Unmasking Political Question Evasions

Iaguta Alen-Mihael[1]

[1] Babeș-Bolyai University, Cluj-Napoca 400591, Romania
[2] alen.iaguta@stud.ubbcluj.ro

Abstract. CLARITY competition of the SemEval-2026 challenge represents a set of tasks which take into account the automatic classification of the clarity and evasive character. Task 6 of this competition is focused on identifying if politicians answers are directly targeting the question they are given or if they are using different forms of ambiguity or avoidance. This paper is exploring this challenge through a supervised learning matter and it is evolving some iterations of this model to outperform the classification.

Keywords: Clarity · Evasion · Classification

1 Introduction

Political discourse plays a central role in creating a public opinion in the democratic process. Between interviews, debates and press conferences, politicians are sometimes confronted with direct questions referring to public politics, governmental decisions or ideological positions. However, the answers given are not always clear or direct, they are frequently characterized by ambiguity, ignorance or changing the subject. These strategies can influence the perception of the public and can reduce transparency of political communication

In this context, automatic analysis of the clarity and the evasive character represents an important challenge for NLP (Natural Language Processing). The automatic identification of the scale of an answer to a question, or the way a question was answered, can contribute to a better understanding of discourse behavior of political actors and to the development of monitoring instruments of public speech.

This paper is focusing towards the CLARITY task from the SemEval 2026 competition, which follows classifying pairs of question-answer by clarity and evasion level used. The big issue in the study lies here: To What extent can AI-Models differentiate, in a reliable manner, clear answers and evasive answers in political discourse?

The importance of this problem lies not only from scientific perspective, but also on it's practical implications. From an academic point of view, CLARITY task contributes to the research of discourse analysis, pragmatic classification and the understanding communicational intentions. From an application point of view, the results can be used by journalists, political analysts and media organizations to further evaluate the quality of political communication and the level of responsibility of the public actors.

The target audience for this research includes NLP researchers, computational social studies, textual analysis researchers, also mass-media professionals interested in monitoring political discourse. In addition, this study may be relevant for institutions which want to promote transparency and responsible communication.

2 Related Work

The study towards the clarity of the answers in political discourse lies at the intersection between discourse analysis and linguistic pragmatics. In literature, the terms of answer equivocation and ambiguity was studied thoroughly in social and linguistic sciences, applying this knowledge in the analysis of political discourse and interviews. *Equivocation* is defined as a communication strategy where the speaker avoid providing a clear answer to a question, adopting instead, vague replies or multiple other possible interpretations. This category of answers was associated with diverse rhetorical techniques and was classified in elaborate typologies by researchers in the domain.

Collaborators proposed one of the many influential evasion typologies in political discourse, highlighting 3 main reply categories:

- **Clear Replies** - the asked information is given completely and without doubt
- **Clear Non-Replies** - the solicited information is not given at all, the replies can be completely unrelated or redirected to other topics
- **Ambivalent Replies** - partial answer was provided, ambiguous, or offers just a part of the asked information

This tradition lies on the base of modern taxonomy today and serves as a conceptual foundation for automatic classification approaches

2.1 NLP for Ambiguity, Clarity & Answerability

As NLP applications evolve, a lot of attention has been given to detecting ambiguity, if an answer is valid or not, in the means of standard datasets:

SQuAD 2.0 is a classic example, which extends *Stanford Question Answering Dataset* in order to include questions which can not be answered correctly from the provided text. Here, QA models not only need to identify the correct paragraph, but they also need to determine if an answer exists.

Although this is not focused entirely on human intent to avoid a question, the approach is relevant for this paper in question, as models need to be able to distinguish clear vs non-clear replies.

2.2 Detecting uncertainness in text

NLP researching towards uncertainness or ambiguity, addresses similar text classification problems based on the degree of clarity.

Using key words like *maybe, I believe, it seems* are strategies commonly interpreted as avoiding precision. Spontaneous speech impediments or prepositions like *uh* and *um* offer a delay in answers which shows signs of vagueness and inaccuracies. [CT02]

Recent work from NLP community uses transformation models to identify *hedging*[3] in academic, medical or political discourse, ultimately demonstrating how useful classifiers are for obtaining a grade of certainty/clarity in phrases. This offers methodological perspectives towards obtaining text characteristics for classification, sustaining the idea for detecting levels of clarity, although it is not identical with the intended equivocation.

Modern models, based on Transformer architectures, like BERT and RoBERTa, have proved their capacity to identify these linguistic niches, by analyzing semantic context of words. These models learn to recognize patterns associated with uncertainness. For example, using modal verbs in speculative expressions or in indirect formulations of answers, this makes them suitable for classification tasks, especially in finding the clarity in answers.

2.3 Rhetorical strategy in politics

In political science, modern rhetorical theories have associated evasion with persuasion and agenda control.[Lak73] Deliberate evasion is an used instrument not only for hiding information, but also to influence public perception. Politicians happen to follow these terms very strictly, as they *redirect the scope* as an avoidance technique. Intentional equivocation can be efficient in the context of political debates, as when the goal is to control the flow of speech. Common techniques used:

- **Ignoring the question** - the politician, literally, doesn't answer the question
- **Refusing to answer** - different from ignoring as here, politicians show evasive affirmations
- **Equivocation as secondary argument** - redirect the topic to related topics
- **Repeating** - going back to an already mentioned point in the interview/conversation without clarifying the initial question

[3] HEDGING definition: a way of avoiding giving a direct answer or opinion, a way of controlling or limiting a loss or risk

3 Data

The used dataset can be found here. This dataset contains question/answer pairs, extracted from political interviews, together with labels that show the clarity of the answer and the type of evasion present (if it exists).

The dataset is built to answer directly to the issue that the project is referring to - automatic classification of clarity and evasion technique in political discourse. This is an official public resource, manually labeled, which is used to train and evaluate AI models. It was collected from presidential interviews from the official Whitehouse website, having a total of 287 unique interviews, from which a total of 3445 questions and responses have been extracted. ChatGPT was used to decompose the originalQA[4] pairs.

3.1 Annotation process

The following labels present in the dataset have been labeled manually by multiple human annotators, which have analyzed every QA pair and they determined the grade of clarity and evasion corresponding to the pair. To ensure quality and consistency between labels, because each human annotator worked independently, the Fleiss' Kappa coefficient was used. This coefficient represents a measuring statistic used to evaluate the degree of agreement between annotators, when they classify a set of QA pairs into discrete categories.

Table 1. Fleiss' κ values

	Clear Reply	**Clear Non-Reply**	**Ambiguous Reply**
Clear Reply	1.00	0.97	0.65
Clear Non-Reply	0.97	1.00	0.71
Ambiguous Reply	0.65	0.71	1.00

The values obtained in the annotation process, suggests that labels are consistent and viable. Values nearing 1 are found in labels that are relatively simple to agree upon, for example between "Clear Reply" and "Clear Non-Reply". However, "Ambiguous Reply" category struggles with agreement between annotators, showing values between 0.65 and 0.71 to other categories. This shows that "Ambivalent/Ambiguous Reply" category struggled the most for annotators, indirectly showing that they have the most occurrences in the dataset labeling process.

[4] QA = short for question answer

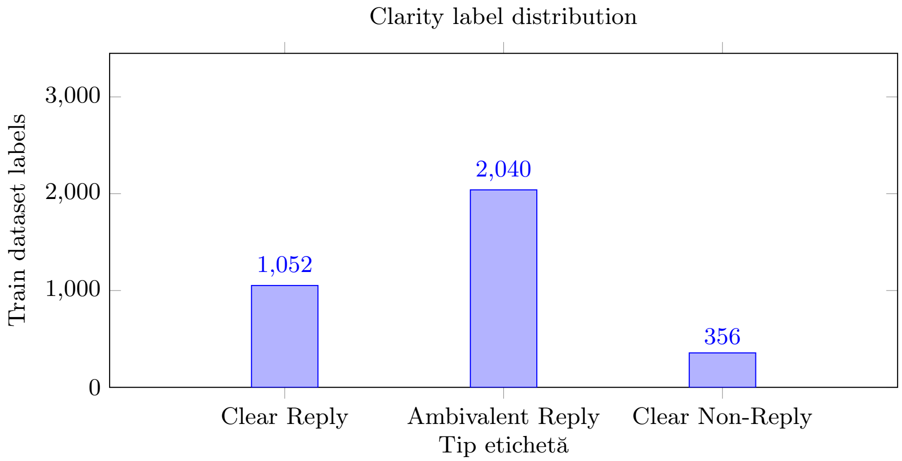


Fig. 1. Clarity labels in QEvasion dataset, showing numbers for each class: *Clear Reply*, *Ambiguous Reply* și *Clear Non-Reply*.

3.2 Validation set agreements

The dataset is split between training and test/validation set, each of them having a specific role in the development process of the ML models. The validation set is used to evaluate the performance of the model during training and in order to adjust hyper-parameters, offering thus a realistic estimation of the models capacity to generalize new, never-seen before data.

An important characteristic of this dataset is that the validation set does not contain the evasion label, as ultimately, this absence is intentional and reflects the structure of the SemEval 2026 competition and the research problem, where the objective is to develop capable models which can predict these labels automatically. More specifically, evasion labels are reserved exclusively for the training set.

3.3 Preparing the dataset

In order to continue the experiments proposed, the dataset was preprocessed and prepared using Google Colab, which offers support for executing code and integrating libraries, specific to NLP and DL. Using this space, it allowed access to certain computational resources like, GPU acceleration, making the whole experiment preparation much easier.

The first step in the preparation stage was to load the dataset, using the library *Hugging Face Datasets*, but only a few attributes we're selected for this research:

- **interview_question**
- **interview_answer**

- **clarity_label**
- **evasion_label**

The fields, *interview_question* and *interview_answer* have been concatenated in order to create a singular text entry, such that the model can analyze the complete context of the interaction. This approach allows the model to understand the semantic relation between question and answer, which is essential for detecting evasive behavior.

The next step in preparing the dataset was to take a deeper look at the labels:

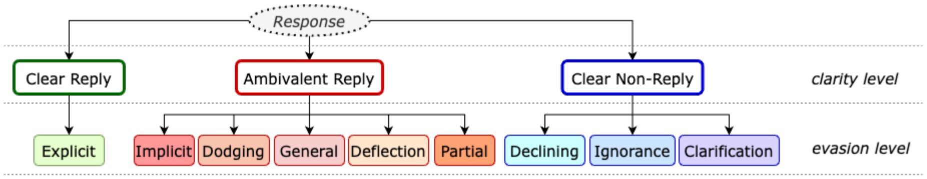


Fig 2. Full branch of response classification categories

However, AI models need numerical labels, thus every clarity and evasion label has been switched to a number, values from 0-2, and 0-8 respectively.

After the label conversion, the text has been processed using a specific tokenizer for RoBERTa model. One more optimization detail used was to remove unnecessary fields like *url, date, annotator_id, etc.*

4 Methodology

For solving classification tasks on the degree of clarity and the level of evasion, a DL[5] method was used, the *Transformer* architecture was leveraged, especially Facebook AI RoBERTa. This choice was motivated by the high performance results, demonstrated by the Transformer models in many natural language processing tasks, including text classification, semantic analysis and detecting intent.

The model used, RoBERTa-base, is a pretrained model on high volume of text data, which allows it to learn complex contextual representations of the language. For this experiment, the model was adapted for a multi-class classification task, by adding a final level of classification which produces probabilities for each class.

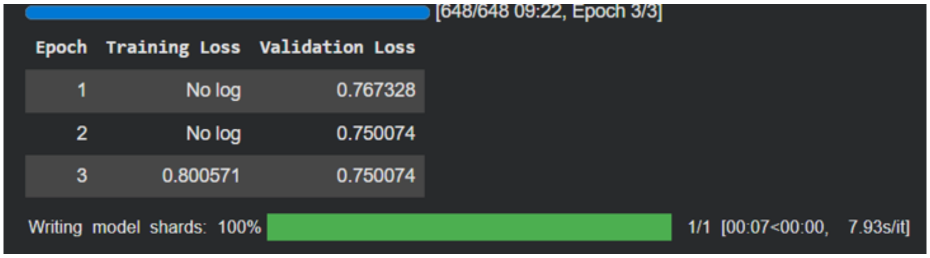
For the first task, the model was trained to classify answers from one out of three clarity classes, ultimately to analyze the relation between clarity and evasion.

[5] DL = short for Deep Learning

4.1 Test configurations & evaluations

The primary parameters used to determine an optimal state we're the following; **learning rate** between *2e-5 and 3e-5*, on a number of **epochs** between three to six, using **batch sizes** of 16, having a **maximum length sequence** of 384 tokens.

For evaluating the performance of the model a validation strategy has been used, based on the validation set. By training the model on the training set and evaluation every epoch per validation set:



[648/648 09:22, Epoch 3/3]

Epoch	Training Loss	Validation Loss
1	No log	0.767328
2	No log	0.750074
3	0.800571	0.750074

Writing model shards: 100% 1/1 [00:07<00:00, 7.93s/it]

Fig 3. Example of how the validation strategy was used (ignore values as they are not accurate)

This approach allows to monitor the performance better so that the overfitting can be spotted, thus a better model can be selected, for which the *Macro F1-score* metric was used. This is the most suitable metric for classification problems, especially, multi-class problems as it is offering a balanced evaluation.

Macro F1 treats all classes equally, regardless of their distribution. For training the *Cross Entropy Loss function* was used, standard for multi-class classification and for optimizing *AdamW*, as it is most recommended for most Transformers. The optimizer was not explicitly defined, as it was automatically created in the *Trainer* class from *Transformers* library.

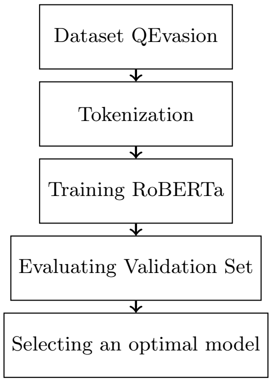


Fig 4. Methodological process used

5 Results and Discussion

The overall final goal of these experiments was to evaluate the capacity of the RoBERTa-base and RoBERTa-large models, to identify correctly the three clarity classes, likewise for the nine evasion labeled techniques. Additionally, this configuration should have a say in to the research question proposed, this being if a Transformer model can efficiently learn linguistic patterns associated to clear, ambiguous or evasive replies.

A clearer analysis of the results showcase that the models differ from the three clarity classes.

5.1 Task I: Clarity

A closer look at the results shows that the model performance differs across the three clarity classes. The extreme classes, namely “Clear Reply” and “Clear Non-Reply”, were classified with higher accuracy, while the “Ambiguous Reply” class showed the lowest performance.

This difference can be explained by the linguistic nature of these classes. Clear replies usually contain explicit linguistic indicators, which makes their classification easier. In contrast, ambiguous replies are characterized by a higher degree of variability and lack of clarity, which makes their classification more difficult.

These results are consistent with previous studies, which have shown that ambiguity detection is a complex task, even for human annotators.

Table 2. The results of RoBERTa models for classifying clarity

Model	Epoch	Accuracy	Precision	Recall	Macro F1
RoBERTa-base	1	0.68	0.54	0.43	0.44
	2	0.66	0.56	0.55	0.50
	3	0.66	0.59	0.54	0.49
	4	0.69	0.58	0.51	0.53
	5	0.64	0.53	0.47	0.50
	6	0.65	0.54	0.55	0.55
RoBERTa-base (final)	-	**0.69**	**0.59**	**0.55**	**0.55**
RoBERTa-large	1	0.68	0.40	0.37	0.34
	2	0.66	0.54	0.55	0.52
	3	0.68	0.62	0.58	0.53
	4	0.69	0.71	0.53	0.59
	5	0.68	0.58	0.53	0.55
	6	0.67	0.56	0.56	0.57
RoBERTa-large (final)	-	**0.69**	**0.71**	**0.58**	**0.59**

The results presented in Table 3 highlight the performance of the RoBERTa-base and RoBERTa-large models for the task of clarity level classification. In the case of the RoBERTa-base model, it can be observed that the accuracy remains relatively stable over the six epochs, ranging between 0.64 and 0.69. The best final performance of this model was obtained at a Macro F1 score of 0.55, together with an accuracy of 0.69, a precision of 0.59 and a recall of 0.55.

There is also variation between epochs in Precision and Recall scores, suggesting some instability in the learning process. For example, in the third epoch, the model achieves a precision of 0.59, but the recall remains at 0.54, indicating that the model is more conservative in its predictions and may omit some relevant examples. This discrepancy between precision and recall is typical for unbalanced datasets, where some classes are more frequent than others.

Regarding the RoBERTa-large model, the results indicate an overall superior performance compared to the RoBERTa-base variant. Although in the first epoch the performance is relatively low, with a Macro F1 score of only 0.34, the model shows a significant improvement in subsequent epochs. The final performance of the RoBERTa-large model achieves a Macro F1 score of 0.59, a precision of 0.71, a recall of 0.58 and an accuracy of 0.69. The significant increase in precision indicates that the model is much more efficient in correctly identifying classes, reducing the number of false positive predictions.

The difference in performance between the two models can be explained by the more complex architecture of the RoBERTa-large model, which contains a larger number of parameters and, implicitly, a greater ability to learn complex semantic representations, having a total runtime of approximately 20 minutes for each model (3+ minutes for each epoch).

5.2 Task II: Evasion

The results obtained for the classification of the evasion level with the RoBERTa-base model are presented in the table above. A progressive increase in performance is observed as the model goes through the training epochs. At epo epoch 1, Macro F1 is only 0.17, indicating the difficulty of the model to correctly differentiate the 9 evasion classes in the initial phase. As training progresses, all metrics-accuracy, precision, recall, and Macro F1-record a significant increase, and at epoch 6 the model reaches a Macro F1 value of 0.58, with an accuracy of 0.63 and balanced precision and recall, reflecting better learning of the subtle linguistic patterns that characterize evasive responses.

The final performance of the model, reported as RoBERTa-base (final), confirms that the model can predict with moderate accuracy the level of avoidance, achieving 0.58 for Macro F1. This suggests that the task of classifying evasion is much more complex than the task of clarity, due to the large number of classes and the linguistic subtleties involved. The distribution of predictions in the CSV file shows that the model was able to differentiate between the main classes, but there is still confusion between some levels, which could be improved by larger models or class balancing techniques.

Table 3. The results of RoBERTa-base for classifying evasion

Model	Epoch	Accuracy	Precision	Recall	Macro F1
RoBERTa-base	1	0.34	0.16	0.23	0.17
	2	0.41	0.38	0.36	0.30
	3	0.50	0.48	0.47	0.40
	4	0.57	0.57	0.50	0.51
	5	0.61	0.58	0.56	0.55
	6	0.63	0.57	0.58	0.58
RoBERTa-base	-	**0.63**	**0.58**	**0.58**	**0.58**
RoBERTa-large	1	0.30	0.07	0.20	0.08
	2	0.39	0.28	0.34	0.26
	3	0.46	0.44	0.46	0.39
	4	0.55	0.57	0.50	0.49
	5	0.62	0.59	0.57	0.56
	6	0.64	0.59	0.60	0.60
RoBERTa-large	-	**0.64**	**0.59**	**0.57**	**0.60**

The performance of the model, reported as RoBERTa-base , confirms that the model can predict with moderate accuracy the level of avoidance, achieving 0.58 for Macro F1. This suggests that the task of classifying evasion is much more complex than the task of clarity, due to the large number of classes and the linguistic subtleties involved. The distribution of predictions in the CSV file shows that the model was able to differentiate between the main classes, but there is still confusion between some levels, which could be improved by larger models or class balancing techniques.

The distribution of the levels of evasion predicted by the RoBERTa-base model shows significant variability across the nine classes investigated. This distribution balances our perspective on model performance: even though Macro F1 increased over training, the inequality between class frequencies suggests that Transformer models such as RoBERTa-base more readily capture common patterns, but may perform more poorly in recognizing subtle or rare evasion tactics. A deeper dive on this distribution is to take a look at the question and answer itself. Two QA pairs were extracted at random to show the results predicted:

- **Q:** *What kind of punishment would you like to see imposed on North Korea, short of some sort of condemnation from the U.N. Security Council?"*
  A: *"Okay"*
  Row 99 - Predicted: *Dodging*
  How was this interpreted in a human manner? Answer given was really short and vague, it does not reply to the question in any way about the sanctions and so on. The model classified this reply as *dodging*, which, semantically is correct, this person clearly avoids offering an opinion or a concrete solution. This strategy is typical for evasive political discourse, where the subject is avoided or redirected, and the answer *"Okay"* is a clear sign of refusing to answer.

- **Q:** *What is the connection between Iraq and the topic being discussed?"*
  A: *"What did have to do with what?"*
  Row 237 - Predicted: *Clarification*
  In this case, the reply given does not offer an argument or an opinion, but it solicits clarifying the question. The model picked correctly, because the speaker was trying to understand what the question means before actually answering. This is different than dodging, here the speaker tries to clarify the ambiguity.

Comparing the predictions of the two models for classifying the level of evasion, it can be observed that RoBERTa-large tends to make more frequent adjustments with respect to the model base, especially on the more subtle classes such as Implicit, Declining to answer or Claims ignorance. In many cases, RoBERTa-large converts RoBERTa-base predictions into more generalized or precise classes, such as changing a General to Dodging or Explicit to Dodging, suggesting that the large model can better capture these techniques.

The distribution of predictions shows that the dominant classes remain Explicit, Dodging and Claims ignorance, while rare classes, such as Clarification or Deflection, are harder to recognize, but the large model tends to identify them more often than RoBERTa-base. This highlights a trade-off between the model's ability to learn complex patterns and the difficulty of balanced classification across all 9 classes.

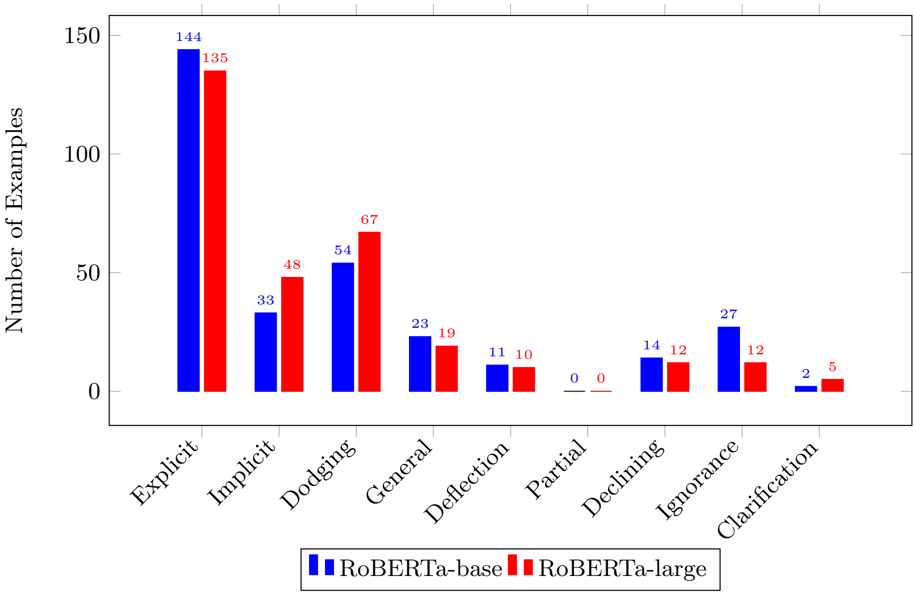


Fig 5. Distribution of evasion

The analysis of the distribution of predictions clearly shows that the most common avoidance strategies are Explicit, Dodging and Claims ignorance, while the rarer classes, such as Partial or Clarification, appear much less frequently in the dataset. It is important to note that no Partial responses were identified in the data, suggesting that participants completely or clearly avoided the questions, and did not provide partial or incompletely formulated answers. Comparing the two models, RoBERTa-large tends to make subtle adjustments on the less frequent classes, better capturing fine-grained variation between avoidance types, but the overall distribution remains similar to that obtained by RoBERTa-base. This observation confirms that the models can distinguish most of the dominant strategies, but rare classes remain harder to detect, which may influence the overall accuracy and interpretability of the results.

6 Conclusions

The following study had as an objective, investigating the possibility to classify automatically the level of clarity in answers and to analyze the evasive behavior in political discourse, using QEvasion dataset and a model based on the Transformer architecture. The main research question lies on the way a pretrained model was visualized, how RoBERTa, developed by Facebook AI, can learn to distinguish between different levels of clarity on a question answer pair. The experimental results showed that the model obtained a Macro F1-score of approximately 0.60 on the validation set, which shows a good capacity to identify linguistic patterns associated to clarity. These results confirm that Transformer models are efficient for semantic classification tasks, likewise in previous researching studies done in the domain of natural language processing.

At the same time, these results showcase some limitations, especially in differentiating ambiguous replies, which suggests that this class is more difficult to automatically detect, mainly because of its intermediate character and high linguistic variability. However, the performance obtained demonstrates that the approach used is fitting for this task and can represent a base for future developments. Additionally, using a manually labeled dataset, validated throughout annotators, contributes to the relevance of results and to the consistency of the analysis. This aspect is highlighted also in recent studies showing detecting evasion in political discourse. In conclusion, the results obtained sustain the assumption that Transformer based models can be used successfully in order to analyze clarity in answers, although improving performance can remain an open direction for other researchers.

References

CT02. Herbert H Clark and Jean E Fox Tree. Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111, 2002.

Lak73. George Lakoff. Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of philosophical logic*, 2(4):458–508, 1973.