RES2 at IMPOLS 2026: Implicit Content Classification via Span-Awareness, Hierarchical Multi-Task Learning and Synthetic Data

Salvatore Spezia[1,†], Salvatore Ferrara[1,†], Roberto Dioguardi[1,†], Ernesto Davì[1,†], Irene Siragusa[1,*] and Roberto Pirrone[1]

[1]Department of Engineering, University of Palermo, Viale delle Scienze, Edificio 6, Palermo, 90128, Sicily, Italy

Abstract

In this report, we present the proposed approach of the RES2 team for the multimodal IMPOLS challenge at EVALITA 2026 campaign. We used a text-only approach based on UmBERTo Language Model, enhanced with a Span-Aware Architecture, which concatenates both global and local dense representations. Due to the scarcity of the data set, we implemented a Span-Preserving Data Augmentation with Gemini 3.0 and a Hierarchical Multi-Task Learning Strategy. The proposed system won in subtasks 2 and 3 and reached the 2nd place in subtask 1. In addition, we document the exclusion of the audio data, which resulted ineffective due to misalignment issues, speaker bias, and insufficient discriminative power.

Keywords

Span-Aware Transformers, Hierarchical Multi-Task Learning, Implicit Content Detection, Political Speech, Language Models

1. Introduction

Political discourse is a highly strategic field of human communication in which persuasion often relies on implicit rather than explicit meaning. The use of linguistic strategies such as presuppositions and implicatures enables speakers to convey questionable or manipulative content, defined as *non-bona fide* by bypassing the critical evaluation of the listeners and reducing their own assertive responsibility [1].

In the context of the EVALITA 2026 campaign [2], the Implicit Content in Political Speech Task (IMPOLS) [1] focuses on the development of automatic recognition systems for implicit content in Italian political speeches for both textual and audio data. The complexity of the task lies in the subtle nature of these mechanisms, as can be seen in the data set of the challenge extracted from IMPAQTS corpus [3].

Implicit content is not determined only by the words used, but by the interaction between a specific *local linguistic trigger* and the surrounding pragmatic context. The challenge is divided into three different subtasks, all evaluated in terms of macro F1-score, for binary detection of implicit content (subtask 1), binary classification tasks for implicatures and presuppositions (subtask 2) and implicature multi-class classification (subtask 3). For all three subtasks, the target segment which must be classified is annotated.

Recent progresses in Natural Language Processing (NLP) field are mainly led by generative Large Language Models (LLMs) but they struggle to capture fine-grained pragmatic nuances in zero-shot, few-shot, and Chain-of-Thought (CoT) settings, particularly in languages other than English [4]. Furthermore, the availability of annotated data for such specific phenomena is limited.

EVALITA 2026: 9th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Feb 26 – 27, Bari, IT

*Corresponding author.

†These authors contributed equally.

salvatore.spezia@community.unipa.it (S. Spezia); salvatore.ferrara05@community.unipa.it (S. Ferrara); roberto.dioguardi@community.unipa.it (R. Dioguardi); ernesto.davi@community.unipa.it (E. Davì); irene.siragusa02@unipa.it (I. Siragusa); roberto.pirrone@unipa.it (R. Pirrone)

0009-0005-8434-8729 (I. Siragusa); 0000-0001-9453-510X (R. Pirrone)

In this paper we present the system developed by the RES2 team for the three proposed subtasks in the IMPOLS challenge at EVALITA 2026, which employs a text-only strategy involving *ad hoc* fine-tuned version on UmBERTo [5] encoder-only Language Models (LMs). The core idea was to focus on both the global context and the local trigger to detect and classify the pragmatic ambiguities in the given speech.

The main contributions of this work can be summarized as follow:

1. A **Span-Aware Architecture** which concatenates the global representation of the input (Global Context Vector) with the target utterance representation (Local Target Vector), following a span-aware strategy;
2. A **Span-Preserving Data Augmentation Strategy** to mitigate the scarcity of samples for subtasks 2 and 3 with Gemini 3.0 [6], which paraphrases the only context while preserving the integrity of linguistic triggers;
3. A **Hierarchical Multi-Task Learning strategy** involving the hierarchical structure of the three tasks to lead the training of complex features through a weighted loss;
4. An analysis over the audio data, which highlights how the speaker bias tends to override useful prosodic signals, thereby justifying our final selection of a uni-modal approach.

The proposed system reached the 1st place in both subtasks 2 and 3 and the 2nd place in subtask 1, assessing the validity of the proposed approach against a general-purpose LLM.

This report is arranged as follows: Section 2 describes the proposed system along with data augmentation process, while details regarding the experimental setup are in Section 3. Obtained results and related discussion are reported in Section 4 and 5. Concluding remarks are in Section 6.

2. Description of the system

A unique transformers-based encoder-only model [7] has been developed for all the three subtasks. We opt for UmBERTo [5] as the LM backbone of the proposed system, which have been adapted to the target political speech domain, following a Layer-Wise fine-tuning strategy, differentiated for each subtask.

Given the pragmatic nature of the challenge, where implicit content is often activated by specific linguistic triggers within a broader context, we hypothesized that the standard `[CLS]` token alone would be insufficient to capture the necessary linguistic nuances. Thus, we propose a Span-Aware approach that combines the global semantic representation of the utterance with a focal representation of the target span. Furthermore, to mitigate data scarcity in subtask 3, we introduced an LLM-based Span-Preserving Data Augmentation strategy and a Hierarchical Multi-Task Learning framework.

The proposed Span-Aware Architecture (Figure 1) concatenates global and local feature, thus enhancing the capabilities of the overall system, which can be aware of both the target utterance and its context. In the proposed methodology, instead of relying exclusively on the features associated with the `[CLS]` token, we make the model explicitly Span-Aware. Specifically, the model is designed to be sensitive to a text portion delimited and annotated as the one which carries an implicit content, following the principle that the direct representation of spans enhances semantic modeling [8].

Formally, given the input sequence $X = \{x_1, \dots, x_n\}$, we add two special delimiter tokens `<TGT_START>` e `<TGT_END>` to mark the span target in the given textual input, following strategies based on explicit markers to highlight mentions or spans within transformer-based architectures [9]. Let $H \in \mathbb{R}^{n \times d}$ output of the last hidden state of the encoder, where d is the embedding size, which is equal to 768 for UmBERTo LM.

Then, we extract the two global and local embedding representations:

- **Global Context Vector** $\mathbf{h}_{cls}$ corresponds to the embedding of the `[CLS]` token (H_0);
- **Local Target Vector** $\mathbf{h}_{span}$ obtained through a mean pooling operation over the embeddings of the token enclosed between the previously defined delimiter tokens, as in [9];

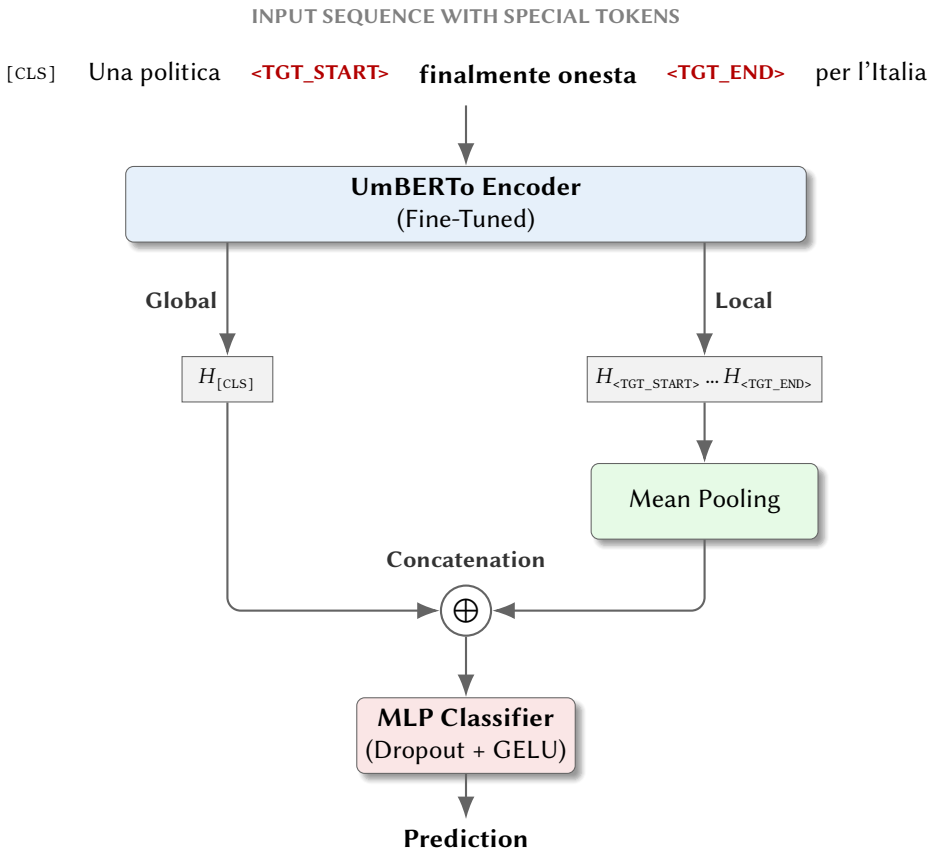

Figure 1: Architecture of the proposed Span-Aware model. The global [CLS] embedding $\mathbf{h}_{cls} \in \mathbb{R}^d$ is concatenated with the mean-pooled target-span representation $\mathbf{h}_{span} \in \mathbb{R}^d$ (tokens between <TGT_START> and <TGT_END>), producing $\mathbf{h}_{final} = [\mathbf{h}_{cls}; \mathbf{h}_{span}] \in \mathbb{R}^{2d}$, which is fed to an MLP classifier.

Thanks to a concatenation operation, the final representation is obtained as $\mathbf{h}_{final} = \mathbf{h}_{cls} \oplus \mathbf{h}_{span}$, which serves as an input to a Multi-Layer Perceptron (MLP) classifier which uses a dropout and a GELU [10] activation function to calculate the final prediction. On doing this, the model can use both the linguistic trigger and the overall context to calculate the final prediction.

In the context if the IMPOLS shared-task, target trigger span boundaries are provided by the organizers as part of the annotated input. Accordingly, at inference time our approach conditions on these gold spans by inserting the <TGT_START> and <TGT_END> markers, and we do not address the upstream problem of automatically localizing trigger spans. We are aware that this could be a limitation in real-case scenario of the proposed system. In this case, an addition target detection method should be included in the pipeline as pre-processing step on the textual input, before it is injected in the proposed system.

In the following subsections, we will deeply focus on the data augmentation phase (Sections 2.1) and the hierarchical multi-task learning strategy to address subtask 3 (Sections 2.2).

2.1. Span-Preserving Data Augmentation

Although the IMPOLS data set has balanced classes, the hierarchical nature of the challenge leads to a severe contraction in data volume as one progresses toward the more fine-grained subtasks. Starting from 1,824 samples in subtask 1, availability is halved to 912 for subtask 2 and is further reduced to only 456 examples for subtask 3.

This scarcity represents a critical bottleneck that exposes models to a high risk of overfitting. To mitigate this problem in the context of subtasks 2 and 3, we implemented a synthetic data augmentation pipeline using Gemini 3.0 [6]. On doing this, we incremented the cardinality of the training set and the

generalization capabilities of the model, making it robust to lexical variations.

In contrast to standard paraphrasing, we adopted a Span-Preserving approach, in which we ensured that the generative models do not change the trigger, with the following prompt:

> *Riscrivi il contesto e la frase circostante per variare lessico e sintassi, ma mantieni inalterato lo span target originale che attiva l'implicatura*

With this constraint we prevent the generative model from modifying the trigger, e.g. substitute *purtroppo* with *sfortunatamente* may change the implicit nature of the given sentence or its class [11]. Two variations have been created for each sample in training sets for subtasks 2 and 3, thus tripling the size of the available training sample.

An ablation study (Section 4.1) deeply explores the impact of the proposed data augmentation strategy and confirms its positive impact of performances over subtasks 2 and 3, with a particularly large improvement on subtask 3.

2.2. Hierarchical Multi-Task Learning

Limited to subtask 3, which is a multi-class classification task of particularized and generalized conversational, or conventional implicatures, we involved the inherent hierarchical and coarse-to-fine structure of the three subtasks in the IMPOLS challenge. Specifically, subtask 3 presupposes the distinction between implicature and presupposition (subtask 2), which in turn depends on identifying the presence of implicit content (subtask 1), consistent with a hierarchical classification framework [12].

Therefore, we adopted a Multi-Task Learning (MTL) approach using an hard parameter sharing strategy. We trained a single model with three independent classification heads (e.g. three different MLP Classifiers), jointly trained from scratch and sharing the same Span-Aware encoder. On doing this, an inductive transfer strategy is implemented, in which the supervision of correlated tasks guide the learning of more robust and generalizable dense representations [13].

Since the main focus was towards increasing the performances over subtask 3, subtasks 1 and 2 are used as auxiliary tasks, in an auxiliary-task learning approach [14]. In this case, the coarse grained tasks, semantic correlated to the target one, serve as stabilization factor during the training phase, and as improving factor for the overall performance. Hence, this strategy acts as a scaffolding structure for the effective learning of feature for the complex fine-grained classification in subtask 3.

Consequently, the loss function $\mathcal{L}_{tot}$ is defined as the weighted sum of the Cross-Entropy Loss of each subtask:

$$\mathcal{L}_{tot} = \lambda_1 \mathcal{L}_{task1} + \lambda_2 \mathcal{L}_{task2} + \lambda_3 \mathcal{L}_{task3}$$

We defined the following constraint $\lambda_3 \gg \lambda_1, \lambda_2$, to highlight the focus on subtask 3. Based on preliminary experiments on the validation set, we set the weights to $\lambda_1 = 1.5$, $\lambda_2 = 0.5$ e $\lambda_3 = 5.0$.

3. Experimental Setup

The data set of the IMPOLS challenge is derived from IMPAQTS [3], a multimodal corpus collecting Italian political speeches. Following a hierarchical division, the three perfectly balanced data sets for each task has been built.

We used `UmBERTo-commoncrawl-cased-v1` [5] as the LM backbone of the proposed system. UmBERTo is a LM based on RoBERTa [15] architecture and trained on an Italian web crawling corpus. We decided to use UmBERTo since (i) it employs SentencePiece [16] as tokenizer, which better manages the morphological complexity of the Italian language compared with standard WordPiece, and (ii) applies the Whole Word Masking during its pre-training phase, which preserves the semantic integrity of the concepts.

Since the semantic complexity of the task and the small training set, we opted for a Stratified K-Fold Cross-Validation, with K=5, instead of a fixed custom train-validation split. Within this strategy, the ensemble obtained model resulted to be much more stable and robust.

We adopted an hyperparameter optimization strategy, differentiated for each task: a classic grid search for subtask 1, while a bayesian optimization with Optuna [17] combined with a manual refinement for subtasks 2 and 3. At training time, our scope was towards the maximization of the macro F1-score. The final hyperparameters used for each subtask are summarized in Table 1.

Table 1
Overview of the used hyperparameters. Values marked with (*) were determined experimentally without automated search.

Hyperparameter	Subtask 1	Subtask 2	Subtask 3
Optimizer	AdamW*	AdamW*	AdamW*
Learning Rate	2×10^{-5}	1.31×10^{-4}	4.5×10^{-5}
Batch Size	16	8*	8*
Accumulation Step	-	2	2
Max Epochs	12*	15*	20*
Unfrozen Layers (Last N)	4	6	8
Dropout	0.20	0.15	0.25
Weight Decay	0.02	0.05	0.20
Label Smoothing	-	0.10	0.15
MTL Loss Weights ($\lambda_1, \lambda_2, \lambda_3$)	-	-	1.5, 0.5, 5.0
Hidden MLP Size	768*	128*	256

To adapt the LM to the target political speech domain, we used a Layer-Wise fine-tuning strategy, differentiated for each subtask. In particular, we trained the last 4 transformers layers for subtask 1, the last 6 layers for subtask 2 and the last 8 for subtask 3. The number of trained layers has been empirically found as the optimal balance between the number of training samples for each subtask and their complexity.

3.1. Training Details

Developed architectures were trained using Google Colab instances equipped with one GPU Tesla T4 and a local machine with NVIDIA GeForce RTX 5070 GPU. Models were implemented in PyTorch [18] incorporating stabilization techniques such as Mixed Precision (FP16), Early Stopping (with patience set to 3–5 epochs), and a linear scheduler with warm-up.

Starting from the developed and trained models, following a Stratified 5-Fold Cross-Validation strategy, the final model has been obtained as an aggregation of the 5 folds through Logit Averaging. For binary classification tasks, the final prediction is obtained by applying a sigmoid function to the mean of the logits and a 0.5 threshold:

$$\hat{y}_{bin} = \mathbb{I}\left(\sigma\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{z}_k\right) > 0.5\right)$$

A softmax function is applied for the multi-class classification task to the mean of the logits, and the predicted label is the one with the highest probability.

$$\hat{y}_{multi} = \text{argmax}\left(\text{softmax}\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{z}_k\right)\right)$$

where $K = 5$ and $\mathbf{z}_k$ is the logits vector of the k-th model. On doing this, we highly reduced the prediction variance among different folds.

4. Results

In Table 2 are reported the official result over the test set of the IMPOLS challenge [1].

Table 2
Official result over the test set (macro F1-score).

System	Subtask 1	Subtask 2	Subtask 3
baseline	0.4602	0.3352	0.2942
kenji-endo	**0.9496**	0.2226	-
RES2 (ours)	0.9089	**0.8762**	**0.6369**

In [1] a simple zero-shot prompt on Qwen2.5-7B-Instruct model [19] has been proposed as baseline for all three subtasks. The proposed RES2 system, overcome the baseline for all subtasks, reaching the 2nd place in the subtask 1, with a macro F1-score of 0.9089, resulting an improvement of 98% to the baseline model. As for subtasks 2 and 3, RES2 model reached the 1st place in the rank with almost tripling and doubling the scores obtained by the baseline, with 0.8762 and 0.6369 as macro F1-score. Obtained results assess the overall validity of the proposed architecture based on encoder-only models, along with the proposed MTL and data augmentation strategies.

During the development phase, we evaluated the performance of each fold over the corresponding validation set, which results are reported in Table 3. There can be highlighted the robustness of the proposed approach due to a small variance among different models, which reached its peak of $\sigma \approx 0.03$ for subtask 3, the most complex one.

Table 3
Average performance over the validation sets of the Stratified 5-Fold Cross-Validation

	Macro F1-score (mean ± std)
Subtask 1	0.9803 ± 0.0053
Subtask 2	0.8252 ± 0.0257
Subtask 3	0.7165 ± 0.0261

4.1. Ablation Study: Impact of Data Augmentation

To evaluate the impact of the proposed span-preserving data augmentation in subtasks 2 and 3, we trained the model with the only IMPOLS data and compared the obtained results with the performances of the model trained with the augmented data, while keeping unchanged the remaining hyperparameters.

Table 4
Ablation study on span-preserving data augmentation in subtasks 2 and 3 (macro F1-score), reporting both validation (Average Stratified 5-Fold CV) and official test results.

	Validation		Test	
Strategy	**Subtask 2**	**Subtask 3**	**Subtask 2**	**Subtask 3**
w/o augmentation	0.8015	0.5892	0.8666	0.4106
w/ augmentation	**0.8252**	**0.7165**	**0.8762**	**0.6369**
Δ (w/ − w/o)	+0.0237	+0.1273	+0.0096	+0.2263

Results reported in Table 4, show consistent improvements, with a particularly large gain on subtask 3, where data scarcity is most severe, thus assessing the positive impact of the proposed data augmentation strategy.

5. Discussion

In this section we will explore the impact of diverse encoder-only models as backbone of the developed Span-Aware architecture (Section 5.1) and its related architectural choices (Section 5.2), and our analysis of the audio data in the IMPOLS data set and our test within audio-only (Section 5.3) and multimodal models (Section 5.4).

5.1. Analysis over the LMs backbone

The proposed model is the best one developed during the training phase, in which we tested other LMs backbones with the proposed Span-Aware architecture. Specifically, we analyzed BERT-it cased (dbmdz/bert-base-italian-cased) [20], mDeBERTa-v3 (microsoft/mdeberta-v3-base) [21], XLM-RoBERTa (FacebookAI/xlm-roberta-base) [22] and GilBERTo (idb-ita/gilberto-uncased-from-camembert) [23], in addition to UmBERTo. With reference to the subtask 2, in Table 5 are reported the obtained results with the diverse LMs backbone under analysis.

Table 5
Comparisons of the performances of the LMs backbones over subtask 2 over the validation sets of the Stratified 5-Fold Cross-Validation. Bold values marks best obtained results.

Backbone Model	Macro F1-score (mean ± std)
BERT-it	0.7850 ± 0.0226
mDeBERTa-v3	0.8029 ± 0.0135
XLM-RoBERTa	0.8070 ± 0.0304
GilBERTo	0.8134 ± 0.0155
UmBERTo	**0.8252 ± 0.0257**

The analysis and the obtained results show that UmBERTo lightly overpass both multilingual models (mDeBERTa-v3 and XLM-RoBERTa) and Italian-only models (GilBERTo and BERT-it), thus it was selected our LM backbone.

5.2. Analysis over the Span-Aware architecture

The final developed method involved the simple concatenation $\mathbf{h}_{cls} \oplus \mathbf{h}_{span}$, we also explored other methodologies involving attention pooling and a self-attention or cross-attention layer to merge the two vector representation. This resulted in lower performances in the training set (macro F1-score of 0.73 for subtask 2) and an unnecessary computational complexity.

Another strategy was to consider $\mathbf{h}_{cls}$ and $\mathbf{h}_{span}$ as distinct feature maps which can be combined with a Convolutional layer (CNN). This idea was promising and competitive in subtask 1, reaching a score in the training set was equal to 0.9562, but still lower compared with the simple concatenation (0.9803).

In conclusion, comparing the results obtained during the development phase (Table 3) with the official competition results (Table 2), it is possible notice that the proposed system does not suffer from overfitting, despite the highest results with the training set. We assume that its generalization capabilities reside in both the system architecture and training strategy implemented, comprehensive of the proposed data augmentation strategy. The simple combination of UmBERTo as the LM backbone and a basic feature concatenation, resulted to be more effective than other methodologies, which may require additional computational resources and an increasing number of parameters.

While conditioning on gold trigger spans is reasonable in the IMPOLS shared-task scenario, it limits direct applicability in real-world settings where such annotations are unavailable. In practical deployments, the proposed classifier would need to be preceded or jointly coupled with an automatic trigger-span detection component. Assessing robustness under automatically predicted spans is left for future work.

5.3. Analysis of the multimodal data set

The multimodal nature of the IMPAQTS data set [3] offered, at least theoretically, the opportunity to leverage paralinguistic cues, such as prosody, intonation and pauses, to disambiguate implicit content. However, before finalizing the text-only architecture described in Section 2, we conducted an exploratory analysis and a set of audio-only probes on the acoustic modality.

A preliminary inspection revealed a non-negligible misalignment between the audio tracks and the provided transcriptions. Through a semi-automated verification procedure based on segment duration, we excluded 110 samples (approximately 6% of the original data set) exhibiting a complete mismatch. This reduction introduces an initial limitation for the audio modality compared to text, which is instead fully preserved.

To quantify the contribution of audio, we extracted 345 acoustic features for each trigger-associated segment, combining prosodic descriptors (e.g., F0, jitter, shimmer) via Praat [24], spectral features (e.g., MFCC and spectral centroid) via Librosa [25], and the standard eGeMAPSv02 set [26] using OpenSMILE [27]. Moreover, to mitigate inter-speaker variability, in addition to the speaker-level normalization, we adopted an intra-speaker local baseline comparison. For each trigger segment we consider the same-duration baseline window within the same recording, and we define delta features Δ as the difference between segment and baseline features. This representation aims to subtract speaker-specific prosodic baselines and capture local acoustic variations around the trigger.

To assess speaker bias and estimate the actual discriminative power of acoustic features with respect to the task labels, we trained a Random Forest classifier under two 5-cross-validation schemes. Specifically we considered (i) a stratified K-Fold (stratified CV), consistent with the text-only development setting, and (ii) Speaker-Aware K-Fold (speaker CV), ensuring that all segments from the same speaker fall within the same fold. Table 6 reports macro-F1 scores obtained using both raw segment features and Δ features.

Table 6
Audio-only Random Forest probes (macro F1): comparison between Stratified CV and Speaker CV, using raw segment features and delta features Δ (segment − baseline).

	Segment (raw)		**Delta (Δ)**	
	Stratified CV	**Speaker CV**	**Stratified CV**	**Speaker CV**
Subtask 1	**0.742**	0.529	0.553	0.511
Subtask 2	**0.646**	0.494	0.547	0.558
Subtask 3	**0.568**	0.391	0.405	0.324

The results highlight two complementary patterns. On the one hand, Δ features yield overall low performance, especially on subtask 3, indicating limited discriminative capacity with respect to pragmatic classes. Although they capture systematic local shifts relative to the baseline (e.g., short-term changes in voicing/pauses and energy), such variations are not sufficiently informative to robustly separate the target categories.

On the other hand, raw segment features achieve higher scores under stratified CV, but drop sharply under grouped CV, revealing substantial speaker leakage. The model tends to exploit speaker-idiosyncratic traits, such as timbre, prosodic range, articulation habits, rather than generalizable prosodic cues linked to the pragmatic phenomenon. Overall, these findings explain why, despite measurable local variations around triggers, the audio modality provides limited benefits for pragmatic classification and remains vulnerable to speaker-specific bias, thus motivating our final choice of a text-only approach, which outperforms the audio-only approaches (Table 6).

5.4. Analysis over the developed multimodal approaches

Before finalizing the proposed text-only architecture, we investigated several multimodal approaches to test the hypothesis that paralinguistic features could resolve pragmatic ambiguities.

The audio pipeline has been built on Wav2Vec 2.0 [28] in its large-xlsr-53-italian version, preferred over WavLM [29] for its better phonetic sensibility towards the Italian language. To implement the proposed Span-Aware approach, we isolated the acoustic component associated with the trigger using WhisperX [30] to implement a forced alignment at word-level. Additionally, we refined the temporal alignment through fuzzy string matching to compensate for transcription discrepancies.

The integration between local and global representation has been implemented with cross-attention. More in detail, two paradigms have been investigated:

1. **Cross-Modal Attention Network** we used a cross-attention mechanism in which textual representation from UmBERTo serves as Query, while audio feature extracted from Wav2Vec and enriched with prosodic descriptors, serves as Key and Value, as in Multimodal Transformer (MulT) [31]. This approach, similar to CM-BERT [32] and CCMT [33], dynamically manages the temporal misalignment.
2. **MTAMW Re-implementation** we implemented the Multimodal Transformer with Adaptive Modality Weighting (MTAMW) architecture [34], designed to adaptively weight the importance of both modalities. Unfortunately, obtained results were comparable to, or slightly lower than, those achieved with standard cross-attention.

Despite the architectural sophistication, empirical results demonstrated the ineffectiveness of the acoustic component for this specific task. A simple audio-only model based on a fine-tuned version of Wav2Vec 2.0 reached a macro F1-score of only 0.5917 on subtask 1. This result over the Stratified CV split, resulted significantly lower compared with text-only approaches, on the same split. On the other hand, the multimodal architecture involving both modalities and a cross-attention merging strategy, obtained a macro-F1 score of 0.9636 in subtask 1. Notwithstanding the promising results, the removal of the audio component at inference time, led to no significant changes in the final result.

These experiments showed the strength of the UmBERTo textual encoder in capturing the semantic context, and, by contrast, the additional contribution provided by the audio, results orthogonal but also redundant and noisy. The orthogonality is given by the diverse nature of information acquired and coded in audio signals compared with the textual ones, while redundancy and noisiness are provided by the alignment unreliability and speaker bias. Overall, despite their potentialities, audio features resulted to be not significantly discriminative: therefore, to maximize the computational efficiency and the robustness of the entire system, we opted for a text-only model (Section 2).

6. Conclusion

In this work we presented the proposed model by RES2 team for the IMPOLS challenge in the context of EVALITA 2026 campaign. The developed system won the competition for subtasks 2 and 3 and reached the 2nd place for the 1st subtask. The strengths of the proposed system rely on the Span-Awareness strategy, which combines both the target trigger with its context using UmBERTo to extract embeddings. The definition and usage of a Span-Preserving Data Augmentation with Gemini 3.0, allows us to increment the size of the training set, providing also more stability and robustness to the final trained model. In addition, the proposed pipeline confirms that, for some tasks in specialized domains, the usage of fine-tuned encoder-only models can overcome generalist models such as generative LLMs.

Moreover, the analysis on the audio data and its usage within the textual transcription in a multimodal setup, highlighted the presence of a severe speaker bias. In these cases, the extracted acoustic features guide the models towards the identification of the speaker, rather than focusing on prosodic features of the implicature. On doing this, the audio signal results redundant or noisy relative to the text, and its usage in a multimodal architecture is limited.

In future works we will focus on a better fine-grained audio-text alignment to isolate the local prosodic variation in correspondence to the triggers, as well as in developing an automatic trigger detection methods. Another interesting aspect will be the analysis of performance of generative LLMs, e.g. with the adoption of Retrieval-Augmented Generation (RAG) techniques. With this strategy,

it would be possible to enhance the context to provide to the model with encyclopedic information regarding speakers, political affiliations, and socio-historical backgrounds. This direction is particularly relevant when implicitness is not fully recoverable from a single lexical trigger, but emerges from broader contextual and world-knowledge cues typical of political discourse. Furthermore, we envision the implementation of fine-tuning based on reasoning paths to inject explicit pragmatic explanations that justify the nature of implicit content. We aim that this strategy may enhance logical inference capabilities of the model, along its precision in discriminating between complex linguistic categories.

Declaration on Generative AI

During the preparation of this work, the author(s) used Gemini 3 in order to: Grammar and spelling check, Text Translation. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] L. Gregori, W. Paci, V. Saccone, IMPOLS at EVALITA 2026: Overview of the IMplicit contents in POLitical Speech Task, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.

[2] F. Cutugno, A. Miaschi, A. P. Aprosio, G. Rambelli, L. Siciliani, M. A. Stranisci, Evalita 2026: Overview of the 9th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.

[3] G. Comandini, G. Ferrari, A. Panunzi, Impaqts: A multimodal corpus of implicit manipulatory strategies in political speech, in: Proceedings of the Ninth Italian Conference on Computational Linguistics (CLiC-it 2023), CEUR-WS.org, Venice, Italy, 2023. URL: https://ceur-ws.org/Vol-3596/.

[4] W. Paci, A. Panunzi, S. Pezzelle, They want to pretend not to understand: The limits of current llms in interpreting implicit content of political discourse, in: Findings of the Association for Computational Linguistics: ACL 2025, Association for Computational Linguistics, Vienna, Austria, 2025, pp. 15569–15593. doi:10.18653/v1/2025.findings-acl.804.

[5] L. Parisi, S. Francia, P. Magnani, Umberto: an italian language model trained with whole word masking, https://github.com/musixmatchresearch/umberto, 2020.

[6] Google DeepMind, Gemini 3.0: A new era of intelligence, 2025. URL: https://deepmind.google/gemini, technical Report. Retrieved from https://deepmind.google/gemini.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is All you Need, in: Advances in Neural Information Processing Systems, 2017.

[8] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, O. Levy, SpanBERT: Improving pre-training by representing and predicting spans, Transactions of the Association for Computational Linguistics 8 (2020) 64–77. URL: https://aclanthology.org/2020.tacl-1.5/. doi:10.1162/tacl_a_00300.

[9] L. Baldini Soares, N. FitzGerald, J. Ling, T. Kwiatkowski, Matching the blanks: Distributional similarity for relation learning, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 2895–2905. URL: https://aclanthology.org/P19-1279/. doi:10.18653/v1/P19-1279.

[10] D. Hendrycks, K. Gimpel, Gaussian error linear units (gelus), 2023. URL: https://arxiv.org/abs/1606.08415. arXiv:1606.08415.

[11] D. I. Beaver, B. Geurts, Presupposition, Stanford Encyclopedia of Philosophy, 2011. URL: https://plato.stanford.edu/entries/presupposition/, first published 2011; substantive revision 2021.

[12] C. N. Silla, A. A. Freitas, A survey of hierarchical classification across different application domains, Data Mining and Knowledge Discovery 22 (2011) 31–72.

[13] R. Caruana, Multitask learning, Machine Learning 28 (1997) 41–75.
[14] L. Liebel, M. Körner, Auxiliary tasks in multi-task learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2018, pp. 1–10.
[15] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, et al., RoBERTa: A Robustly Optimized BERT Pretraining Approach, arXiv preprint arXiv:1907.11692 (2019).
[16] T. Kudo, J. Richardson, Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 66–71. doi:10.18653/v1/D18-2012.
[17] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, Anchorage, AK, USA, 2019, pp. 2623–2631. doi:10.1145/3292500.3330701.
[18] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, et al., Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems, volume 32, Curran Associates, Inc., 2019, pp. 8024–8035. URL: https://proceedings.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.
[19] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, et al., Qwen Technical Report, arXiv preprint arXiv:2309.16609 (2023).
[20] S. Schweter, Italian bert and electra models, 2020. URL: https://github.com/dbmdz/berts. doi:10.5281/zenodo.4271493.
[21] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, in: Eleventh International Conference on Learning Representations (ICLR), 2023. URL: https://openreview.net/forum?id=sE7-XhLxHA.
[22] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, et al., Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451.
[23] D. Atzeni, M. Guredda, A. Moretti, Gilberto: a pre-trained language model for italian, 2020. URL: https://github.com/idb-ita/GilBERTo.
[24] P. Boersma, D. Weenink, Praat: doing phonetics by computer [computer program], 2024. URL: http://www.praat.org/, retrieved from http://www.praat.org/.
[25] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, O. Nieto, librosa: Audio and music signal analysis in python, in: K. Huff, J. Bergstra (Eds.), Proceedings of the 14th Python in Science Conference, 2015, pp. 18–25. doi:10.25080/Majora-7b98e3ed-003.
[26] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, et al., The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing, IEEE Transactions on Affective Computing 7 (2016) 190–202. doi:10.1109/TAFFC.2015.2457417.
[27] F. Eyben, M. Wöllmer, B. Schuller, opensmile: The munich versatile and fast open-source audio feature extractor, in: Proceedings of the 18th ACM International Conference on Multimedia, ACM, Firenze, Italy, 2010, pp. 1459–1462. doi:10.1145/1873951.1874246.
[28] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems (NeurIPS), volume 33, 2020, pp. 12449–12460. URL: https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf.
[29] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, et al., Wavlm: Large-scale self-supervised pre-training for full stack speech processing, IEEE Journal of Selected Topics in Signal Processing 16 (2022) 1505–1518. doi:10.1109/JSTSP.2022.3188113.
[30] M. Bain, J. Huh, T. Han, A. Zisserman, Whisperx: Time-accurate speech transcription of long-form audio, in: INTERSPEECH 2023, ISCA, 2023, pp. 148–152. URL: https://doi.org/10.21437/Interspeech.2023-78. doi:10.21437/Interspeech.2023-78.
[31] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, R. Salakhutdinov, Multimodal trans-

former for unaligned multimodal language sequences, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 6558–6569. doi:10.18653/v1/P19-1656.

[32] K. Yang, H. Xu, K. Gao, Cm-bert: Cross-modal bert for text-audio sentiment analysis, in: Proceedings of the 28th ACM International Conference on Multimedia, ACM, Seattle, WA, USA, 2020, pp. 521–528. doi:10.1145/3394171.3413690.

[33] N.-C. Ristea, A. Anghel, R. T. Ionescu, Cascaded cross-modal transformer for audio-textual classification, arXiv preprint arXiv:2401.07575 (2024). URL: https://arxiv.org/abs/2401.07575.

[34] Y. Wang, J. He, D. Wang, Q. Wang, B. Wan, X. Luo, Multimodal transformer with adaptive modality weighting for multimodal sentiment analysis, Neurocomputing 572 (2024) 127181. doi:10.1016/j.neucom.2023.127181.