Report

for Semeval Task 6.1

Team members:

- Toma Oprea Lacrimioara
- Oprea Stefan

State of the art (generated by AI + manually)

1. Description of the Problem

1.1 Context and Motivation

In democratic systems, political accountability relies on the ability of journalists and the public to elicit clear information from elected officials through Question-Answering (Q&A) sessions. However, in high-stakes environments like presidential debates or press conferences, politicians often employ equivocation—a strategic communication style characterized by being intentionally non-straightforward or ambiguous.

The CLARITY task (SemEval-2026 Task 6) addresses the computational gap in detecting these dodges. While evasion has been studied for decades in political science, it remains a "frontier" problem for Natural Language Processing (NLP) due to the subtle, high-level semantic reasoning required to distinguish between a "lengthy but informative" reply and a "lengthy but evasive" one.

1.2 Task Definition

The core problem is defined as a hierarchical classification of response quality given a specific question context.

Clarity-Level Classification (Level 1): A three-way classification task where a response must be labeled as:

Clear Reply: The response directly addresses the question constraints.

Ambiguous: The response partially addresses the question or uses vague language.

Clear Non-Reply: The response is a total evasion or a refusal to answer.

Evasion-Level Classification (Level 2): A fine-grained classification into 9 distinct evasion techniques. Based on the Thomas et al. (2024) taxonomy, these include:

Attacking the questioner

- Answering a different question

- Questioning the question’s premise
- Pleading ignorance or confidentiality
- Declining to answer (and others).

2. Review of the State of the Art (SOTA)

2.1 Theoretical Foundations

Research in this field originated with Bavelas et al. (1990) and the Equivocation Theory, later refined by Bull and Mayer (1993), who identified 30 manual tactics used by politicians to elude questions. For decades, analysis was limited to manual coding by political scientists, which lacked the scalability needed for modern digital media.

2.2 Current Computational Approaches (2024–2025)

State-of-the-art approaches have recently shifted from keyword-based irrelevance detection to deep contextual reasoning:

Large Language Models (LLMs): Current benchmarks utilize models like GPT-4o, Llama 3.1, and Mistral 7B. Recent findings (January 2026) indicate that Chain-of-Thought (CoT) prompting is essential for this task; models that explicitly "reason" through the question's core request before classifying the answer show significant accuracy gains (up to 74% in topic alignment).

Cross-Encoder Architectures: Unlike Bi-Encoders that process and separately, SOTA systems for CLARITY often use DeBERTa-v3 in a Cross-Encoder setup. This allows the model to capture the fine-grained interaction between the and tokens, identifying "pivots" where the politician shifts from the topic.

Natural Language Inference (NLI) Framing: Modern systems often treat evasion as an "entailment" problem. If the response does not logically entail a direct answer to the question's presuppositions, it is flagged. Research by Thomas et al. (EMNLP 2024) demonstrated that hierarchical models (predicting Level 1 and Level 2 simultaneously) outperform independent classifiers.

Quality Metrics beyond F1: The current SOTA also investigates "Answer Quality" through self-supervised learning. Models are trained to rank "observed" answers against "distractor" answers; if an observed answer is easily confused with a random one, it is quantitatively marked as having low semantic engagement (evasive).

3. Implemented approaches

3.1. First aproach

At first, we wanted to train some existing models, such as BERT, using the dataset provided by the organisers.

3.2. Second approach

Our second approach was to use Autogluon as an orchestrator. This library trains several models, and selects the best one.

It provides several presets, with better presets containing larger models:

- medium (default)
- high
- high v. 1.5
- best
- best v. 1.5
- extreme

It also allows you to specify a time limit for training.

We ran three experiments with Autogluon and submitted the results on Codabench.

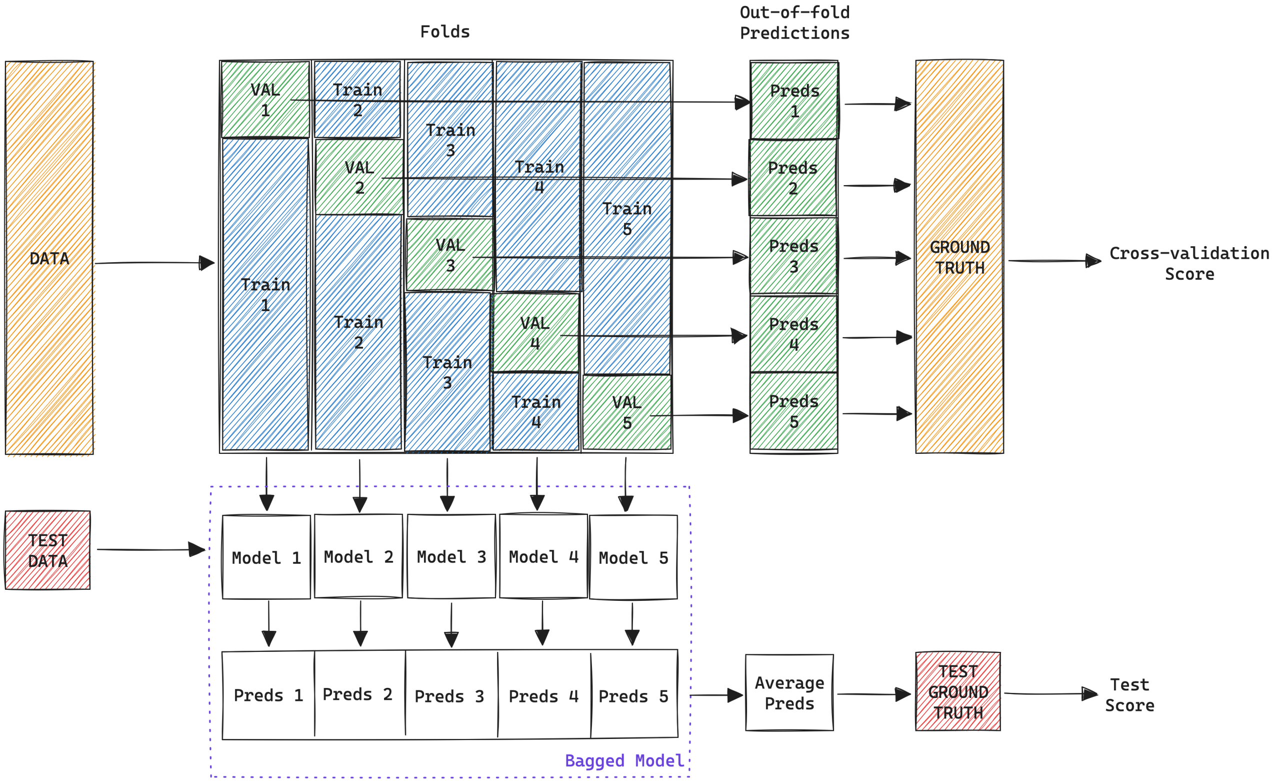


Fig. 1 - Autogluon architecture

4. Experimental results

4.1. Dataset

A training dataset was provided by the organizers (as well as a small testing dataset), and it was mandatory to use only this dataset in training the models. The dataset contains about 3.5k rows.

It contained the following columns:

- 'title',
- 'date',
- 'president',
- 'url',
- 'question_order',
- 'gpt3.5_summary',
- 'gpt3.5_prediction',
- # 'question',
- 'annotator_id',
- 'annotator1',
- 'annotator2',
- 'annotator3',
- 'inaudible',
- 'multiple_questions',
- 'affirmative_questions',
- 'index',
- 'evasion_label',
- 'interview_question',
- 'interview_answer',
- 'clarity_label'.

In the evaluation dataset, with 236 rows, only these columns were populated:

- 'url',
- 'index',
- 'interview_question',
- 'interview_answer'.

Our model computed the 'clarity_label' fields, and we submitted them on Codabench.

The datasets can be accessed here:

- training dataset
- evaluation dataset

4.2. Exploratory data analysis

We ran a test with Autogluon, submitting the training csv as-is, and it was heavily biased. This was due to many columns being unpopulated in the evaluation dataset. For this reason, we dropped all columns from the training dataset, except interview_question and interview_answer, before submitting it to Autogluon.

4.3. Experimental setting

We ran all experiments in a free Google Colab environment, with the following specs:

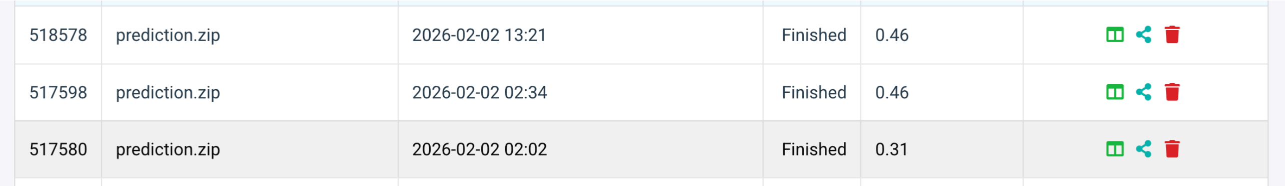
- T4 GPU instance
- 8 GB CPU memory
- 15 GB GPU memory
- 112 GB Disk space

4.4. Evaluation metrics

The evaluation metric used was F1-score, ran on Codabench.

We performed three experiments:

- Autogluon preset medium, time limit 1 minute - received an F1 score of 0.31
- Autogluon preset high_v150, time limit 20 minutes - received and F1 score of 0.46
- Autogluon preset high_v150, time limit 2 hours - received and F1 score of 0.46



518578	prediction.zip	2026-02-02 13:21	Finished	0.46	▥ ⋖ 🗑
517598	prediction.zip	2026-02-02 02:34	Finished	0.46	▥ ⋖ 🗑
517580	prediction.zip	2026-02-02 02:02	Finished	0.31	▥ ⋖ 🗑

4.5. Discussion

The first experiment was more of a smoke test.
For the second experiment, we upgraded the preset. But we saw in the logs that, for some models tried by Autogluon, it was reducing the model size, due to the modest compute environment.
For this reason, we decided not to attempt a larger model, but ran another experiment with the preset high_v150, and giving it a larger time limit.

The second and third experiment achieved the same F1 score. This is due to Autogluon training a very good model in both experiments, and it selected the same model both times. The compute environment was definitely a bottleneck, which prohibited us from training some larger models.
As a step-up, to achieve a stronger score, it would be necessary to upgrade to a more powerful (paid) compute environment.