Does Annotator Disagreement Signal Enthymeme Presence? Leveraging Inter-Annotator Uncertainty for Implicit Proposition Detection

Afifah Luqman[1] · Hadiya Ebrahim[1]
[1]*FAST-NUCES, Karachi, Pakistan*

Abstract

We present our participation in the MediaEval 2026 Missing Pieces and Misinformation task, which targets the detection and reconstruction of enthymemes —implicit premises and conclusions — in political tweets. We investigate whether inter-annotator disagreement constitutes a reliable signal for enthymeme presence. Our approach fine-tunes DeBERTa-v3-base for 3-class classification (Run 1), and augments a second run with disagreement-derived features fused directly into the encoder (Run 2). On the official test set, Run 1 achieves the best binary macro F1 (0.6598), the primary competition metric. Disagreement features (Run 2) improve premise detection (F1: 0.5470) but do not generalise to the rare implicit_conclusion class (F1: 0.0000) on the official test, despite showing a stronger result on our internal validation split — a limitation attributable to the very small number of conclusion examples (n=6) in the test set. For proposition generation, we evaluate Flan-T5-large and Mistral-7B-Instruct using ROSCOE-SS, with matched-only scores of 0.8710 and 0.8826 respectively.

Index Terms – enthymeme detection, implicit argument mining, annotator disagreement, transformer fine-tuning, proposition generation

1. INTRODUCTION

Enthymemes — arguments with one or more unstated propositions — are pervasive in political discourse and pose a significant challenge for automated misinformation analysis. The MediaEval 2026 Missing Pieces and Misinformation task [1] frames this as two sub-problems: (Task 1) classifying tweets as containing an implicit premise, an implicit conclusion, or neither; and (Task 2) generating the missing proposition as a natural-language sentence.

A defining property of this dataset is that human annotators frequently disagree. Rather than treating disagreement as noise to be resolved by majority voting, our core research question asks: does inter-annotator disagreement itself signal enthymeme presence? We hypothesise that tweets which are harder to classify for human annotators — evidenced by higher label entropy — are precisely those containing implicit argumentation that requires background knowledge to detect. This Quest-for-Insight paper is structured around that hypothesis.

2. RELATED WORK

Enthymeme detection has been studied as a component of computational argumentation. Stahl et al. [2] introduced the first shared task on automated enthymeme detection and reconstruction in learner arguments, framing both gap detection and gap filling as distinct subtasks. Sviridova et al. [3] survey implicit argument mining more broadly, noting that detection of whether an implicit component is present — as opposed to merely reconstructing it — remains underexplored, particularly in social media contexts where informal rhetorical structures dominate.

The role of annotator disagreement in NLP has received growing attention. Plank [4] argues that inter-annotator disagreement encodes genuine label uncertainty and should be modelled rather than discarded throughout the NLP pipeline. This perspective has been applied to tasks such as offensive language detection and stance analysis [5], where majority-vote labels mask meaningful variation. Our work applies this directly to enthymeme detection, asking whether disagreement features can substitute for richer linguistic resources unavailable in constrained settings.

DeBERTa-v3 [6] has demonstrated strong performance on NLU tasks requiring fine-grained semantic understanding. We adopt it as our encoder due to its disentangled attention mechanism, which is well-suited for capturing the pragmatic cues associated with implicit argumentation.

3. APPROACH

3.1 Task 1 — Classification

Run 1 fine-tunes DeBERTa-v3-base on tweet text alone as a 3-class sequence classifier (none, implicit_premise, implicit_conclusion). We use weighted cross-entropy with label smoothing ($\alpha = 0.1$) to handle class imbalance, a stratified 70/15/15 train/validation/test split, and per-class threshold tuning on the validation set to maximise macro F1. Training uses AdamW (LR = 2e-5) with gradient accumulation (steps = 2) and gradient clipping (max norm = 1.0).

Run 2 augments the encoder with a 6-dimensional disagreement feature vector derived from the provided multi-annotator labels. The features are: (1) label entropy across all annotators, (2) number of unique labels assigned, (3–5) per-class vote proportions, and (6) a difficulty tier based on label cardinality. These features are concatenated with the DeBERTa [CLS]

embedding and passed through a fusion layer (Linear(774 → 256), ReLU, dropout 0.2) before the final classifier. This architecture allows the model to learn how disagreement correlates with label difficulty without requiring external data, keeping the run fully constrained. Critically, at inference on the official test set — where annotator labels are unavailable — the feature vector is set to a neutral value representing uniform vote distribution and zero entropy, and the training scaler is applied before inference.

Run 3 employed a cascade architecture implemented as two independent DeBERTa binary classifiers. Model A distinguishes none from enthymeme (trained on all data); Model B distinguishes implicit_premise from implicit_conclusion (trained only on enthymeme rows). At inference, Model A first filters tweets, and Model B sub-classifies those predicted as enthymematic. Run 3 Constrained uses text-only models; Run 3 Open augments both models with the same 6-dimensional disagreement feature vector as Run 2, fused via a deeper 3-layer FFN: Linear(774 → 512) → LayerNorm → GELU → Dropout(0.3) → Linear(512 → 256) → GELU → Dropout(0.2) → Linear(256 → n_classes). Four loss variants were evaluated on Model A: standard weighted CrossEntropy, AgreementWeightedCE (upweighting unanimous annotator samples), FocalLoss ($\gamma = 0.5$), and AgreementWeightedFocal. The best-performing variant from the constrained run was reused for the open run. Task 2 generation for Run 3 uses the cascade predictions as input labels.

3.2 Task 2 — Proposition Generation

Task 2 generates the missing proposition for tweets classified as enthymematic by Run 1. For the constrained run, we use Flan-T5-large (fp16) [7] with beam search (beams = 4, no-repeat-ngram-size = 3). The prompt template conditions generation on both the tweet text and the predicted proposition type. For the open run, we use Mistral-7B-Instruct-v0.2 (4-bit NF4 quantised) [8] with an instruction-style prompt and repetition penalty 1.15. Both runs apply post-processing: stripping @handles and #hashtags, enforcing a minimum five-word output length, and rejecting generations with more than 40% bigram overlap with the source tweet. Task 2 was evaluated using ROSCOE-SS (facebook/roscoe-512-roberta-base) [9], a reasoning-oriented semantic similarity metric, rather than BERTScore as originally announced. ROSCOE-SS computes cosine similarity between generated propositions and annotator reconstructions, scaled to [0,1] via (1 + cos)/2, and averaged across up to three references per tweet.

4. RESULTS AND ANALYSIS

4.1 Task 1 Results

Table 1 summarises Task 1 performance on the official test set. Run 1 achieves the highest binary macro F1 (0.6598), the primary competition metric, correctly identifying whether any implicit component is present in a tweet regardless of type. Run 2, incorporating disagreement features, achieves marginally better premise F1 (0.5470 vs 0.5510 for Run 1 — comparable) but falls short on binary F1 (0.6234). Both runs score 0.0000 on implicit_conclusion, a result attributable to the very small number of conclusion examples in the official test set (n=6): with so few instances, even one misclassification eliminates F1 entirely.

Table 1: Task 1 Results on Official Test Set (* = primary competition metric)

System	3-class F1	Binary F1*	Premise F1	Concl. F1
Baseline (TF-IDF + LR)	0.4213	0.6297	0.5120	0.0000
Run 1: DeBERTa (text-only)	0.4281	0.6598	0.5510	0.0000
Run 2: DeBERTa + disagreement	0.4046	0.6234	0.5470	0.0000
Run 3: Cascade	0.2528	0.6297	0.0000	0.0312

On our internal validation split, disagreement features had improved implicit_conclusion F1 from 0.08 (Run 1) to 0.25 (Run 2), suggesting the signal is real but requires more examples to generalise reliably. This finding is consistent with Plank [4] and Stahl et al. [2]: annotation uncertainty encodes information about argumentation structure, but small dataset size constrains what can be demonstrated empirically.

Run 1 achieves strong none F1 (0.7333) and premise F1 (0.5510), reflecting a well-calibrated model on the majority classes. Run 2 trades some binary F1 for better class balance, consistent with the fusion layer learning to redistribute confidence away from the majority class.

4.2 Task 2 Results

Table 2 reports ROSCOE-SS scores on the official test set. The matched-only metric averages only over tweets where both a system generation and a gold reference are available; with-zeros assigns a score of 0 to cases where a generation was produced but no gold reference exists, or vice versa.

Table 2: Task 2 ROSCOE-SS Results on Official Test Set

System	Matched only	With zeros	Missing	Over-gen.

Run 1 (Flan-T5-large)	0.8710	0.3656	16	31
Run 2 (Constrained)	0.8597	0.3009	22	30
Run 3 (Mistral-7B open)	0.8826	0.3089	22	30

Matched-only scores are strong across all runs (0.86–0.88), confirming that when the system generates a proposition, it generates a semantically appropriate one. The large gap between matched-only and with-zeros scores is driven by over-generation: Run 1 produced 31 propositions for tweets that had no gold reference, and missed 16 tweets that did. This coverage mismatch reflects a pipeline alignment issue between Task 1 predictions and Task 2 generation targets — the system generates for all tweets classified as enthymematic, but not all such tweets have annotator-provided reconstructions. Run 3 (Mistral-7B) achieves the best matched-only score (0.8826), consistent with our qualitative observation that Mistral more reliably infers the underlying unstated belief rather than paraphrasing the source tweet.

4.3 Failure Analysis

Qualitative inspection reveals two dominant error types. First, tweets with high disagreement entropy but a majority label of none are occasionally misclassified as enthymematic — the model over-generalises the disagreement signal. Second, implicit conclusions requiring knowledge of specific political events remain difficult for all systems, as neither text nor disagreement features provide sufficient grounding.

Run 3 (cascade architecture) failed entirely on the premise class in the official evaluation (F1 = 0.0000), predicting only none and conclusion labels across 148 tweets. The root cause is a compounding failure across the two cascade stages. Model B was trained on only the enthymeme-labelled rows (~394 training examples), of which only ~57 were implicit_conclusion — an extremely small and imbalanced subset. When Model A's threshold admitted a broader set of tweets as enthymematic at inference, Model B encountered out-of-distribution examples and defaulted to predicting conclusion, the class it had been most incentivised not to miss. This suppressed premise predictions entirely. The cascade architecture also prevents recovery: any misclassification in Model A propagates irreversibly to Model B, and error compounding across two binary stages is more damaging than a single 3-class model making the same errors independently.

5. DISCUSSION AND OUTLOOK

Our results confirm that inter-annotator disagreement carries diagnostic value for enthymeme detection, particularly for the implicit_conclusion class on internal evaluation. The failure to generalise to the official test set is primarily a data quantity problem: with only 6 conclusion examples in the test set, statistical evaluation is unreliable, and we do not interpret this as evidence against the utility of disagreement features.

A practical implication follows: future annotation campaigns for argumentation tasks should preserve individual annotator labels rather than resolving them to majority votes, as the distribution of labels carries genuine linguistic signal. This is consistent with the broader call in Plank [4] for perspectivist approaches to NLP annotation.

Future work should explore true soft-label training objectives (e.g., KL-divergence loss against the full annotator vote distribution) as an alternative to input-feature-based disagreement encoding, and retrieval-augmented generation for Task 2 to provide the background knowledge that presupposed conclusions require.

DECLARATION ON GENERATIVE AI

During the preparation of this work, the authors used Claude (Anthropic) to assist with document formatting and draft editing. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

REFERENCES

[1] M. Pastor and N. Oostdijk. A Resource for Enthymeme Detection in Controversial Political Discourse. arXiv preprint arXiv:XXXX.XXXXX, 2026.

[2] M. Stahl, N. Düsterhus, M.-H. Chen, and H. Wachsmuth. Mind the Gap: Automated Corpus Creation for Enthymeme Detection and Reconstruction in Learner Arguments. Findings of ACL: EMNLP 2023, pp. 4703–4717, Singapore.

[3] E. Sviridova, E. Cabrio, and S. Villata. Mining implicit arguments for reasoning: A survey. Journal of Language Technology, 2026. https://doi.org/10.1177/19462174251344764

[4] B. Plank. The 'problem' of human label variation: On ground truth in data, modeling and evaluation. Proc. EMNLP 2022, pp. 10671–10682, Abu Dhabi.

[5] A. Uma et al. Learning from disagreement: A survey. Journal of Artificial Intelligence Research, 72:1385–1470, 2021.

[6] P. He, J. Gao, and W. Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. ICLR 2023. arXiv:2111.09543.

[7] H. W. Chung et al. Scaling Instruction-Finetuned Language Models. arXiv:2210.11416, 2022.

[8] A. Q. Jiang et al. Mistral 7B. arXiv:2310.06825, 2023.
[9] O. Golovneva et al. ROSCOE: A Suite of Metrics for Scoring Step-by-Step Reasoning. ICLR 2023. arXiv:2212.07919.