clarity labels and the three most frequent fine-grained evasion labels in the training set.

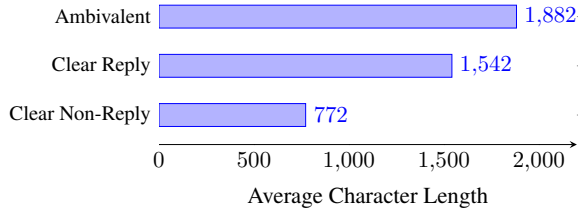


Figure 3: Average answer length (in characters) by clarity label, computed over the training set.

methodology requires specific attention. Unlike the training dataset, the test dataset contains three annotator labels. We follow the official task evaluation metric: if the model's prediction matches any of the three labels, we classify that prediction as correct. This applies to all metrics.

Each input sequence was formed as `Question:` Q `Answer:` A, where Q and A denote the interview question and answer, respectively.

Dataset Statistics. The dataset consists of 3,448 QA pairs with 4 distinct presidents, along with significant class imbalance (Table 1).

Text Length Distribution. The average answer length varies significantly across clarity labels. As shown in Figure 3, *Clear Non-Replies* are substantially shorter than other categories, a feature that likely aids the model in distinguishing this class.

4 Methodology and Results

This section presents the methodology and experimental results for each of the three modeling approaches explored in this work.

We begin with a flat baseline that treats clarity and evasion classification as independent standard multi-class problems, establishing a reference point for comparison.

We then describe two structured approaches: a hierarchical model that leverages the linguistic distinctiveness of clarity classes through two binary classifiers, and a multitask learning framework that jointly optimizes for both tasks using a shared encoder.

For each approach, we report accuracy, precision, recall, and macro F_1 averaged over five random seeds. Hyperparameter configurations for all models are provided in Appendix 4.

4.1 Baseline Flat Classification

As a simple baseline, we train two independent BERT-base classifiers (Devlin et al., 2019) to directly predict the clarity and evasion labels without exploiting the hierarchical structure of the taxonomy. In this setup, the task is treated as a standard multi-class text classification problem, where each answer is assigned exactly one category: either one of the three clarity levels or one of the nine evasion categories.

As shown in Table 2, the flat BERT classifier achieves a macro F_1 of 0.577 on the clarity task and 0.341 on the evasion task. These results establish a strong reference for the structured models described below. While accuracy on the clarity task reaches 0.630, fine-grained categories and semantic overlap make it harder for the evasion task, with accuracy of only 0.405. This flat classification approach serves as a baseline for comparison with the hierarchical and multitask models presented in the following sections.

4.2 Hierarchical Approach

To improve classification at the clarity level, we employed a hierarchical strategy and a more powerful BERT model – RoBERTa (Liu et al., 2019). First, we trained a binary model to distinguish one clarity class from the others. Then, a second model is trained to differentiate between the remaining two classes.

We consider two variants: one where a binary classifier first isolates the *Clear Reply* class, and another where it first isolates the *Clear Non-Reply* class. We hypothesized that isolating either the *Clear Replies* or *Clear Non-Replies* first would give better results, rather than separating the ambiguous *Ambivalent* class.

4.3 Clear Reply vs. Rest

In the first hierarchical variant, a BERT-base binary model separates *Clear Reply* examples from all others before a second model resolves the remaining two classes. As reported in Table 2, this pipeline achieves a macro F_1 of 0.556 on the clarity task – slightly below the flat baseline, suggesting that *Clear Reply* does not provide a sufficiently distinctive signal to benefit from early isolation.

4.4 Clear Non-Reply vs. Rest

In the second variant, the binary stage instead isolates *Clear Non-Reply* examples. This choice is motivated by the notably shorter average answer length of this class

Task	Approach	Backbone	Acc	Prec	Rec	Macro F_1 (Avg $\pm$ Std)
Clarity	Traditional ML (KNN)‡	TF-IDF features	0.6299	0.4708	0.4044	0.4154 ± 0.0000
	Flat baseline	BERT-base	0.6300	0.6450	0.5650	0.5770 ± 0.0075
	Hier. (Clear Reply vs. Rest)	BERT-base	0.7091	0.6093	0.5574	0.5564 ± 0.0255
	Hier. (Clear Reply vs. Rest)	RoBERTa-base	0.5883	0.5171	0.6198	0.5355 ± 0.0155
	Hier. (Clear Non-Reply vs. Rest)	BERT-base	0.7097	0.6276	0.5671	0.5766 ± 0.0111
	Hier. (Clear Non-Reply vs. Rest)†	**RoBERTa-base**	**0.7039**	**0.6216**	**0.6270**	$\mathbf{0.6029 \pm 0.0143}$
Evasion	Traditional ML (Log. Reg.)‡	TF-IDF features	0.3247	0.2618	0.2380	0.2200 ± 0.0000
	Flat baseline	BERT-base	0.4053	0.4274	0.3210	0.3406 ± 0.0128
	MTL with Independent Heads	BERT-base	0.4080	0.4266	0.5090	0.4241 ± 0.0216
	MTL Joint (unified label space)	BERT-base	0.3615	0.3577	0.4781	0.3732 ± 0.0124
	MTL Constrained (masked)	**BERT-base**	**0.4349**	**0.4746**	**0.5183**	$\mathbf{0.4380 \pm 0.0225}$

Table 2: Comprehensive comparison of all modeling approaches across the clarity and evasion tasks. Each row reports mean accuracy, precision, recall, and macro F_1 ($\pm$ standard deviation across five random seeds). Best result per task is in **bold**; † marks the overall best system. ‡Traditional ML results are deterministic (Std = 0); the best configuration per task is shown (no weights for clarity; with class weights for evasion).

(772 characters, versus 1,542 and 1,882 for *Clear Reply* and *Ambivalent*, respectively; see Figure 3), which reduces semantic ambiguity and makes the binary decision easier. With BERT-base we achieve F_1 = 0.577 (on par with the flat baseline). Replacing BERT with RoBERTa increases performance to F_1 = 0.603, our highest clarity score overall (Table 2). The gain over the *Clear Reply*-first variant confirms that early isolation of the most linguistically distinctive class is key to hierarchical performance.

4.5 Multitask Learning Approach

Following the success of multitask learning (MTL) frameworks in capturing cross-task correlations (e.g., Barić et al., 2023), we employ an MTL strategy to leverage commonalities across classification tasks, thereby improving robustness and reducing overfitting.

In our initial baseline setup, we employed a shared encoder to transform input text into latent embeddings, upon which we attached separate, independent task-specific heads for predicting clarity and evasion levels. Each head has its own output layer and loss function, while the encoder weights are shared across all tasks. The overall loss function is the weighted sum of the individual task losses:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{evasion}}\,\mathcal{L}_{\text{evasion}} + \lambda_{\text{clarity}}\,\mathcal{L}_{\text{clarity}}$$

where λ_{evasion} and λ_{clarity} are loss weights for each task.

However, we observed that this architecture frequently produced invalid label combinations (e.g., predicting a *Clear Reply* alongside a *Dodging* evasion technique), as the independent heads do not inherently respect the hierarchical constraints of the taxonomy. To address this structural incoherence, we extended our study to include two additional modeling strategies. Here, C denotes the *clarity label* and E denotes the *evasion label*:

- **Unified label space.** We replaced the independent heads with a single classification head that predicts the Cartesian product of all valid label pairs, explicitly modeling the joint distribution $P(C, E)$ and making illegal pairs impossible.
- **Constrained inference.** We retained the efficient independent training objective but augmented the model with a post-hoc validity mask during inference. This method computes the joint probability $P(C, E) = P(C) \cdot P(E)$ and forces structurally invalid pairs to zero before selecting the final prediction.

As shown in Table 2, the independent MTL baseline achieves an evasion F_1 of 0.424. The unified joint model drops to 0.373, suffering from data sparsity introduced by fragmenting labels into a large Cartesian product space. The constrained model not only eliminates all structurally invalid predictions but also achieves the highest evasion F_1 of 0.438, with accuracy 0.435, precision 0.475, and recall 0.518. This confirms that enforcing logical constraints during inference helps filter noise without incurring the sparsity cost of joint training.

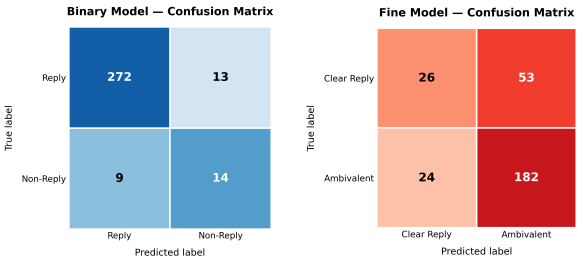
5 Error Analysis

We analyze errors along three axes: hierarchical error propagation, structural validity in multitask predictions, and task-specific cartographic difficulty (Swayamdipta et al., 2020). This separates failures caused by the label hierarchy from failures caused by pragmatic ambiguity.

5.1 Hierarchical Approach

The main weakness of the hierarchical pipeline is its tendency to propagate errors. Figure 4 illustrates this with the confusion matrices.

A *Clear Non-Reply* instance misclassified by the first binary model as *Rest* is routed to the fine-grained model, where it can no longer be correctly classified. Consequently, mistakes at the top level cascade downstream, limiting the overall accuracy and F_1 scores. If the binary model could be improved significantly, or by introduc-



(a) Binary stage (b) Fine-grained stage

Figure 4: Confusion matrices for the hierarchical pipeline on the RoBERTa model.

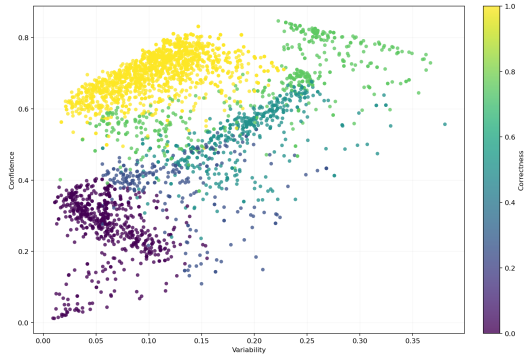


Figure 5: Dataset cartography map for the hierarchical pipeline (*Clear Non-Reply* vs. *Rest*, RoBERTa-base), aggregated across both stages.

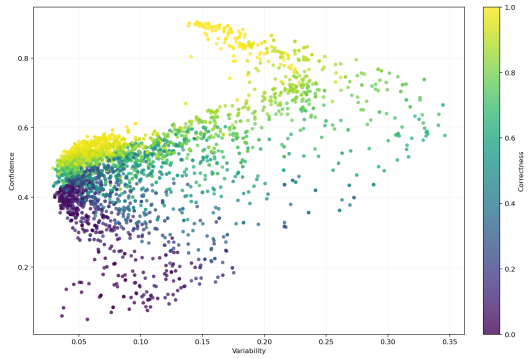


Figure 6: Task-specific cartography map for the clarity decision, calculated on the multitask model for better comparison between the two tasks, color denotes clarity correctness.

ing a fallback model, the propagation errors could be reduced.

Nine *Clear Non-Reply* examples are routed to *Rest* and therefore cannot recover the correct label, while thirteen examples predicted as *Clear Non-Reply* never enter the fine-grained model.

5.2 Multitask Approach

In our multitask analysis, we examined the clarity and evasion labels predicted by the independent-head model. We observed that this model exhibits a distinct failure mode: it often produces locally plausible but structurally invalid clarity-evasion pairs, such as *Clear Reply* with *Dodging*. To address this, we determined whether each predicted pair was legal under our taxonomy and compared these errors against the constrained decoder outputs. This evaluation allowed us to disentangle structural inconsistencies from ordinary semantic classification mistakes.

	Expl.	Deft.	Dodge	Gen.	Impl.	Part.	Ign.	Clar.	Decl.
Clear Reply	52	1	0	11	6	3	1	0	0
Ambivalent	2	10	19	102	44	9	2	0	2
Clear Non-Reply	0	0	2	0	0	2	17	4	19

Figure 7: Predicted clarity–evasion label pairs for the independent multitask model in one test run. Red cells mark structurally invalid combinations, while the numbers show the test examples assigned to each predicted pair in a single run.

Figure 7 quantifies the issue: 32 of 308 predictions (10.4%) are illegal in the run shown. The constrained decoder removes these invalid pairs across all five seeds, but only at the structural level; remaining errors still depend on distinguishing fine-grained evasion strategies.

Most illegal predictions are not random combinations: 22 of the 32 involve *Clear Reply* paired with a non-*Explicit* evasion label. The largest single illegal cell is *Clear Reply* + *General* (11 cases), followed by *Clear Reply* + *Implicit* (6 cases). This suggests that the clarity head often recognizes answer-like surface form, while the evasion head still reacts to topic shifts or generalized political language. Constrained decoding therefore improves ontology consistency, but the semantic conflict remains.

5.3 Dataset Cartography

We use dataset cartography (Swayamdipta et al., 2020) to aggregate confidence, variability, correctness, and forgetfulness across epochs and seeds for the constrained multitask BERT-base model. For the task-specific maps, we recompute the same metrics separately from the clarity and evasion probabilities produced by the multitask model. Figure 5 additionally shows RoBERTa hierarchical clarity cartography for comparison. We use a natural-break split in which high-variability examples are treated as *Ambiguous*; low-variability examples are then separated into *Easy-to-learn* and *Hard-to-learn* by confidence. This yields 1,345 easy examples (39.0%), 818 ambiguous examples (23.7%), and 1,285 hard examples (37.3%). Table 3 summarizes the qualitative motifs behind these regions, while Figures 6 and 8 show the corresponding task-specific cartographies.

These categories reflect how the model behaved during training (based on confidence and variability), not just how common the labels are. Easy examples are not only more frequent; they are examples where surface form and pragmatic function agree. Ambiguous examples often contain competing evidence across the question and answer, so the model can be correct in some epochs and wrong in others. Hard examples form the most important diagnostic group: they are frequently learned with low confidence and low variability, which indicates systematic misunderstanding rather than mere seed instability.

Figure 5 shows how a hierarchical model produces

Region	Motive	Evidence	Examples	Effect
Easy	Direct or explicit answers dominate.	The majority of inspected examples were direct or explicit answers.	*Q:* Would you encourage Congress to pay for school testing? *A:* I would if they want. *Q:* Why did you come back, Mr. President? *A:* I did not even think about not coming back.	The model learns stable lexical and discourse cues when the answer gives a clear commitment or a conventional explicit response.
Ambig.	The text often supports more than one reading.	Inspected examples often combined direct-looking answers, multi-question context, or non-replies with answer-like content.	*Q:* Did you discuss the election with her? *A:* No, I did not. I did not. *Q:* What exactly is un-American about those programs? *A:* We are going to do a report; those people are gone.	Ambiguity usually comes from competing targets: a reply can answer one part of the prompt while evading another, or a refusal can be followed by substantive policy material.
Hard	The label requires pragmatic inference beyond surface form.	Most inspected hard cases required implicit rhetorical inference or exposed model reliance on surface cues.	*Q:* Is the peace process over? *A:* Who? *Q:* Can the President go around Congress? *A:* You ever hear the word obstruction?	Hard examples are not merely noisy; many require identifying whether a short reply, counterquestion, topic shift, or long answer functions as a specific evasion strategy.

Table 3: Qualitative explanation of why examples fall into each cartography region. Examples are cropped to the minimal span needed to show the relevant behavior.

a diffuse distribution with high variance of confidence and variability. Rather than sharply separating *Easy-to-learn* and *Hard-to-learn* instances, the hierarchical model produces a broad band of confidence predictions.

The task-specific maps (Fig. 6 and 8) reveal distinct difficulty profiles for clarity and evasion. Clarity is mostly learnable: 2,050 examples fall in the easy region (59.5%), while only 520 are hard (15.1%). Evasion is substantially harder, with 1,548 hard examples (44.9%) and only 642 examples in the ambiguous region (18.6%), showing that the model struggles much more with the specific evasion nuances than the basic clarity task.

The category-level distributions in Figure 9 support the same conclusion. *Dodging* is the most concentrated hard label: 64.4% of its examples are hard under evasion-only cartography. *Implicit* is also difficult, with 57.0% in the hard region. Labels with overt cues, such as *Clarification*, *Declining to answer*, and *Claims ignorance*, move more often to the ambiguous region, suggesting unstable confidence rather than uniformly low learnability.

In general, structural constraints prevent invalid output, but the remaining largest errors are pragmatic.

By comparing the HTC and MTL cartography maps on the clarity task (Fig. 5 & 6) with the results in Table 2, we observe how HTC produces a more polarized map, with clearly clustered *Easy-to-learn* and *Hard-to-learn* examples, while MTL appears more ambiguous and noisy, forming a large central cluster where confidence is approximately 0.5 and variability remains low. This can be explained by the hierarchical model's two-stage decision process, which enforces sharper separations between classes, whereas the multitask setup, due to its shared representation and greater flexibility, blends decision boundaries and results in less distinct confidence patterns. As a result, HTC exhibits clearer class separation and interpretability, while MTL trades this structure for improved flexibility on harder, fine-grained cases.

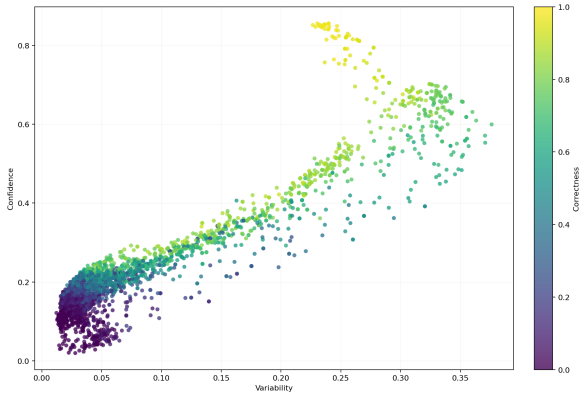


Figure 8: Task-specific cartography map for the evasion decision, calculated on the multitask model. The larger hard-to-learn region indicates that fine-grained pragmatic labels are the main source of remaining errors.

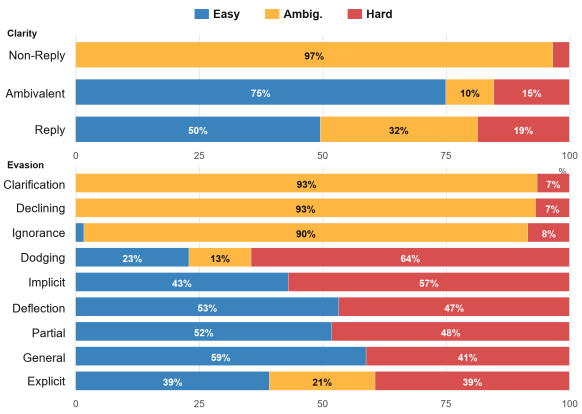


Figure 9: MTL distribution of clarity (top) and evasion (bottom) labels across cartography regions.

6 Conclusion

In this work, we demonstrate that the inherent hierarchical structure of political question evasion can be effectively exploited to improve classification performance. Our findings show that the linguistic distinctiveness of

the *Clear Non-Reply* class makes it a prime candidate for early isolation in cascaded pipelines, leading to our best macro F_1 of 0.603.

Furthermore, we addressed the prevalent issue of structural inconsistency in multitask learning by implementing a constrained inference mask that ensures logical alignment with the taxonomy.

While dataset cartography revealed significant semantic overlap in the *Ambivalent* category, it also highlighted a more distinct and interpretable representation of uncertainty in our hierarchical approach.

Limitations

The QEvasion dataset's focus on four U.S. presidents in English restricts the generalisation of the findings to diverse political systems, other languages, or different discourse formats like parliamentary debates.

Significant class imbalance, where the Clear Non-Reply class comprises only 10.3% of the data, further limits accuracy on classifying minority labels and increases evaluation variance. Our hierarchical pipeline is vulnerable to top-level error propagation, where initial clarity misclassifications irrevocably affect downstream evasion labeling.

Additionally, our multitask experiments were restricted to standard architectures, omitting advanced structural learning methods such as label-aware attention or graph-based decoding.

The official "match-any" evaluation metric may provide an optimistic estimate of performance compared to stricter agreement-based measures.

Bound by Google Colab's compute limits, this study used BERT-base encoders; however, future research should scale these efforts by adopting larger LLMs and more advanced ensemble architectures.

Acknowledgements

We'd like to thank TakeLab for their mentorship, constructive feedback, and support throughout the development of this work.

References

Ana Barić, Laura Majer, David Dukić, Marijana Grbeša-zenzerović, and Jan Snajder. 2023. Target two birds with one SToNe: Entity-level sentiment and tone analysis in Croatian news headlines. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 78–85, Dubrovnik, Croatia. Association for Computational Linguistics.

Peter Bull. 2003. The microanalysis of political communication: Claptrap and ambiguity. *The Microanalysis of Political Communication: Claptrap and Ambiguity*, pages 1–220.

Artemis Dampa. 2025. Investigating hierarchical structure in multi-label document classification. In *Proceedings of the 9th Student Research Workshop associated with RANLP 2025*, pages 10–19.

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

Steve Hanneke and Mingyue Xu. 2026. When more data doesn't help: Limits of adaptation in multitask learning. *arXiv preprint arXiv:2601.20774*.

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.

Eleonora Mancini, Federico Ruggeri, and Paolo Torroni. 2024. Multimodal fallacy classification in political debates. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 170–178, St. Julian's, Malta. Association for Computational Linguistics.

Carlos N Silla and Alex A Freitas. 2011. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22(1-2):31–72.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. *Preprint*, arXiv:2009.10795.

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.

Aziguli Wulamu, Lyu Zhengyu, Kaiyuan Gong, Yu Han, Zewen Wang, Zhihong Zhu, and Bowen Xing. 2025. HTML: Hierarchical topology multi-task learning for semantic parsing in knowledge base question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9307–9321, Vienna, Austria. Association for Computational Linguistics.

Component	BERT-base	MTL (DeBERTa-v3)
Model	BERT-base	DeBERTa-v3-base
Optimizer	AdamW	AdamW
Learning Rate	2×10^{-5}	1.5×10^{-5}
Epochs	2	8
Batch Size	16	8 (accum. = 4)
Weight Decay	0.0	0.01
Loss	Cross-entropy	$\mathcal{L}_{\text{evasion}} + 0.5 \cdot \mathcal{L}_{\text{clarity}}$
Class Weighting	None	Inverse-frequency balancing
Seeds	42, 43, 44, 45, 46	42, 123, 456, 789, 1337

Table 4: Baseline flat classification and multi-task learning configurations.

Component	Clear Reply vs Rest	Clear Non-Reply vs Rest
Model	RoBERTa-base	RoBERTa-base
Optimizer	AdamW	AdamW
Binary Learning Rate	2×10^{-5}	1×10^{-5}
Fine Learning Rate	2×10^{-5}	1×10^{-5}
Epochs	3	4
Batch Size	16	16
Weight Decay	0.0	0.02
Loss	Cross-entropy	Cross-entropy
Seeds	42, 43, 44, 45, 46	Random seeds

Table 5: Hierarchical RoBERTa configurations using two-stage binary and fine-grained classification.

A Hyperparameters

Tables 4 & 5 show the hyperparameters used in experimental setups.

B Traditional Machine Learning Baselines

We evaluated a set of traditional machine learning models that do not rely on transformer-based architectures. These models serve as a performance reference point before the transformer-based approaches described in the main paper. Full results for the clarity task (Table 6) and the evasion task (Table 7) are provided below; the best configuration per task is also included in the overall comparison in Table 2.

Scenario	Model	Avg Acc.	Avg Prec.	Avg Rec.	$F_1 \pm$ Std
No Weights	Nearest Neighbors	0.6299	0.4708	0.4044	**0.4154 ± 0.0000**
No Weights	Linear SVM	0.6688	0.2229	0.3333	0.2672 ± 0.0000
No Weights	RBF SVM	0.6688	0.2229	0.3333	0.2672 ± 0.0000
No Weights	Decision Tree	0.6747	0.5322	0.3698	0.3386 ± 0.0004
No Weights	Random Forest	0.6688	0.2229	0.3333	0.2672 ± 0.0000
No Weights	Neural Net	0.6630	0.6963	0.3667	0.3496 ± 0.0111
No Weights	AdaBoost	0.6526	0.6393	0.3459	0.3119 ± 0.0000
No Weights	Naive Bayes	0.6688	0.2229	0.3333	0.2672 ± 0.0000
No Weights	Logistic Regression	0.6721	0.7118	0.3841	0.3795 ± 0.0000
With Weights	Nearest Neighbors	0.6299	0.4708	0.4044	**0.4154 ± 0.0000**
With Weights	Linear SVM	0.5812	0.2953	0.4570	0.3405 ± 0.0000
With Weights	RBF SVM	0.5779	0.6620	0.3685	0.3682 ± 0.0000
With Weights	Decision Tree	0.5844	0.3505	0.3892	0.3367 ± 0.0000
With Weights	Random Forest	0.2234	0.3464	0.3819	0.1944 ± 0.0432
With Weights	Neural Net	0.6656	0.6918	0.3638	0.3431 ± 0.0078
With Weights	AdaBoost	0.6526	0.6393	0.3459	0.3119 ± 0.0000
With Weights	Naive Bayes	0.6364	0.4435	0.4023	0.4008 ± 0.0000
With Weights	Logistic Regression	0.4968	0.3919	0.4001	0.3919 ± 0.0000

Table 6: Performance of traditional machine learning models on the clarity classification task, with and without class-weights. Best F_1 in each scenario is in **bold**.

C Baseline Flat Classification

We additionally analyze the behavior of the flat BERT baseline using Dataset Cartography.

The flat BERT baseline's cartography (Fig 10 & 11) reveals the Ambivalent samples are predominantly *Easy-to-learn* (75%), perhaps largely due to their 59.2% prevalence. In contrast, *Clear Non-Replies* are almost completely *Hard-to-learn*, making up only a tenth of the dataset. This suggests that without hierarchical cues, the model struggles to consistently identify minority classes despite their linguistic and semantic distinctiveness.

Scenario	Model	Avg Acc.	Avg Prec.	Avg Rec.	$F_1 \pm$ Std
No Weights	Nearest Neighbors	0.3117	0.2280	0.1522	**0.1665 ± 0.0000**
No Weights	Linear SVM	0.3734	0.0415	0.1111	0.0604 ± 0.0000
No Weights	RBF SVM	0.3701	0.0816	0.1142	0.0701 ± 0.0000
No Weights	Decision Tree	0.3831	0.1207	0.1187	0.0823 ± 0.0000
No Weights	Random Forest	0.3714	0.0562	0.1107	0.0612 ± 0.0025
No Weights	Neural Net	0.3747	0.1295	0.1191	0.0906 ± 0.0066
No Weights	AdaBoost	0.3636	0.1883	0.1191	0.0894 ± 0.0000
No Weights	Naive Bayes	0.3734	0.0415	0.1111	0.0604 ± 0.0000
No Weights	Logistic Regression	0.3636	0.2365	0.1305	0.1203 ± 0.0000
With Weights	Nearest Neighbors	0.3117	0.2280	0.1522	0.1665 ± 0.0000
With Weights	Linear SVM	0.1591	0.0263	0.1240	0.0434 ± 0.0000
With Weights	RBF SVM	0.3636	0.1578	0.1367	0.1229 ± 0.0000
With Weights	Decision Tree	0.3636	0.1279	0.2328	0.1235 ± 0.0000
With Weights	Random Forest	0.3045	0.1741	0.1406	0.0904 ± 0.0100
With Weights	Neural Net	0.3779	0.1371	0.1202	0.0903 ± 0.0067
With Weights	AdaBoost	0.3630	0.1884	0.1191	0.0895 ± 0.0003
With Weights	Naive Bayes	0.2825	0.1806	0.1646	0.1425 ± 0.0000
With Weights	Logistic Regression	0.3247	0.2618	0.2380	**0.2200 ± 0.0000**

Table 7: Performance of traditional machine learning models on the fine-grained evasion classification task, with and without class-weights. Best F_1 in each scenario is in **bold**.

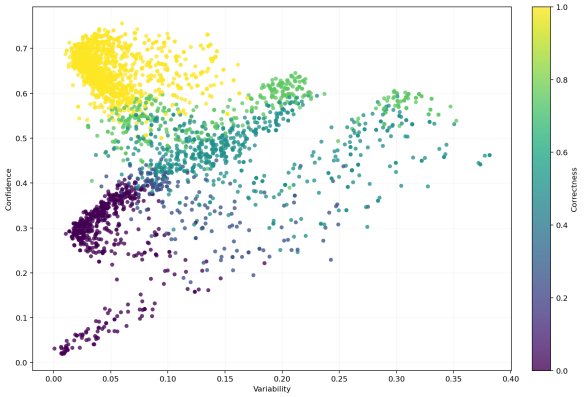


Figure 10: Dataset cartography map for the baseline model.

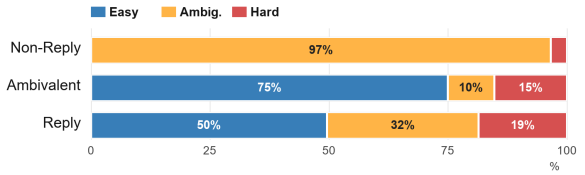


Figure 11: Flat baseline distribution of clarity labels across cartography regions.

D Hierarchical Approach

Figure 12. illustrates the distribution of clarity labels for the HTC model.

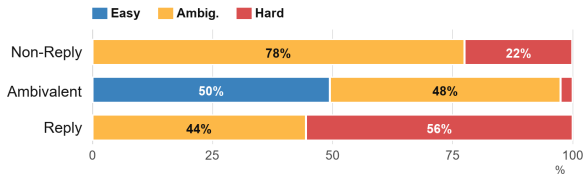


Figure 12: HTC distribution of clarity labels across cartography regions.