Baseline Methods for the Missing Pieces and Misinformation Task of MediaEval 2026

Martial Pastor[1,*], Stefan Kok[2]

[1]Centre for Language Studies, Radboud University, Nijmegen, The Netherlands
[2]Institute for Computing and Information Sciences, Radboud University, Nijmegen, The Netherlands

Abstract

This paper presents the baseline methods for the Missing Pieces and Misinformation task at MediaEval 2026. The task involves detecting enthymemes—arguments with an unstated premise or conclusion—in tweets about Covid and Immigration (Tasks 1A and 1B), and generating the missing component as a natural language sentence for enthymematic tweets (Task 2). For enthymeme detection, we establish lexical baselines using TF-IDF representations paired with a Support Vector Classifier and Logistic Regression, and fine-tune DeBERTa-v3-base with hard majority-vote labels. For implicit component generation, we prompt a pre-trained T5-base model without any task-specific fine-tuning to establish a lower bound for Task 2. Results confirm that neural models outperform lexical baselines on classification, and that generation quality can be meaningfully assessed via ROSCOE-SS semantic similarity scoring.

1. Introduction

Implicit argumentation is pervasive in online discourse and plays a central role in the spread of misinformation [1]. When a proposition is left unstated, readers must reconstruct it themselves, leading them to perceive the implied message as their own inference and therefore to accept it more readily [2]. This makes enthymemes—arguments where either a premise or a conclusion is left implicit—a particularly effective vehicle for persuasion and misinformation in social media.

The Missing Pieces and Misinformation shared task provides a dataset of 1,483 tweets on Covid and Immigration, each annotated by five independent annotators [3]. Task 1A is binary classification (`enthymeme`/`none`); Task 1B extends this to three classes (`implicit_premise`, `implicit_conclusion`, `none`); Task 2 requires generating the missing component as a single natural language sentence for tweets identified as enthymemes. Performance on Tasks 1A and 1B is measured using macro-F1 on hard majority-vote labels and cross-entropy loss against soft labels derived from the full annotation distribution [4, 5]; Task 2 outputs are evaluated for semantic similarity to gold reconstructions using ROSCOE-SS [6].

This paper presents publicly released baselines intended to assist participants and to anchor the evaluation space. All implementation details will be made available alongside the dataset release.

MediaEval'26: Multimedia Evaluation Workshop, June 15–16, 2026, Amsterdam, Netherlands and Online
*Corresponding author.
martial.pastor@ru.nl (M. Pastor); stefan.kok@ru.nl (S. Kok)

2. Method

2.1. Enthymeme Detection (Tasks 1A and 1B)

Lexical Baselines. We evaluate TF-IDF representations with unigram and bigram features using a Linear Support Vector Classifier (SVC) [7] and Logistic Regression (LR) [8]. These models serve as surface-level baselines that operate purely on lexical co-occurrence statistics, without any contextual representation.

DeBERTa Fine-tuning. We fine-tune DeBERTa-v3-base [9] on both Task 1A (binary) and Task 1B (three-class) using majority-vote hard labels with class-weighted cross-entropy loss and layer-wise learning rate decay.[1] Class weighting compensates for the moderate label imbalance in the dataset. Model selection is performed via 5-fold cross-validation on the development set (n=1333); final evaluation is conducted on a held-out test set (n=148).

2.2. Implicit Component Generation (Task 2)

T5-base Prompt Baseline. We include a T5-base baseline [10] to establish a lower bound for Task 2. The pre-trained T5-base checkpoint is used directly, with no task-specific fine-tuning, to assess what a general-purpose sequence-to-sequence model can produce without any adaptation to the enthymeme domain.

T5-base is a 12-layer encoder-decoder Transformer with hidden dimension 768, feed-forward size 3,072, 12 attention heads, and a vocabulary of 32,128 SentencePiece tokens, totalling approximately 220M parameters. For each tweet in the test set whose majority-vote label is `implicit_premise` or `implicit_conclusion` (51 of 148 tweets), the model is prompted with the template:

```
generate implicit {label}: {tweet}
```

where `{label}` is the gold majority-vote label. Generation uses beam search (beam size 4, max 64 new tokens). The remaining 97 tweets with majority label `none` receive no generation. No development set tuning was performed; pre-trained weights are used as-is.

3. Results

3.1. Task 1A: Binary Enthymeme Detection

TF-IDF+SVC collapses to majority-class prediction on the test set (enthymeme F1 = 0.000), confirming that surface lexical patterns are insufficient for detecting enthymematic content. TF-IDF+LR recovers some enthymeme cases but remains substantially below the neural model. DeBERTa achieves the best macro-F1 on both partitions (dev: 0.676; test: 0.647), with balanced improvement across both classes. The dev-to-test gap is small and stable, suggesting that stratified sampling yields representative partitions.

We also report mean cross-entropy loss against soft labels derived from the annotation distribution (lower is better; 0 = perfect match to the soft distribution). The soft label for each instance is a normalized frequency vector over annotator judgments: e.g., if 3 of 5 annotators label an instance `enthymeme` and 2 label it `none`, the soft label is $\mathbf{q} = [0.6, 0.4]$, and loss is $\mathcal{L}_{\text{CE}} = -\sum_i q_i \log p_i$.

[1] 5 epochs, batch size 16, max sequence length 128, peak learning rate 2e-5, linear warmup over 10% of steps, layer-wise decay factor 0.9.

			None			Enthymeme		
Model	**Acc**	**Mac-F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Development (5-fold CV, n=1333)								
TF-IDF + SVC	0.667	0.425	0.667	0.993	0.798	0.684	0.029	0.055
TF-IDF + LR	0.658	0.629	0.759	0.708	0.732	0.495	0.561	0.526
DeBERTa	0.694	**0.676**	0.805	0.706	0.752	0.543	0.671	0.600
Test (held-out, n=148)								
TF-IDF + SVC	0.642	0.391	0.651	0.979	0.782	0.000	0.000	0.000
TF-IDF + LR	0.601	0.575	0.716	0.649	0.681	0.433	0.510	0.469
DeBERTa	0.669	**0.647**	0.773	0.701	0.735	0.517	0.608	0.559

Table 1
Task 1A results (binary classification). Best macro-F1 per partition in **bold**.

Model	**Dev CE Loss**	**Test CE Loss**
TF-IDF + SVC	1.243	1.318
TF-IDF + LR	0.874	0.921
DeBERTa	**0.621**	**0.658**

Table 2
Task 1A cross-entropy loss against soft annotation distribution. Lower is better.

DeBERTa yields the lowest cross-entropy loss across both partitions, indicating that its predicted probability distributions are closer to the empirical annotation distribution than those of the lexical baselines. The gap between SVC and DeBERTa is particularly pronounced: SVC's near-constant majority-class predictions assign near-zero probability to the minority class, incurring high loss on instances where annotators were genuinely divided.

3.2. Task 1B: Three-Class Classification

			None		Impl. Premise		Impl. Conclusion	
Model	**Acc**	**Mac-F1**	**R**	**F1**	**R**	**F1**	**R**	**F1**
Development (5-fold CV, n=1333)								
TF-IDF + SVC	0.651	0.312	0.991	0.787	0.021	0.041	0.000	0.000
TF-IDF + LR	0.632	0.521	0.714	0.720	0.487	0.471	0.381	0.373
DeBERTa	0.671	**0.598**	0.718	0.741	0.584	0.542	0.513	0.512
Test (held-out, n=148)								
TF-IDF + SVC	0.628	0.281	0.974	0.771	0.000	0.000	0.000	0.000
TF-IDF + LR	0.608	0.492	0.651	0.681	0.451	0.427	0.373	0.368
DeBERTa	0.648	**0.571**	0.694	0.718	0.549	0.511	0.471	0.484

Table 3
Task 1B results (three-class classification). Best macro-F1 per partition in **bold**.

Task 1B is substantially harder than Task 1A: macro-F1 drops across all models, and the distinction between `implicit_premise` and `implicit_conclusion` proves particularly challenging for lexical baselines. TF-IDF+SVC again degrades to near-majority prediction on

the test set. DeBERTa achieves the strongest performance, though recall on implicit conclusion remains low, pointing to an inherent difficulty in surface-level signals that discriminate premise from conclusion roles. Cross-entropy loss against the soft three-class distribution (e.g., $\mathbf{q} = [0.6, 0.2, 0.2]$ for a 3/1/1 annotation split) follows the same ordering as Task 1A, with DeBERTa achieving the best calibration.

3.3. Task 2: Missing Component Generation

Model	n (matched)	ROSCOE-SS
T5-base (no fine-tuning)	50	0.8236

Table 4
Task 2 generation results. ROSCOE-SS computed over matched instances with at least one gold reconstruction.

Of the 51 test tweets with a majority-vote label of `implicit_premise` or `implicit_conclusion`, 50 have at least one available gold reconstruction (211 reference texts in total, 3–5 per tweet). For each such instance, ROSCOE-SS [6] is computed using `facebook/roscoe-512-roberta-base`, a RoBERTa-base model fine-tuned with a contrastive SimCSE objective to produce calibrated sentence embeddings suited to reasoning content. The score for a single instance is obtained by encoding both the generated proposition and each gold reference via mean-pooling over the model's last hidden states, computing pairwise cosine similarities, mapping each to $[0, 1]$ via $\frac{1+\cos(\mathbf{p},\mathbf{r})}{2}$, and averaging over all available references. The final reported score is the mean of these per-instance scores over the 50 matched test tweets.

The T5-base prompt baseline achieves a ROSCOE-SS score of 0.8236. Given that this model receives no task-specific training and operates solely through a template prompt, this score constitutes a reasonable lower bound. It also reflects the semantic regularity of the task: many enthymemes in the dataset involve relatively predictable stances on Covid or immigration, and a general-purpose model with broad pretraining can approximate the propositional content of missing components without dedicated supervision. Participants with fine-tuned generation models should aim to exceed this threshold.

4. Conclusion

We have presented lexical and neural baselines for the Missing Pieces and Misinformation task at MediaEval 2026. For enthymeme detection (Tasks 1A and 1B), DeBERTa fine-tuned on hard majority-vote labels consistently outperforms TF-IDF-based baselines across accuracy, macro-F1, and cross-entropy loss against soft annotation distributions. The three-class task (1B) is considerably harder, with premise/conclusion disambiguation posing a particular challenge for surface-level models. For implicit component generation (Task 2), an untuned T5-base prompt baseline achieves a ROSCOE-SS score of 0.8236 over matched test instances, establishing a lower bound that fine-tuned systems should seek to exceed. All code and data splits will be released publicly to support participant development.

Acknowledgments

This work was produced as part of the HYBRIDS project, a Marie Skłodowska-Curie Doctoral Network funded by the European Union under grant no. 101073351 and UKRI Horizon Funding Guarantee, and the AI-CODE project under Horizon Europe grant agreement no. 101135437.

Disclaimer

The dataset contains social media posts that may include offensive or controversial language, included solely for scientific research purposes and not reflecting the views of the task organizers.

References

[1] E. Lombardi Vallauri, L. Baranzini, D. Cimmino, F. Cominetti, C. Coppola, G. Mannaioli, Implicit argumentation and persuasion: A measuring model, Journal of Argumentation in Context 9 (2020) 95–123. doi:10.1075/jaic.00009.lom, publisher: John Benjamins.

[2] A. Reboul, A relevance-theoretic account of the evolution of implicit communication, Studies in Pragmatics 13 (2011) 1–19.

[3] A. Flaccavento, Y. Peskine, P. Papotti, R. Torlone, R. Troncy, Automated detection of tropes in short texts, Proceedings of the 31st International Conference on Computational Linguistics (2025) 5936–5951. Abu Dhabi, UAE. Association for Computational Linguistics.

[4] B. Plank, D. Hovy, A. Søgaard, Linguistically debatable or just plain wrong?, Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (2014) 507–511. Baltimore, Maryland. Association for Computational Linguistics.

[5] A. N. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, Learning from disagreement: A survey, Journal of Artificial Intelligence Research 72 (2021) 1385–1470. doi:10.1613/jair.1.12752.

[6] O. Golovneva, M. Chen, S. Poff, M. Corredor, L. Zettlemoyer, M. Fazel-Zarandi, A. Celikyilmaz, ROSCOE: A suite of metrics for scoring step-by-step reasoning (2023). URL: https://openreview.net/forum?id=xYlJRpzZtsY.

[7] C. Cortes, V. Vapnik, Support-vector networks, volume 20, 1995, pp. 273–297. doi:10.1007/BF00994018.

[8] D. R. Cox, The regression analysis of binary sequences, Journal of the Royal Statistical Society: Series B (Methodological) 20 (1958) 215–232. doi:10.1111/j.2517-6161.1958.tb00292.x.

[9] P. He, X. Liu, J. Gao, W. Chen, DeBERTa: Decoding-enhanced BERT with disentangled attention, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=XPZIaotutsD.

[10] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67. URL: http://jmlr.org/papers/v21/20-1307.html.