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

Political accountability relies heavily on the ability of journalists and the public to elicit clear answers from leaders, yet evasion remains a pervasive strategy in political discourse that obscures true intentions. This paper addresses the automatic detection of such evasions within the context of SemEval 2026 Task 6. We explore the hierarchical nature of political answers, where responses are first categorized by clarity (e.g., Ambivalent, Clear Reply) and subsequently by specific evasion techniques (e.g., Dodging, Deflection). We compare three modeling strategies: a flat baseline, a hierarchical cascade, and a multitask learning approach. Our experiments demonstrate that a hierarchical RoBERTa-based model achieves the best performance (macro F1 $\approx$ 0.60), particularly by leveraging the structural distinctiveness of "Clear Non-Replies". Conversely, we find that standard multitask learning frequently produces structurally invalid label combinations in approximately 25% of predictions. We demonstrate that applying a constrained inference mask eliminates these errors entirely while improving F1 performance, whereas a fully joint training approach falters due to data sparsity. Finally, we employ Dataset Cartography to diagnose training dynamics, revealing that while hierarchical models suffer from error propagation, they offer a more calibrated representation of uncertainty compared to flat classifiers.

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, as well as natural language processing (NLP).

This work focuses on detecting and categorizing evasions of political questions as defined in SemEval-2026 Task 6 (Thomas et al.). We follow the taxonomy proposed by Thomas et al. (Thomas et al., 2024a), which forms the basis of this shared task. 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.

This hierarchical label structure poses several challenges, including strong class imbalance, semantic overlap between evasion categories, and ambiguity in borderline cases(Thomas et al., 2024b). To address these challenges, we explore two approaches that explicitly leverage the task structure: a hierarchical classification approach and a multitask learning approach.

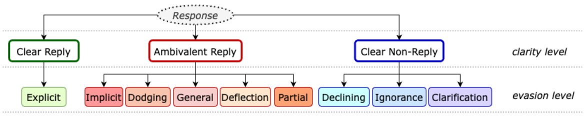

Figure 1: Taxonomy tree

2 Data Analysis and Evaluation

We use the dataset provided for SemEval-2026 Task 6 (Thomas et al.), which consists of political interviews in English language and builds upon the taxonomy introduced by Thomas et al. (Thomas et al., 2024a). Alongside LLM prompting, they also compared encoder models. Given the nature of the problem, the evaluation 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.

2.1 Dataset Statistics

We summarize the core dataset metrics and label distributions in Table 1 and Table 2.

Metric	**Value**
Total QA pairs	3,448
Unique URLs	287
Distinct Presidents	4
Multiple Questions	2.5%
Affirmative Questions	22.4%
Inaudible	1.3%

Table 1: Basic dataset statistics and boolean flags.

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

Table 2: Distribution of clarity and top evasion labels.

2.2 Text Length Distribution

The average answer length varies significantly across clarity labels. As shown in Figure 2, *Clear Non-Replies* are substantially shorter than other categories, a feature that likely aids the model in distinguishing this class.

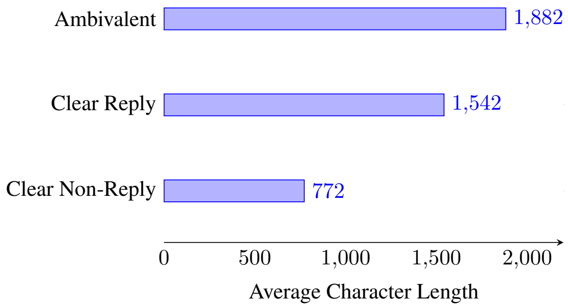

Figure 2: Average answer length by clarity label.

In our upcoming evaluations, we measured the model's performance using accuracy, precision, recall, and macro F1 score. For each metric, we calculated the minimum and maximum values, as well as the mean and its standard deviation.

3 Traditional Machine Learning Baselines

In this section, we evaluate a set of traditional machine learning models that do not rely on transformer-based architectures.

These models, including Logistic Regression, Support Vector Machines (SVM), Random Forests, k-Nearest Neighbors, and Neural Networks, serve as classical baselines for the clarity classification task. The goal is to establish a performance reference point before introducing more advanced transformer-based approaches in the following sections.

Scenario	Model	Avg Acc.	Avg Prec.	Avg Rec.	F1 ± 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 3: Performance of traditional machine learning models on the clarity classification task. The best results are in **bold**.

Scenario	Model	Avg Acc.	Avg Prec.	Avg Rec.	F1 ± 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 4: Performance of traditional machine learning models on the evasion classification task. The best results are in **bold**.

4 Baseline Flat Classification

As a simple baseline, we train two independent BERT-based classifiers 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.

This flat classification approach serves as a baseline for comparison with the hierarchical and multitask models presented in the following sections.

4.1 Clarity Classification

Metric	Min	Max	Avg ± Std
F1 Score	0.5677	0.5850	0.5770 ± 0.0075
Accuracy	0.6150	0.6550	0.6300 ± 0.0150
Precision	0.6250	0.6650	0.6450 ± 0.0140
Recall	0.5550	0.5800	0.5650 ± 0.0100

Table 5: Base clarity metrics

4.2 Evasion Classification

Metric	Min	Max	Avg ± Std
F1 Score	0.3200	0.3550	0.3406 ± 0.0128
Accuracy	0.3900	0.4221	0.4053 ± 0.0135
Precision	0.3960	0.4520	0.4274 ± 0.0255
Recall	0.3040	0.3360	0.3210 ± 0.0142

Table 6: Base evaluation metrics

5 Hierarchical Approach

To improve classification at the clarity level, we employed a hierarchical strategy. 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 applied this approach in two ways: first, by distinguishing the *Clear Reply* class from all others, and second, by distinguishing the *Clear Non-Reply* class from all others. We hypothesized that isolating either the *Clear Replies* or *Clear Non-Replies* first, would give better results, rather than separating the ambiguous *Ambivalent* class.

5.1 Clear Reply vs. Rest

Metric	Min	Max	Avg ± Std
F1 Score	0.5081	0.5784	0.5564 ± 0.0255
Accuracy	0.6948	0.7208	0.7091 ± 0.0089
Precision	0.5934	0.6248	0.6093 ± 0.0126
Recall	0.5243	0.5913	0.5574 ± 0.0234

Table 7: Clear Reply vs. Rest metrics (BERT-base).

5.2 Clear Non-Reply vs. Rest

Metric	Min	Max	Avg ± Std
F1 Score	0.5660	0.5972	0.5766 ± 0.0111
Accuracy	0.7013	0.7208	0.7097 ± 0.0078
Precision	0.6052	0.6472	0.6276 ± 0.0150
Recall	0.5459	0.5886	0.5671 ± 0.0155

Table 8: Clear Non-Reply vs. Rest metrics (BERT-base).

Metric	Min	Max	Avg ± Std
F1 Score	0.5859	0.6189	0.6029 ± 0.0143
Accuracy	0.6818	0.7370	0.7039 ± 0.0215
Precision	0.5768	0.7012	0.6216 ± 0.0455
Recall	0.5945	0.6844	0.6270 ± 0.0308

Table 9: Clear Non-Reply vs. Rest metrics (RoBERTa-base).

We observe that the second approach achieved better scores, likely because the average length of *Clear Non-Reply* answers was considerably shorter than that of the other classes.

6 Multitask Learning Approach

We also explored a multitask learning (MTL) approach, where a single BERT model is trained to perform multiple related tasks simultaneously. The main idea behind MTL is that by sharing representations between tasks, the model can leverage common patterns in the data, leading to better generalization and reduced overfitting. This approach has previously been applied to structurally related NLP tasks, such as jointly modeling sentiment and tone, where shared representations were shown to improve performance (Barić et al., 2023).

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.

Further, we can define our overall loss function as the sum of the individual task losses, weighted according to the relative importance of each task:

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

where λ_{evasion} and λ_{clarity} are loss weights for the total loss of each task. This allows the model to learn from multiple signals simultaneously.

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 and enforce valid predictions, we extended our study to include two additional modeling strategies:

- Unified label space: In this approach, we replaced the independent heads with a single classification head that predicts the cartesian product of all valid label pairs. This approach explicitly models the joint distribution $P(C, E)$, making the generation of illegal pairs impossible.
- Constrained inference: Alternatively, we retained the efficient independent training objective but augmented the model with a post-hoc validity mask

during inference. This method calculates the joint probability $P(C, E) = P(C) \times P(E)$ and eliminates structurally invalid pairs by forcing their probabilities to zero before selection.

By comparing these three architectures, we aimed to determine the optimal balance between structural validity and predictive performance.

6.1 Multitask Evasion Results

The comparison of F1 scores across the three multitask strategies is presented in Table 10. While the joint model successfully eliminated illegal predictions, it suffered from data sparsity due to the fragmentation of labels, resulting in a lower F1 score. Conversely, the constrained model not only ensured 100% structural validity but also achieved the highest predictive performance, outperforming both the independent baseline and the joint model. This confirms that enforcing logical constraints helps filter out noise and enhances the model's reliability.

Model	Avg F1 ± Std
Independent (Baseline)	0.4241 ± 0.0216
Joint (Dependent)	0.3732 ± 0.0124
Constrained (Masked)	**0.4380** ± **0.0225**

Table 10: Comparison of multitask strategies (evasion task F1 score)

Detailed performance metrics for the optimal constrained approach are provided in Table 11.

Metric	Min	Max	Avg ± Std
F1 Score	0.4166	0.4617	0.4380 ± 0.0225
Accuracy	0.4111	0.4533	0.4349 ± 0.0216
Precision	0.4332	0.5378	0.4746 ± 0.0556
Recall	0.4872	0.5699	0.5183 ± 0.0450

Table 11: Metrics for constrained multitask model

7 Overall Results

Method	Task	Min	Max	F1 (Avg ± Std)
Baseline Flat	Clarity	0.5677	0.5850	0.5770 ± 0.0075
Hierarchical	Clear Reply vs Rest	0.5081	0.5784	0.5564 ± 0.0255
Hierarchical	Clear Non-Reply vs Rest (BERT)	0.5660	0.5972	0.5766 ± 0.0111
Hierarchical	Clear Non-Reply vs Rest (RoBERTa)	0.5859	0.6189	**0.6029** ± **0.0143**
Baseline Flat	Evasion	0.3200	0.3550	0.3406 ± 0.0128
Multitask	Evasion	0.4166	0.4617	**0.4380** ± **0.0225**

Table 12: Overall comparison of F1 scores across different modeling approaches. The best result is in **bold**.

8 Error Analysis

To better understand the performance of our models, we employed *Dataset Cartography* (Swayamdipta et al., 2020), a method for diagnosing datasets based on training dynamics. This approach tracks the confidence, variability, and correctness of model predictions across training epochs, allowing us to categorize samples as *Easy-to-learn*, *ambiguous*, or *Hard-to-learn*. By visualizing these categories, we can identify regions in the dataset where the model struggles and understand which examples contribute most to errors or uncertainty.

It is important to note several limitations of our study. The dataset is relatively small, consisting of interviews with only four presidents, and it originates from a domain-specific context, which may introduce model biases and limit generalization.

Based on the dataset cartography projection, three characteristic regions can be clearly identified: *Easy-to-Learn* samples are associated with high confidence and low variability, indicating that the model consistently predicts them correctly. *Hard-to-Learn* samples exhibit low confidence and low variability, suggesting systematic misclassification. Finally, *Ambiguous* samples are characterized by higher variability and moderate to high confidence, reflecting inherent ambiguity or semantic overlap between classes.

8.1 Bert baseline

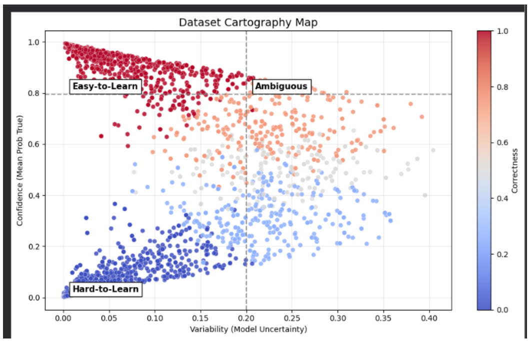

Figure 3: BERT baseline dataset cartography map

8.2 Hierarchical approach

The hierarchical approach generally improved classification for the *Clear Non-Reply* class, likely due to the shorter average answer length, which reduces semantic ambiguity.

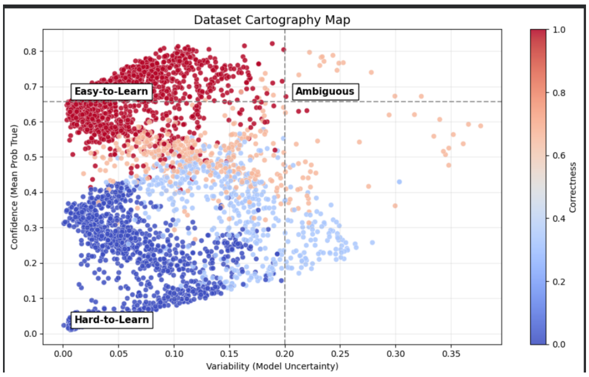

Figure 4: Hierarchical dataset cartography map.

Comparing the hierarchical and BERT baseline cartography maps reveals distinct training dynamics and differences. Firstly, the BERT baseline map resembles a "V" shape, splitting the dataset distinctly between two groups: what's hard to learn and what's easy to learn where the variability is low, with few outliers in the middle. The model is either pretty confident or not confident at all. As the variability increases, polarizations of confidence disappear.

In contrast, the hierarchical approach does not exhibit the characteristic "V" shape. Instead, the cartography map reveals a more diffuse and continuous distribution of samples across confidence and variability. Rather than sharply separating *Easy-to-learn* and *Hard-to-learn* instances, the hierarchical model produces a broader band of confidence predictions with a slightly decreased variability compared to the BERT baseline map. This effect can be attributed to error propagation in the hierarchical pipeline. A *Clear Non-Reply* instance that is misclassified by the first binary model as "Rest" is routed to the fine-grained model, where it can no longer be correctly labeled. Similarly, *Ambivalent* and *Clear Reply* examples misrouted at the binary level suffer from the same issue. Consequently, mistakes at the top level cascade downstream, limiting the overall accuracy and F1 scores of the hierarchical approach. This phenomenon highlights a trade-off inherent to hierarchical classification: while it enforces structural label constraints, it is sensitive to errors in early-stage predictions. If the binary model could be improved significantly or by introducing a *fallback model*, the propagation errors could be reduced significantly.

Overall, the absence of a "V" shape in the hierarchical cartography highlights a key trade-off: while the hierarchical model sacrifices some decisiveness, it gains a more calibrated representation of uncertainty, particularly for semantically overlapping or structurally constrained labels.

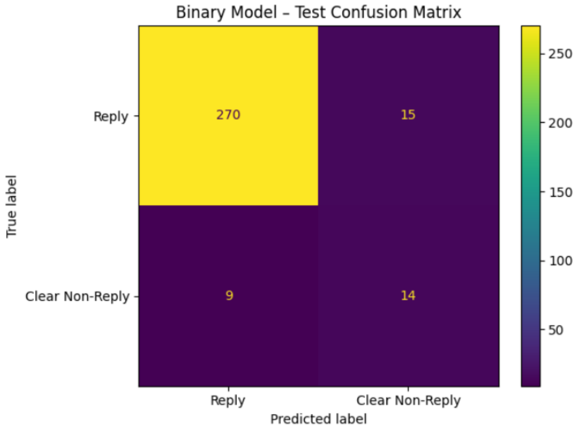

Figure 5: Confusion matrix for hierarchical binary model.

The confusion matrix in Figure 5 highlights significant class imbalance and several misclassifications. Nine examples were incorrectly classified as *Reply (Rest)* at the binary level, which are then routed to the fine-grained model and, as discussed above, are misclassified again. Additionally, fifteen examples that were predicted as *Clear Non-Reply* never enter the fine-grained model, effectively bypassing the downstream correction step. This illustrates how errors at the top level of the hierarchical pipeline can propagate impacting overall classification performance.

The confusion matrix in Figure 6 shows the predictions of the fine-grained model.

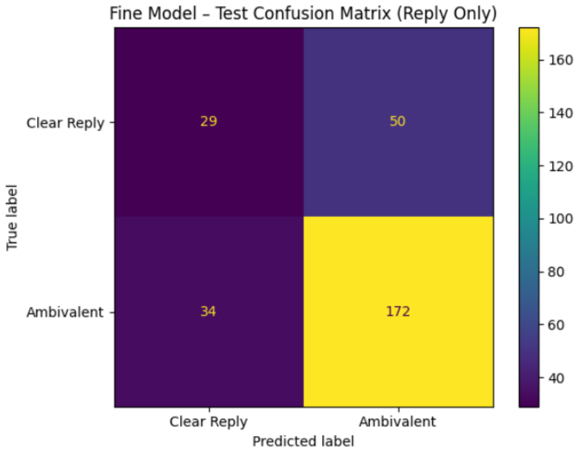

Figure 6: Confusion matrix for hierarchical fine-grained model.

8.3 Multitask approach

Multitask models exhibited distinct training dynamics compared to the hierarchical approach. While the initial motivation was to leverage shared representations to improve generalization, our experiments revealed that the standard architecture, utilizing independent prediction heads, suffered from structural inconsistencies.

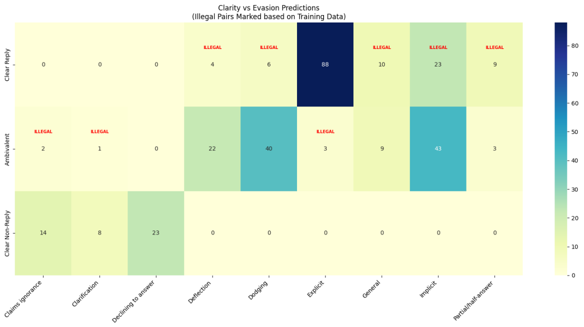

Figure 7: Predictions of clarity and evasion labels showing structurally invalid pairs (marked in red).

As shown in Figure 7, the independent multitask baseline frequently produces structurally invalid label combinations. We quantified this error rate and found that approximately 26.8% of test instances resulted in "illegal" pairs (e.g., predicting *Clear Reply* alongside a *Dodging* technique). This behavior arises because the independent heads optimize separate loss functions for clarity and evasion. Consequently, the model learns independent probabilities $P(\text{clarity}|x)$ and $P(\text{evasion}|x)$, but fails to model their joint distribution $P(\text{clarity}, \text{evasion}|x)$.

To diagnose the learning dynamics, we analyzed the Dataset Cartography map for the multitask model (Figure 8).

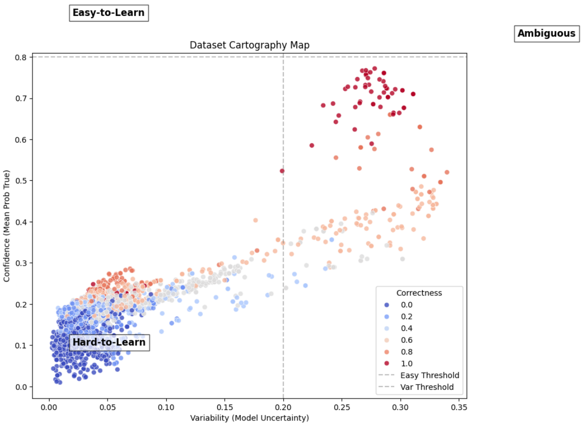

Figure 8: Multitask (independent baseline with two separate heads) dataset cartography map.

The cartography reveals that a substantial proportion of samples fall into the *Hard-to-learn* region with low confidence and low variability (Swayamdipta et al., 2020). This indicates that despite shared representations, the model struggles to distinguish between overlapping linguistic cues in the *Ambivalent* and borderline categories.

These diagnostics explain the performance trade-off observed in Section 6. The standard independent model fails due to the structural invalidity shown in Figure 7. While joint training methods are often recommended for related tasks (Zhang et al., 2023), we found that the joint model fails in our case due to the data sparsity visible in the "Hard-to-Learn" regions of Figure 8. This observation aligns with the findings of Hanneke and Xu (Hanneke and Xu, 2026), who show that more data does not necessarily improve model adaptation in multitask settings. The constrained approach succeeds because it bridges this gap: it allows the model to learn efficient independent representations, avoiding sparsity while enforcing the necessary structural logic during inference.

8.4 Key Observations

- Ambivalent and Clear-Reply answers are the most challenging, due to subtle linguistic cues and semantic overlap.
- Hierarchical modeling benefits classes with clear structural distinctions, particularly *Clear Non-Reply*.
- Multitask learning with independent heads fails to capture logical label dependencies.
- Multitask shares representations effectively but is sensitive to rare labels and ambiguous examples.
- Weighted loss functions help partially, but extreme class imbalance remains a challenge.

9 Conclusion

In this work, we explored multiple approaches for the detection and classification of evasive political answers, focusing on the hierarchical structure of clarity and evasion labels.

Our experiments show that traditional machine learning models provide a reasonable baseline, but transformer-based approaches, especially hierarchical classification with RoBERTa, significantly outperform classical methods. The hierarchical strategy is particularly effective for the *Clear Non-Reply* class, likely due to reduced text length and clearer linguistic patterns.

Multitask learning offers the advantage of shared representations across related tasks, but performance is limited by class imbalance and semantic overlap, particularly in the *ambivalent* category.

Error analysis highlights the challenges inherent in this dataset: subtle evasions, overlapping semantics, and ambiguous examples reduce model confidence and F1 scores. Future work could explore building more complex ensembles that combine models with different architectures, allowing complementary strengths to compensate for different kinds of errors and therefore improving overall performance and robustness. It may also be valuable to investigate additional data augmentation techniques or to incorporate discourse-level features, which could further enhance the detection of evasive answers.

Overall, our study demonstrates that leveraging the hierarchical structure of political question evasion can yield measurable gains, and that careful modeling of class imbalance and ambiguity is crucial for reliable automatic detection.

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.

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

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. 2024a. "i never said that": A dataset, taxonomy and baselines on response clarity classification. *Preprint*, arXiv:2409.13879.

Konstantinos Thomas, Giorgos Filandrianos, Maria Lymperaiou, Chrysoula Zerva, and Giorgos Stamou. 2024b. "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 and 1 others. SemEval-2026 task 6: CLARITY: Unmasking political question evasions.

Zhihan Zhang, Wenhao Yu, Mengxia Yu, Zhichun Guo, and Meng Jiang. 2023. A survey of multi-task learning in natural language processing: Regarding task relatedness and training methods. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 943–956, Dubrovnik, Croatia. Association for Computational Linguistics.