HappyFrogs at SemEval-2026-Task 6:DeltaSHAP: a Shapley Value Framework for Interpreting Political Ambiguity

Sven-Alexander Gal
Babeș-Bolyai University
Cluj Napoca
Romania
sven.gal@econ.ubbcluj.ro

Rodica-Ioana Lung
Babeș-Bolyai University
Cluj Napoca
Romania
rodica.lung@econ.ubbcluj.ro

Abstract

Political ambiguity and response clarity have become increasingly important research topics in computational social science and natural language processing. In this paper, we present a solution to the SemEval 2026 Task 6 "Clarity" Challenge. We propose a novel framework that employs TF–IDF representations and Shapley-value–based feature selection for multi-class classification. Shapley-based feature importances are used both for post-hoc explanation and as an active mechanism for label-specific vocabulary selection. For each label, features exceeding a predefined threshold are retained, label-specific vocabularies are filtered through set differences, and independent one-versus-all classifiers are trained using specific features. Experimental results show that threshold tuning substantially impacts performance, with the best performance achieved at intermediate threshold values(1.4, more exactly). Our findings demonstrate that using the game-theoretic feature selection provides an interpretable approach to clarity classification, offering a flexible methodology for ambiguity-sensitive text analysis.

1 Introduction

Studying political ambiguity has emerged as an important research topic in recent years. A significant study for this research framework is proposed by Thomas et al. (2024), proposing a novel two-level taxonomy designed specifically to evaluate how clearly a response addresses a given question, rather than attempting to infer speaker intent or deception. One of the major contributions of their work is the construction of a human-validated dataset consisting of 3,445 question–answer pairs extracted from U.S. presidential interviews spanning nearly two decades. From a modeling perspective, the study establishes strong baselines across encoder architectures and large language models, demonstrating that clarity classification benefits substantially from hierarchical reasoning. (Thomas et al., 2026)

Despite existing advances, the problem of explicitly identifying and interpreting the features that drive ambiguity classification remains underexplored, particularly in shared-task settings where both performance and explainability are critical. Moreover, existing approaches in the SemEval 2026 Task 6 ("Clarity") largely focus on predictive performance, with limited emphasis on transparent, label-specific feature attribution.

In this paper, we propose DeltaSHAP, a framework for solving Task 6 of the SemEval Conference ("Clarity") that employs a game-theoretic feature-selection approach based on the Shapley value. The Shapley value is used to quantify feature importance and to select features for each label in classification. Performance is evaluated using macro F_1 scores across a variety of settings, and qualitatively by providing keywords identified using the Shapley value-based approach. Thus, the findings can help reveal key characteristics in data using a game-theoretic machine-learning approach.

Thus, we position our approach within the SemEval 2026 shared task as a hybrid interpretability-driven framework that complements standard encoder-based solutions by introducing feature-level reasoning grounded in cooperative game theory. Our main contributions are as follows:

- we introduce DeltaSHAP, a novel Shapley value-based method for label-wise feature selection in ambiguity classification;
- we offer a qualitative analysis of ambiguity cues, revealing task-relevant linguistic patterns through Shapley-derived feature importance.

2 Related work

Natural language processing (NLP) methods have been widely employed to study ambiguity in polit-

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2939–2945

ical texts, where strategic vagueness and framing play a central role. Early work focused on lexical and syntactic ambiguity detection using probabilistic models and word sense disambiguation techniques, while more recent approaches leverage contextual embeddings from transformer-based models such as BERT to capture nuanced semantic uncertainty (Devlin et al., 2019). Topic modeling and framing analysis have also been used to identify ambiguous or multi-interpretable narratives in political discourse (Grimmer and Stewart, 2013), while supervised classifiers detect hedging, vagueness, and equivocation in speeches and manifestos (Recasens et al., 2013). Additionally, stance detection and sentiment analysis help reveal implicit ambiguity by capturing conflicting signals within the same text (Mohammad et al., 2016). More recently, explainable NLP methods, including attention-based models and feature attribution techniques, have been applied to uncover how ambiguity manifests at the token and phrase level, offering deeper insights into strategic communication in politics (Vig and Belinkov, 2019). Together, these approaches highlight NLP's capacity to systematically quantify and analyze ambiguity in political language.

Shapley-based explainability has been widely applied in natural language processing (NLP), extending beyond post-hoc interpretation to feature and vocabulary selection. SHAP methods aggregate feature contributions across instances to derive global importance rankings, enabling explanation-driven selection of informative inputs (Marcílio and Eler, 2020). While such approaches can be competitive with traditional techniques, their effectiveness depends on the alignment between the underlying game-theoretic formulation and the predictive task (Trotskii et al., 2025; Fryer et al., 2021). In NLP, Shapley-inspired frameworks have been adapted to unstructured text. Vocabulary selection can be modeled as a cooperative game, where Shapley values identify influential words and outperform frequency-based baselines (Patel et al., 2021). Similarly, TokenSHAP enables fine-grained token-level attribution in large language models through Monte Carlo estimation (Goldshmidt and Horovicz, 2024). More broadly, SHAP-based methods provide a unifying framework for interpretability across NLP tasks, with connections to attention-based explanations under certain conditions (Mosca et al.; Ethayarajh and Jurafsky, 2021). Despite these advances, theoretical and empirical studies highlight limitations, including computational costs and a mismatch between Shapley axioms and feature-selection objectives (Fryer et al., 2021). Moreover, there is limited work that combines SHAP-based analysis with one-versus-all classification to study specific discourse, such as ambiguity in political speech, leaving a clear direction for further research.

3 DeltaSHAP

DeltaSHAP[1] is a framework for identifying specific features associated with different labels using the concept of Shapley values, feature selection, and a one-versus-all classification mechanism for aggregating results. Algorithm 1 describes an outline of the framework.

Algorithm 1 DeltaSHAP outline.

1: TF-IDF processing;
2: Shapley value for every label (CatBoost);
3: Selecting words within a SHAP threshold for every label;
4: Select features based on differences between sets;
5: Separate classification of every label according to the features chosen using a threshold of the minimum Shapley value (RandomForest);
6: Aggregating the predictions for every label.

In the first step (Alg. 1, line 1), TF-IDF processing is used to extract the most informative features for subsequent analysis, yielding the document-term matrix DTM.

Next (line 2, Alg. 1), a Shapley score is computed for each feature in the DTM, based on (Marcílio and Eler, 2020). The Shapley score is computed separately for each individual label. Thus, for each word in the DTM, three scores indicating its contribution to each label are computed. CatBoost is used as a baseline classifier for computing the Shapley scores (Prokhorenkova et al., 2018).

Subsequently, DTM features (words) are selected based on a SHAP threshold δ for each label (line 3, Alg. 1). The value of δ influences the size of the selected sets, which are to be used to construct feature sets in the subsequent steps. Thus, for each label, we have a feature set containing words identified by the Shapley value as important for classifying that label.

[1] https://github.com/galsven/DeltaSHAP/blob/main/Main1last accessed Apr. 2025

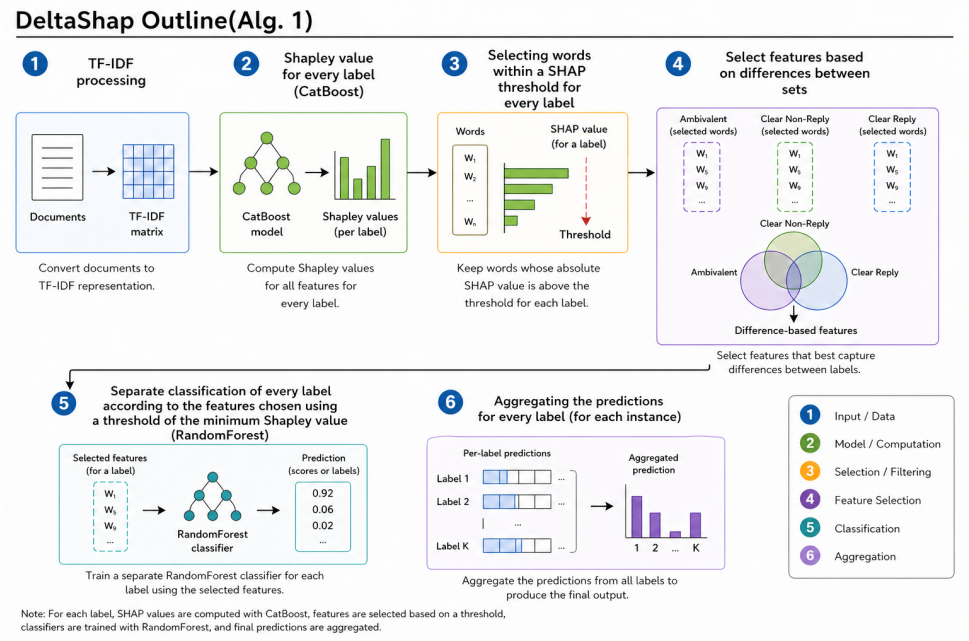

Figure 1: A visual representation of DeltaSHAP: classification is performed using the feature sets specific to each label as identified by Shapley-based importances.

In the following step (line 4, Alg. 1), set differences between the resulting feature sets to identify words that may characterize only a certain label. Thus, for each label from its corresponding feature set, the words that appear in the other two are subtracted.

Afterwards (line 5, Alg. 1), a Random Forest classifier is trained separately for each label, following a one-vs-all strategy, using the features selected based on the minimum Shapley value threshold in line 4. Finally (line 6, Alg. 1), the individual label predictions are aggregated to produce the final classification output by choosing the label with the highest predicted probability among the three models. The entire process is illustrated in Figure 1.

4 Numerical results

4.1 Experimental setup

The dataset curated by the organizers of Task 6, available on Hugging Face, is used as the training data, with the corresponding challenge test dataset. The columns passed to TF-IDF are: 'url', 'question_order', 'interview_question', and 'interview_answer'. The best-known results obtained by **?** previously used a fine-tuned Llama-70b model, achieving an $F1$-score (macro) of 0.68. TF-IDF parameters used are: *max_features*=5000, *ngram_range*=(1,2), and *stop_words*="english". *max_features* limits the vocabulary size to the 5,000 most informative terms according to their TF-IDF scores across the corpus, reducing dimensionality and improving computational efficiency. *ngrams_range* enables the extraction of both unigrams (single words) and bigrams (two consecutive words). Unigrams capture the importance of individual terms, and bigrams enable the model to learn short contextual expressions (e.g., "human rights", "climate change") that may carry stronger semantic meaning than isolated words. *stop_words* removes common English function words such as "the", "is", "and", or "of", which typically appear frequently but provide little discriminative information for classification.

DeltaSHAP was initially tested with δ thresholds ranging from 1 to 10; however, preliminary results (Figure 5 indicated that thresholds between 0.5 and 1.5 may provide a better insight into the data. Apart from δ and the classifiers used in Steps 2 and 5 of Algorithm 1, DeltaSHAP does not require other specific parameters.

To illustrate the effect of using the Shapley value-based mechanism in the DeltaSHAP algorithm, we also provide results from Random Forest and CatBoost as baselines.

4.2 Results and discussions

Figures 2, 3, and 4 illustrate the most significant contribution of the term/words related to each of

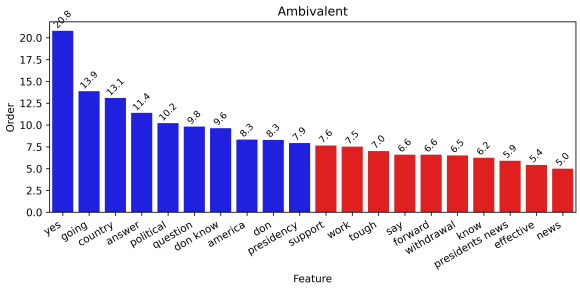

Figure 2: 'Ambivalent' label: terms with the highest SHAP importances, that contribute most to the classification of this label.

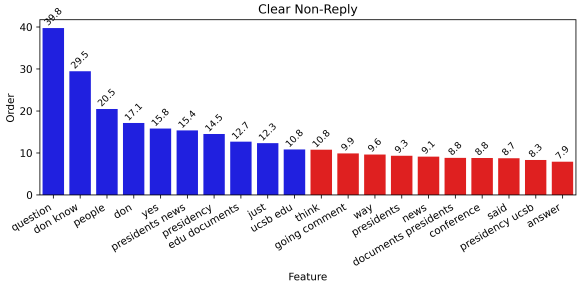

Figure 3: 'Clear Non-Reply' label: terms with the highest SHAP importances, that contribute most to the classification of this label.

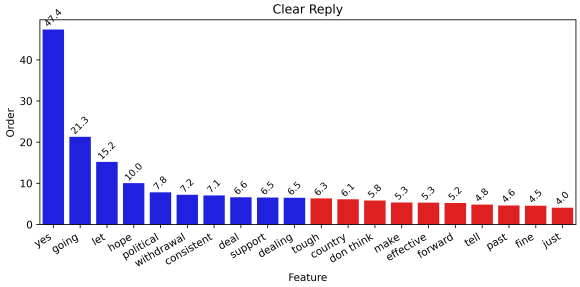

Figure 4: 'Clear Reply' label: terms with the highest SHAP importances, that contribute most to the classification of this label.

the three classes based on the Shapley value analysis. We emphasize the 10 most significant ones for each label in blue. We can observe in Figure 4, that words that have the highest Shapley value for 'Clear Reply' label are strong and affirmative, such as 'yes', 'support' or negative words, such as 'withdrawal'. When considering the 'Ambivalent' label, we find again 'yes' as topping the charts, as well as nouns such as 'question' or 'answer', as we can see in figure 2. At the 'Clear Non-reply' label, we can observe negative expressions such as 'don know' or words expressing something happening that just happened ('just').

Figure 5, presents F_1 scores for δ values ranging from 1 to 10. We observe that the F_1 decreases

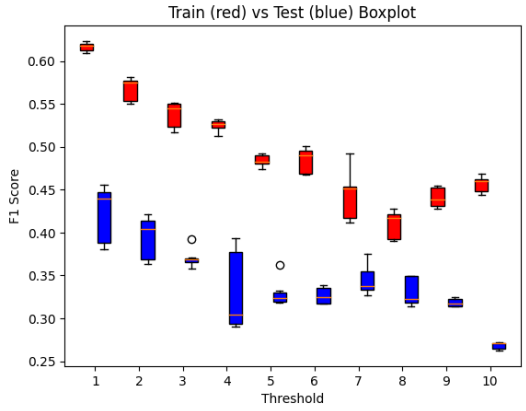

Figure 5: F_1 train/test scores reported for δ values between 1 and 10. The decreasing trend indicates that threshold values near 1 should be explored further.

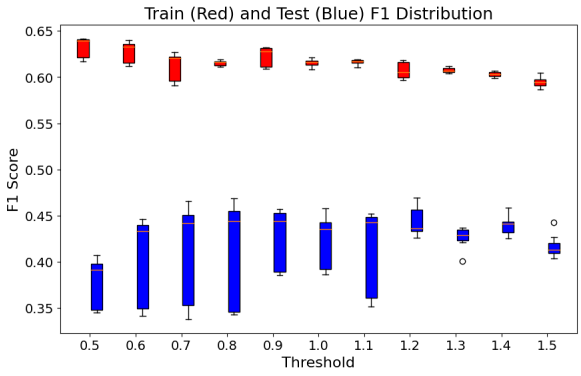

Figure 6: Distribution of F_1 scores among thresholds between 0.5 and 1.5. Differences between train and test F_1 values indicate overfitting, illustrating both a limitation of the approach and the complexity of the data.

as the threshold increases. This led us to further examine results for thresholds ranging from 0.5 to 1.5 to determine an efficient range.

Thus, Figure 6 illustrates boxplots of the distribution of F_1 scores across thresholds between 0.5 and 1.5. An empirical analysis led us to consider 1.4 as the optimal working threshold, as a trade-off among F_1 training values, variance (among the smallest possible variance values), and the number of features selected for each label. These numbers are illustrated in Figure 7, as bars comparing the number of features selected for every label and threshold values from 0.5 to 1.5. As the threshold value increases, the number of features specific to each label decreases, making classification more difficult.

Table 1 reports the mean, median, and standard deviation of F_1 scores reported by DeltaSHAP, using thresholds ranging from 0.5 to 1.5. We high-

Table 1: Train and test F_1 scores (mean, median, and the standard deviation) for different thresholds after 10 runs for each threshold; a $(*)$ indicates DeltaSHAP values can be considered significantly better than both baseline models, and (-) indicates no significant difference. Comparisons are performed with CatBoost results. We considered the alpha as 0.05

DeltaSHAP threshold	Train F_1 Mean	Train F_1 Median	Train F_1 StDev	Test F_1 Mean	Test F_1 Median	Test F_1 StDev
0.5	0.6322	0.6393	0.0109	$0.3780^{(-)}$	$0.3914^{(-)}$	0.0270
0.6	0.6270	0.6326	0.0114	$0.4018^{(-)}$	$0.4330^{(-)}$	0.0481
0.7	0.6116	0.6209	0.0145	$0.4105^{(-)}$	$0.4421^{(-)}$	0.0555
0.8	0.6153	0.6162	0.0029	$0.4110^{(-)}$	$0.4437^{(-)}$	0.0576
0.9	0.6226	0.6279	0.0104	$0.4260^{(*)}$	$0.4438^{(*)}$	0.0331
1.0	0.6154	0.6163	0.0037	$0.4213^{(*)}$	$0.4356^{(*)}$	0.0283
1.1	0.6166	0.6172	0.0027	$0.4112^{(*)}$	$0.4424^{(*)}$	0.0464
1.2	0.6074	0.6054	0.0085	$0.4437^{(*)}$	$0.4359^{(*)}$	0.0155
1.3	0.6075	0.6069	0.0029	$0.4266^{(*)}$	$0.4291^{(*)}$	0.0107
1.4	0.6034	0.6034	0.0027	$0.4403^{(*)}$	$0.4411^{(*)}$	0.0110
1.5	0.5944	0.5942	0.0054	$0.4161^{(*)}$	$0.4128^{(*)}$	0.0114
Baseline models						
RandomForrest	0.8480	0.8480	0.00028	0.3007	0.3007	0.00944
CatBoost	0.7769	0.7769	0.00115	0.3873	0.3873	0.00835

Table 2: List of distinct terms selected by DeltaSHAP for each label, as well as the common terms that are not included.

Ambiguity	'great', 'everybody', 'possible', 'today', 'leave', 'called', 'worst', 'saying', 'asked', 'weapons', 'suspend', 'weren', 'pointed', 'effort', 'met', 'partners', 'happened', 'week', 'trade', 'union', 'people want', 'soon', 'change', 'conversation', 'doing', 'power', 'conflict', 'far', 'various', 'process', 'options', 'americans', 'answer question', 'war', 'create'
Clear Non-Reply	'united states', 'years', 'talking', 'things', 'questions', 'comment', 'falling', 'succeed', 'challenge', 'inaudible', 'north korea', 'congress', 'look', 'potentially', 'good', 'got', 'understanding', 'able', 'help', 'appreciate', 'reason', 'happen', 'hard', 'let know', 'earlier', 'talk', 'analyze', 'haven seen', 'condi', 'know going', 'economy', 'fair', 'thank', 'world', 'fit', 'hear', 'passed', 'region', 'probably', 'think going', 'exactly', 'come', 'going make', 'important', 'launch', 'www',
Clear Reply	'quick', 'negotiate', 'easier', 'conference president', 'secretary', 'anticipate', 'mr', 'fine', 'ambassador', 'literally', 'secretary state', 'civilian', 'signed', 'israel', 'like', 'fuel', 'make sure', 'justice', 'working', 'having', 'don want', 'basis', 'directly', 'anytime', 'innocent', 'seriously', 'strategy', 'legal', 'game', 'want people'
Common terms	'conference', 'house', 'presidency', 'country', 'america', 'yes', 'let', 'going', 'don think', 'don know', 'news', 've', 'work', 'thank mr'

light the value $\delta = 1.4$ as a potential optimal setting, given the low standard deviation of the train and test values. As there is no evidence of significant differences among the results reported for this value and the neighbouring values, we highlight it as an example of a possible interpretation. Nevertheless, the threshold value controls the size of the feature sets, and its choice, considering similar performance metrics values, may be subject to other interpretability-related criteria.

Furthermore, statistical significance tests (t-tests and Wilcoxon nonparametric tests with $\alpha = 0.05$) comparing the mean and median of the F_1 values for DeltaSHAP with the baselines are also included in Table 1. A (*) represents results where DeltaSHAP is significantly better than the baseline model. A $(-)$ symbol indicates no significant difference between the results of baseline algorithms and DeltaSHAP.

The significance tests indicate that, in most cases (from 0.9 to 1.5), the mean and median of the F_1 score from the DeltaSHAP algorithm yield significantly better results than those of CatBoost (and Random Forests). For smaller thresholds (from 0.6 to 0.8), the mean and median of the F_1 scores are higher, but not statistically significant. When the threshold increases and becomes more selective, as seen in 7, the difference between results becomes significant. The fact that in 7 of 11 cases the results differed significantly indicates DeltaSHAP's potential to improve classification accuracy.

For comparison, a set of transformer-based models is also tested (Tunstall et al., 2022). Across five independent runs for each model (using the

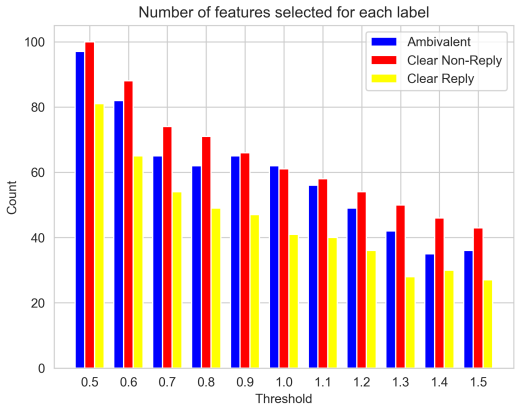

Figure 7: The number of features selected for every label, for thresholds ranging from 0.5 to 1.5, is decreasing. The proportion of features selected for each threshold and each label does not reflect the proportion of instances with different labels in the dataset (which are 'Ambivalent': 59.16%, 'Clear Non-Reply': 30.52%, 'Clear Reply': 10.32%). While this does not change the proportion of labels in the data, it does provide a more balanced set of terms.

best F_1 from each run), RoBERTa achieves the highest average performance, with a mean F_1 of 0.526 and a standard deviation of 0.024, indicating strong and relatively stable results. DistilBERT follows closely with a mean of 0.514 and a notably lower standard deviation of 0.013, making it the most consistent model despite slightly lower peak performance. BERT achieves a mean F_1 of 0.507 with a standard deviation of 0.023, indicating competitive but slightly less stable performance than DistilBERT. In contrast, ELECTRA significantly underperforms, with a mean F_1 of 0.368 and a standard deviation of 0.025, suggesting both lower effectiveness and comparable variability. Overall, the results indicate that RoBERTa offers the best trade-off between performance and robustness, while DistilBERT stands out for its consistency, and ELECTRA appears less suitable for this task in its current configuration. It is important to note that all models were trained using default hyperparameters, without any additional fine-tuning. The same applies to the traditional approaches employed, Random Forest and CatBoost. Moreover, unlike TF-IDF-based analyses, transformer-based architectures do not naturally provide interpretable outputs, such as explicit word-importance lists, thereby limiting direct feature-level interpretability in this setup.

The proposed DeltaSHAP approach has several limitations that should be acknowledged. First, its overall predictive performance remains modest, as reflected by the relatively low F_1 scores, which suggests that there is still room for improvement in this direction. Second, the approach is inherently dependent on TF-IDF representations, meaning it relies on surface-level lexical features rather than deeper contextual or semantic information. This dependence limits its ability to model nuanced language patterns that transformer-based models typically capture more effectively. Additionally, because the quality of the explanations is tied to the underlying feature space, any sparsity or bias in the TF-IDF representation directly affects the reliability of the extracted word importance signals. Finally, while DeltaSHAP provides interpretability advantages, this comes at the cost of reduced flexibility and scalability compared to more modern contextual embedding approaches, making it less suitable for tasks where high performance and rich semantic understanding are required.

Table 2 further lists distinct terms selected by DeltaSHAP for each label, as well as the common terms for $\delta = 1.4$. *Ambiguity* is characterized by general or context-dependent terms related to ongoing discussions or processes, while *Clear Non-Reply* and *Clear Reply* contain more specific conversational, institutional, or action-oriented expressions, with a small set of common terms shared across all categories indicating general discourse context. The common words shouldn't be very many, because it became harder for the classifier to predict the right label.

5 Conclusions and future work

DeltaSHAP proposes a game-theoretic approach to classify political statements into three distinct *clarity* labels, using the Shapley Value and a feature-adaptive, one-versus-all classification mechanism. What makes this approach different is that we use set differences, selecting specific words for each label that are useful for explaining classification and the data. By doing this, we can identify label-specific details for future analyses.

As further work, we aim to refine feature extraction techniques to improve predictive precision and explainability. We also propose extending this pipeline to other topics, such as political radicalization, and providing reliable methods for classifying radical and non-radical content, a growing concern in today's society. Also, we explore integrating transformer-based and LLM approaches with the

feature selection mechanism into a hybrid explainable model that can be further adapted to various specific topics.

Funding This research was supported by the project "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization and Financial Instruments Program, 2021-2027, MySMIS no. 351416.

References

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.

Kawin Ethayarajh and Dan Jurafsky. 2021. Attention Flows are Shapley Value Explanations. *arXiv preprint*. Version Number: 1.

Daniel Fryer, Inga Strumke, and Hien Nguyen. 2021. Shapley Values for Feature Selection: The Good, the Bad, and the Axioms. *IEEE Access*, 9:144352–144360.

Roni Goldshmidt and Miriam Horovicz. 2024. TokenSHAP: Interpreting Large Language Models with Monte Carlo Shapley Value Estimation. *arXiv preprint*. Version Number: 2.

Justin Grimmer and Brandon M. Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3):267–297.

W. E. Marcílio and D. M. Eler. 2020. From explanations to feature selection: assessing shap values as feature selection mechanism. In *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 340–347.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. A Dataset for Detecting Stance in Tweets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3945–3952, Portorož, Slovenia. European Language Resources Association (ELRA).

Edoardo Mosca, Ferenc Szigeti, and Stella Tragianni. SHAP-Based Explanation Methods: A Review for NLP Interpretability. ACL.

Roma Patel, Marta Garnelo, Ian Gemp, Chris Dyer, and Yoram Bachrach. 2021. Game-theoretic Vocabulary Selection via the Shapley Value and Banzhaf Index. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2789–2798, Online. Association for Computational Linguistics.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic Models for Analyzing and Detecting Biased Language. In *ACL*.

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.

Igor Trotskii, Amer Farea, and Frank Emmert-Streib. 2025. Comparative Analysis of Shapley Value-Based Feature Selection.

Lewis Tunstall, Leandro von Werra, and Thomas Wolf. 2022. *Natural Language Processing with Transformers: Building Language Applications with Hugging Face*. O'Reilly Media, Incorporated.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL workshop BlackboxNLP: analyzing and interpreting neural networks for NLP*, pages 63–76.