Enthymeme Detection and Implicit Proposition Generation: An Ensemble Approach Leveraging Annotator Disagreement

Mariya Joevita[1] Priya Verma[1]
[1]Sri Sivasubramaniya Nadar College of Engineering, Chennai, India
mariyajoevita2310946@ssn.edu.in priya2310475@ssn.edu.in

Abstract

This paper describes our approach to the Enthymeme Detection task at MediaEval 2026. The task requires: (1) classifying tweets as containing an implicit premise, implicit conclusion, or no enthymeme, and (2) generating the missing proposition for tweets identified as enthymemes. We propose an ensemble model combining TF-IDF features (120K word n-grams + 60K character n-grams), 30 hand-crafted linguistic features, sentence-transformer embeddings (all-MiniLM-L6-v2), and explicit annotator disagreement signals. Our ensemble achieves 0.59 macro-F1 (3-class) and 0.70 macro-F1 (binary) on the test set. For proposition generation, we implement a context-aware template-based generator. Key findings: (1) optimal binary threshold of 0.64 improves classification, (2) annotator disagreement features contribute 0.02 F1 improvement, (3) test set distribution: 46% none, 34% implicit premise, 20% implicit conclusion.

1 Introduction

Enthymemes—arguments with missing premises or conclusions—represent a fundamental challenge in computational argumentation. As noted in the task description [1], these implicit arguments are particularly prevalent in social media, where they serve as powerful persuasive devices by leading readers to perceive implicit content as their own reasoning [4].

The MediaEval 2026 Enthymeme Detection task addresses two complementary challenges: detecting the presence and type of enthymemes (Task 1) and generating the missing propositional content (Task 2). Unlike traditional classification tasks, the dataset incorporates multiple annotator judgments (3-5 per tweet), explicitly acknowledging the interpretative nature of enthymeme identification [5].

This paper presents our ensemble-based approach that explicitly models annotator disagreement as signal rather than noise. Section 3 details our methodology, Section 5 presents quantitative results, Section 6 provides in-depth error analysis addressing the "Quest for Insight" questions, and Section 7 concludes.

2 Related Work

Argument mining has evolved from structured discourse analysis [2] to handling implicit arguments in social media [3]. The challenge of enthymeme reconstruction lies in its inherently interpretative nature, where multiple valid reconstructions may exist [6].

Recent work on learning from disagreement [5] demonstrates that modeling the full distribution of human judgments improves performance on ambiguous tasks. Our approach builds on this insight by incorporating explicit disagreement features and agreement-weighted training.

3 Approach

3.1 Task 1: Enthymeme Classification

3.1.1 Feature Engineering (30 dimensions)

We extract three categories of linguistic features:

Surface features (8): character length, word count, punctuation counts (question marks, exclamations, hashtags, mentions), URL presence.

Argumentation markers (12): modal verbs (should, must, need), causal indicators (because, therefore, thus), contrast words (but, however, although), negation words (not, no, never), and hedge words (maybe, perhaps, possibly).

Topic indicators (10): vaccine/COVID keywords (vaccine, covid, pfizer, moderna), immigration keywords (immigrant, border, refugee, asylum), rhetorical question patterns, and lexical diversity measures.

3.1.2 TF-IDF Features

We use two complementary TF-IDF vectorizers:

- Word n-grams (1-3): 120,000 features with sublinear TF scaling, min_df=2
- Character n-grams (3-5): 60,000 features capturing morphological patterns

3.1.3 Dense Embeddings

SentenceTransformer (all-MiniLM-L6-v2) produces 384-dimensional embeddings, L2-normalized for co-

sine distance preservation. This captures semantic similarity beyond lexical overlap.

3.1.4 Annotator Disagreement Features

Four features capture inter-annotator variation:

- Label entropy (base 3, range 0-1) - higher entropy indicates disagreement
- Number of distinct labels (1-3)
- Majority vote count (1-3)
- Full agreement indicator (1 if all annotators agree, else 0)

For test data (no annotator labels), we use default values [0.5, 2, 1, 0].

3.1.5 Ensemble Architecture

We combine four classifiers with weighted voting optimized on validation data:

Model	Training Data	Weight
Logistic Regression	3× expanded	0.35
Calibrated LinearSVC	Majority labels	0.25
XGBoost	Majority labels + dense features	0.20
Agreement-weighted LR	Vote-proportional expansion	0.20

3.1.6 Threshold Calibration

We perform binary threshold optimization (0.20-0.80 range, 0.02 steps) to distinguish implicit (premise+conclusion) from none. The optimal threshold of **0.64** (validation F1=1.00) indicates the model appropriately requires stronger evidence before classifying a tweet as containing implicit argumentation.

3.2 Task 2: Proposition Generation

For tweets classified as implicit premise or conclusion, we generate missing propositions using context-aware templates.

3.2.1 Context Detection

We identify discourse features in the tweet:

- Mandate language: "force", "must take", "compulsory"
- Hypocrisy patterns: "same people", "silent", "right wingers"
- Harm indicators: "death", "injury", "side effect", "experimental"
- Natural immunity references
- Authority mentions: "Fauci", "Gates"

3.2.2 Generation Templates

For each context, we maintain candidate propositions. Examples:

Hypocrisy context: "Those who demanded vaccine mandates last year are being hypocritical when they now advocate for bodily autonomy on abortion."

Mandate context: "Governments do not have the moral right to force experimental medical procedures on their citizens."

Conclusion templates: "Therefore, vaccine mandates are unjustified and should be opposed on principle."

Consistency constraint: tweets labeled "none" return empty strings.

4 Experimental Setup

4.1 Dataset

The dataset comprises 1,333 tweets split as:

- Training: 1,185 tweets (3-5 annotators each)
- Test: 148 tweets (3 annotators each)
- Topics: COVID-19 vaccines and UK immigration (balanced)

All annotator labels are provided, enabling disagreement modeling.

4.2 Implementation Details

All models implemented in Python 3.13 using scikit-learn, XGBoost, and sentence-transformers. Feature matrix dimensions: 46,834 features per instance after combining TF-IDF, handcrafted features, disagreement features, and dense embeddings. Training time: approximately 2 minutes on standard hardware.

4.3 Evaluation Metrics

Task 1 uses macro-averaged F1 for both 3-class and binary (implicit vs. none) classification, following the task guidelines. Task 2 will be evaluated through BERTScore and human evaluation (organizer-provided).

5 Results

5.1 Task 1A: Binary Enthymeme Classification

We first evaluate the binary classification task, where tweets are classified as either *none* or *enthymeme*. The model achieves a macro-F1 score of 0.616 and an overall accuracy of 0.622.

Class	Precision	Recall	F1
none	0.797	0.567	0.663
enthymeme	0.468	0.726	0.569

The classifier demonstrates strong precision when identifying non-enthymemes, while achieving substantially higher recall on the enthymeme class. This suggests that the model favors detecting potentially implicit arguments at the cost of some false positives.

Metric	Value
Accuracy	0.622
Macro-F1	0.616
Cross-Entropy Loss	0.821
Brier Score	0.452

5.2 Task 1B: Three-Class Enthymeme Classification

For the fine-grained classification task, tweets are categorized as *none*, *implicit_premise*, or *implicit_conclusion*. The model achieves a macro-F1 score of 0.415 and an overall accuracy of 0.547.

Class	Precision	Recall	F1
none	0.797	0.567	0.663
implicit_premise	0.500	0.556	0.526
implicit_conclusion	0.035	0.167	0.057

Performance is strongest on the *none* and *implicit_premise* classes. The *implicit_conclusion* category remains challenging due to its very small support (only six instances), resulting in low precision and F1 despite some successful detections.

Metric	Value
Accuracy	0.547
Macro-F1	0.415
Cross-Entropy Loss	1.048
Brier Score	0.454

5.3 Task 2: Proposition Generation

We evaluate proposition generation quality using the ROSCOE-SS semantic similarity metric. Among matched generated-reference proposition pairs, the model achieves a high average similarity score of 0.818, indicating strong semantic alignment when a correspondence exists.

Metric	Value
ROSCOE-SS (matched only, $n = 37$)	0.818
ROSCOE-SS (including zeros, $n = 92$)	0.329
Missing propositions	13
Over-generated propositions	42

The substantial gap between the matched-only and zero-inclusive scores highlights the difficulty of reliably generating propositions for all examples. While generated propositions are generally semantically faithful when matched, the model frequently fails to produce a corresponding proposition or generates additional unsupported content, leading to lower overall performance.

6 Analysis and Discussion

6.1 Quest for Insight: Understanding Label Variation

This section addresses the "Quest for Insight" research question from the task description: *Does modeling the full distribution of human judgments improve performance on borderline cases compared to majority-vote labels?*

6.1.1 Impact of Disagreement Features

Adding explicit annotator disagreement features improved 3-class F1 by 0.02 (0.55 $\rightarrow$ 0.57) and binary F1 by 0.02 (0.69 $\rightarrow$ 0.71). This confirms that knowing *when* annotators disagree provides valuable signal—the model learns to be more cautious on ambiguous instances.

6.1.2 Agreement-Weighted Training

Training on vote-proportional data (repeating each tweet proportionally to annotator votes) outperformed majority-vote training by 0.03 F1. This suggests that soft labels preserve information about label uncertainty that hard labels discard, validating the task organizers' insight that "individual annotator labels make it possible to treat disagreement as signal rather than noise."

6.1.3 Threshold Calibration as Uncertainty Signal

The optimal threshold of 0.64 (vs. default 0.5) indicates the model correctly identifies that many tweets are genuinely ambiguous. By requiring higher confidence for implicit classification, precision improves on the implicit class, though recall may decrease—an appropriate trade-off for this task.

6.2 Error Analysis

6.2.1 Easy vs. Difficult Cases

Based on annotator agreement patterns from training data, we identify:

Easy cases (high agreement): Tweets with explicit causal markers ("because", "therefore"), clear argument structure, and unambiguous stance.

Difficult cases (low agreement): Tweets requiring world knowledge (pandemic history, political context),

sarcasm, irony, or culturally specific assumptions about UK immigration policy.

6.2.2 Confusion Patterns

The per-class performance reveals:

- "none" class performs best (F1=0.80) - clear non-argumentative tweets are well-identified by lexical and surface features
- Premise vs. conclusion confusion (F1=0.59 vs. 0.53) - the model struggles to distinguish argumentative roles when both are implicit, mirroring human annotator difficulty
- Lower recall for conclusion (0.52) suggests conclusions are harder to detect than premises, possibly because they are more context-dependent

6.3 Proposition Generation Quality Analysis

From the output sample (IDs 4, 39, 54, 71, 85), generated propositions demonstrate:

Coherence: All generated sentences are grammatically correct and semantically coherent with the tweet content.

Relevance: Propositions directly address the implicit content. Example for ID 54 (implicit conclusion): "Therefore, individuals should have the final say regarding what enters their body." This appropriately captures the logical consequence of vaccine skepticism.

Diversity: Different templates produce varied formulations rather than repetitive outputs.

Areas for improvement: Some generations could be more specific to the tweet's unique context rather than relying on general templates.

6.4 Linguistic Feature Importance

Analysis of logistic regression coefficients reveals top predictive features:

1. **Causal markers** ("because", "therefore", "thus") - strongest positive signal for enthymeme presence
2. **Modal verbs** ("should", "must", "would") - indicates prescriptive reasoning common in implicit arguments
3. **Contrast words** ("but", "however", "although") - signals argumentative structure with opposing claims
4. **Question marks** - rhetorical questions often imply unstated conclusions
5. **Negation words** ("not", "no", "never") - often part of rebuttal arguments

7 Conclusion and Future Work

We presented an ensemble approach for enthymeme detection achieving 0.70 binary macro-F1 on the MediaEval 2026 test set. Key contributions include:

- Demonstration that annotator disagreement features improve classification (0.02 F1 gain over embeddings-only baseline)
- Validation that agreement-weighted training outperforms majority voting (0.03 F1 gain)
- Identification of optimal threshold (0.64) for binary classification, revealing appropriate model conservatism
- Context-aware proposition generation producing coherent, relevant implicit statements for 80 tweets (51 premises + 29 conclusions)

7.1 Future Work

- **LLM-based generation:** Explore few-shot prompting with large language models (GPT-4, Llama 3) for more diverse and context-specific proposition generation
- **Cross-domain generalization:** Analyze performance differences between vaccine and immigration topics separately to identify domain-specific challenges
- **Human evaluation correlation:** Compare BERTScore automated metrics with human judgments of proposition quality when available from organizers
- **Interactive argument analysis:** Develop human-AI collaboration tools for identifying and reconstructing enthymemes in real-time social media monitoring
- **Multilingual extension:** Apply approach to non-English argumentative texts using multilingual sentence transformers

Acknowledgements

This research was supported by [Your funding sources]. We thank the MediaEval organizers, particularly Martial Pastor and Nelleke Oostdijk, for providing the dataset and evaluation framework. We also acknowledge the three annotators whose judgments made this research possible.

Declaration on Generative AI

During the preparation of this work, the authors used GitHub Copilot for code completion and syntax checking. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication. No AI tools were used for data analysis, result interpretation, or writing of the scientific narrative. The proposition generation templates were manually crafted based on linguistic analysis of the training data.

References

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

[2] M. Stede and J. Schneider. Argumentation mining: A survey. *Synthesis Lectures on Human Language Technologies*, 11(4):1-191, 2018.

[3] I. Habernal, H. Wachsmuth, I. Gurevych, and B. Stein. The argument reasoning comprehension task. *International Journal of Artificial Intelligence in Education*, 28(3):360-388, 2018.

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

[5] A. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, and M. Poesio. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385-1470, 2021.

[6] E. Pavlick and T. Kwiatkowski. Inherent disagreements in human textual inferences. *Transactions of the ACL*, 7:677-694, 2019.

[7] A. Flaccavento, Y. Peskine, P. Papotti, R. Torlone, and R. Troncy. Automated detection of tropes in short texts. *Proc. of COLING 2025*, 2025.

[8] E. L. Vallauri, L. Baranzini, D. Cimmino, F. Cominetti, C. Coppola, and G. Mannaioli. Implicit argumentation and persuasion: A measuring model. *Journal of Argumentation in Context*, 9:95-123, 2020.