Missing Pieces and Misinformation: Shared Task on Detecting and Generating Implicit Components in Enthymematic Arguments

Martial Pastor[1,*], Nelleke Oostdijk[1]

[1]Centre for Language Studies, Radboud University, Nijmegen, The Netherlands

Abstract

This paper describes the Missing Pieces and Misinformation task and its accompanying dataset of tweets annotated for implicit argumentative content. The task focused on enthymemes, arguments in which a premise or conclusion is left unstated, a structure both pervasive in natural language and highly effective as a persuasive strategy in online discourse. The tweets address the topics of Covid and Immigration and potentially convey controversial stances through unstated propositions. The task comprised two subtasks: (1) detecting the presence of enthymemes in tweets, and (2) generating a single natural language sentence expressing the implicit component when an enthymeme is present. Each data point was annotated by five independent annotators in order to capture variation in semantic interpretation. We describe the task, dataset, and evaluation framework, and situate the task within research on argumentation mining and misinformation detection.

1. Introduction

Implicit argumentation is a natural characteristic of everyday language. Whether used candidly or deliberately as a rhetorical device, it may obscure fallacies and facilitate the spread of misinformation [1]. Studies in the pragmatics of manipulation have long recognized its persuasive success [2]: readers who must reconstruct an unstated message perceive it as their own and believe it more readily. Online communities form and consolidate around shared worldviews [3], and implicit encoding strategies are particularly effective at embedding questionable content within them [1, 4].

The Missing Pieces and Misinformation shared task investigated implicit argumentation in tweets about Covid and immigration. The task focuses on a particular argumentative structure, the enthymeme, an argument with either a missing premise or a missing conclusion. To support the task, we released a dataset of 1,482 tweets, each annotated for the presence or absence of an enthymeme; where one is present, annotators also provided a natural language reconstruction of the implicit component, capturing the stance, claim, belief, or attitude it conveys. Participants were challenged to develop NLP methods that first determine whether a tweet contains an enthymeme, and then generate the text of the missing component.

A secondary goal of the task was to engage with the interpretive nature of enthymeme annotation. Identifying enthymemes requires inference and semantic interpretation, where human disagreement is not a flaw but an expected outcome [5, 6]. The task inherently allows for multiple plausible readings [7], and collapsing variation into a single correct label discards valuable information about human judgment [8]. Participants were therefore encouraged to treat annotator disagreement as signal rather than noise, and were invited to explore several

MediaEval'26: Multimedia Evaluation Workshop, June 15–16, 2026, Amsterdam, Netherlands and Online
*Corresponding author.
martial.pastor@ru.nl (M. Pastor); nelleke.oostdijk@ru.nl (N. Oostdijk)

open questions: What patterns emerge in label variation across easy, medium, and hard cases? Does modeling the full distribution of human judgments improve performance on borderline cases over majority-vote labels? Can disagreement patterns predict which instances will be hardest for models? And what linguistic or discourse features can improve classification?

2. Background and Related Work

Though the study of enthymemes dates back to Aristotle and the rhetorical traditions of the ancient world [9, 10], it has most recently become relevant as a continuation of argument mining [11], traditionally concerned with the automatic identification and classification of explicit argument structures in text [12, 13]. Findings in this discipline point to a recurring challenge: much argumentative reasoning relies on implicit knowledge and hidden meanings that approaches targeting explicit structure do not capture, and uncovering such implicit meaning requires dedicated pragmatic inference [14].

Research on implicit argumentation has progressed from foundational studies establishing the feasibility of annotating implicit content to the development of computational models [15, 16]. Key benchmarks include ARCT, designed for identifying the warrants that justify reasoning [17], and IRAC, which generates semi-structured causal chains to map background knowledge [18]. While enthymeme detection research has diverged into NLP-based textual methods and symbolic logic-based systems [19, 20], other approaches integrate generative language models with formal logical verification [21]. For implicit component generation, generative AI has led to systems that produce multi-step reasoning chains bridging premises and claims [22, 23], and others that reconstruct the causal chains underlying an argument's unstated reasoning [18, 24].

Our approach focuses on enthymemes where the implicit component is itself a controversial stance or carries (mis)informative content, unlike resources such as ARCT and IRAC, where the missing element is most often neutral commonsense or background knowledge [17, 18]. However, getting into the mind of a potentially controversial author raises the well-known risk of the *straw man fallacy*: the analyst may reconstruct a missing component the arguer did not intend [9]. To guard against this, the annotation framework requires that the reconstructed premise or conclusion restore classical entailment over the full argument, such that the argument's own logical structure constrains reconstruction and limits the space within which annotator interpretation can operate [25]. Prior work in persuasion research and the pragmatics of implicit communication has documented how contentious political content is linguistically encoded in discourse, yet to the best of our knowledge no existing resource rigorously tracks such content through the logical structure of enthymematic arguments. The present dataset was designed to fill this gap.

3. Dataset

The data come from a preexisting dataset of tweets introduced by Flaccavento et al. in their study of trope usage in social media [26]. The authors collected English-language tweets via keyword-based queries targeting vaccines (June 26–27, 2022) and immigration (2019–2022), and cleaned the dataset by stripping links and removing personally identifiable information, without applying a profanity filter.

From this dataset, we selected a subset of 1,483 tweets so as to obtain a balanced distribution of trope vs. no-trope labels, constituting the Missing Pieces dataset. The annotation proceeded in three phases: trial explorative annotations from which elementary guidelines emerged; the recruitment and training of annotators who helped further elaborate the guidelines; and a final

phase in which five annotators (three primary annotators and two validators) were assigned different subsets so as to maximise the number of data points. The released dataset thus contains three original annotations and two validation annotations per data point, with individual labels released prior to any majority vote, making it possible to treat disagreement as signal rather than noise. The full annotation framework is described in a companion paper [25].

For a tweet to be annotated as an enthymeme containing either an `implicit_premise` or `implicit_conclusion` rather than `none`, five criteria must all hold: (1) the missing component must express a stance on COVID-19 vaccines or immigration, not a neutral general claim; (2) at least two premises must share a term; (3) the reconstructed argument must fit a recognisable argumentation scheme; (4) at most one non-commonsense implicit proposition can be reconstructed; if two substantive implicit steps are needed, the tweet is labelled `none`; and (5) at least two explicit argument segments must appear in the tweet; a bare single-proposition tweet cannot be annotated unless loaded language or proposition unpacking yields a second segment (see Appendix A for an example).

4. Task 1: Enthymeme Detection and Classification

Task 1A is a binary classification task: given a tweet, the system must determine whether it contains an enthymeme (`enthymeme`) or not (`none`). Task 1B extends this to three classes: `implicit_premise`, `implicit_conclusion`, or `none`. An implicit premise is a supporting assumption the argument relies upon but leaves unstated; an implicit conclusion is a claim that follows from the stated premises but is never explicitly made. The two subtasks are logically related, since only tweets labeled as enthymemes in Task 1A can receive an implicit-component label in Task 1B, but participants were free to approach them in either order.

For both subtasks we proposed the following runs. **Constrained Run 1:** tweet text only and majority vote label, no external resources. **Constrained Run 2:** tweet text plus raw labels from three annotators, to investigate whether modeling disagreement improves performance on borderline cases. **Open Run:** any external data or models permitted, documented in the working notes.

Evaluation. Participants were asked to submit two outputs per instance: a hard label prediction, used to compute precision, recall, and F1 against the majority-vote gold label, and a probability vector over the label set, used to compute cross-entropy loss against the soft gold label derived from the full annotation distribution. This accounts for the interpretive nature of the task, where disagreement reflects genuine ambiguity rather than noise [27]. The soft gold label is the normalized frequency vector of the five annotations: for Task 1A, three `enthymeme` and two `none` labels yield $\mathbf{q} = [0.6, 0.4]$. Cross-entropy between the predicted distribution $\mathbf{p}$ and $\mathbf{q}$ is $\mathcal{L}_{\text{CE}} = -\sum_i q_i \log p_i$ over the label set, averaged over all test instances; lower is better, with a perfect match scoring 0.

5. Task 2: Missing Component Generation

For tweets classified as containing an enthymeme, the system must generate the missing proposition. The input is the tweet text; the output is a single natural language sentence making the unstated premise or conclusion fully explicit.

Example. Consider the following tweet:

"Deterring the plans of illegal people smugglers is essential to controlled immigration. We should support all plans to stop them."

The full argument reconstructs as:

- Premise 1 **(implicit, to generate)**: Controlled immigration is desirable.
- Premise 2 (explicit): Deterring the plans of illegal people smugglers is essential to controlled immigration.
- Conclusion (explicit): We should support all plans to stop them.

The expected system output is: *"Controlled immigration is desirable."*

Evaluation. System outputs were evaluated with metrics developed for natural language explanations and step-by-step rationales, along two complementary dimensions: **consistency with the source tweet** and **similarity with the gold reconstructions**.

For consistency, we use the Inter-step Correctness metric of RecEval [28], which assesses "global consistency" by verifying the absence of contradictions between the generated component and the source context. The text is broken into granular claims called Reasoning Content Units (semantic subject-verb-object triplets), and a logic model identifies the unit with the highest contradiction risk (P_{contr}) relative to the original tweet. The score $1 - \max(P_{\text{contr}})$ ensures that even a single logical error significantly lowers the result, with 1 indicating perfect compatibility and 0 a direct contradiction.

For similarity against gold reconstructions, we use the Semantic Coverage-Chain ($r \leftrightarrow h$) metric from the Roscoe-SS category [29], which relies on a SimCSE model fine-tuned on reasoning datasets to capture semantic equivalence of logical propositions rather than word overlap. The score $\frac{1+\cos(r,h)}{2}$, the normalized cosine similarity between the embeddings of the generated component and the human reference, lies in $[0, 1]$, with 1 indicating a perfect match with the gold meaning.

6. Submissions and Working Notes

Participants communicated labels and probability distributions for Tasks 1A and 1B, and generated implicit text for Task 2, accompanied by working notes presenting their modeling and design choices (architecture, computational complexity, hyperparameters, and resource utilisation for neural systems). They were also invited to address the open questions posed in Section 1, reflecting on the coherence of the dataset for computational exploitation and on the potential societal impact of such systems.

7. Conclusion

The Missing Pieces and Misinformation shared task addressed the detection of enthymemes and the generation of their implicit components in tweets about Covid and immigration, with an evaluation methodology that treats annotator disagreement as informative signal. We hope it will encourage researchers to further explore how persuasive political content operates through implicit reasoning.

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 the UK Research and Innovation (UKRI) Horizon Funding Guarantee, and the AI-CODE project, funded under the European Union's Horizon Europe research and innovation programme grant agreement no. 101135437.

Disclaimer

The dataset used in this shared task contains social media posts that may include offensive, harmful, or controversial language and viewpoints. These do not reflect the views of the task organizers, and are included solely for the purpose of scientific research.

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] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, W. Quattrociocchi, The spreading of misinformation online, Proceedings of the National Academy of Sciences 113 (2016) 554–559. doi:10.1073/pnas.1517441113.

[4] F. Macagno, Argumentation profiles and the manipulation of common ground: The arguments of populist leaders on Twitter, Journal of Pragmatics 191 (2022) 67–79. doi:10.1016/j.pragma.2022.01.011.

[5] 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.

[6] L. Aroyo, C. Welty, Truth is a lie: Crowd truth and the seven myths of human annotation, AI Magazine 36 (2015) 15–24. doi:10.1609/aimag.v36i1.2564.

[7] E. Pavlick, T. Kwiatkowski, Inherent disagreements in human textual inferences, Transactions of the Association for Computational Linguistics 7 (2019) 677–694. doi:10.1162/tacl_a_00293.

[8] 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.

[9] D. Walton, Media argumentation: Dialectic, persuasion and rhetoric, Cambridge University Press (2007). doi:10.1017/CBO9780511619311, new York. ISBN: 978-0-521-70030-6.

[10] F. Macagno, D. Walton, Emotive language in argumentation, Cambridge University Press (2014). New York. ISBN: 978-1-107-67665-7.

[11] J. Lawrence, C. Reed, Argument mining: A survey, Computational Linguistics 45 (2019) 765–818. doi:10.1162/coli_a_00364.

[12] M. Lippi, P. Torroni, Argumentation mining: State of the art and emerging trends, ACM Transactions on Internet Technology 16 (2016) Art. 10. doi:10.1145/2850417.

[13] I. Habernal, I. Gurevych, Argumentation mining in user-generated web discourse, Computational Linguistics 43 (2017) 125–179. doi:10.1162/COLI_a_00276.

[14] S. Oswald, Commitment attribution and the reconstruction of arguments, The Psychology of Argument: Cognitive Approaches to Argumentation and Persuasion (2016) 17–32. London: College Publications. Studies in Logic and Argumentation.

[15] F. Boltužić, J. Šnajder, Fill the gap! Analyzing implicit premises between claims from online

debates, Proceedings of the Third Workshop on Argument Mining (ArgMining2016) (2016) 124–133. doi:10.18653/v1/W16-2815, berlin, Germany. Association for Computational Linguistics.
[16] M. Becker, M. Staniek, V. Nastase, A. Frank, Enriching argumentative texts with implicit knowledge, Natural Language Processing and Information Systems: 22nd International Conference on Applications of Natural Language to Information Systems, NLDB 2017 10260 (2017) 84–96. doi:10.1007/978-3-319-59569-6_9, springer, Cham. Lecture Notes in Computer Science.
[17] I. Habernal, H. Wachsmuth, I. Gurevych, B. Stein, The argument reasoning comprehension task: Identification and reconstruction of implicit warrants, Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) 1 (Long Papers) (2018) 1930–1940. doi:10.18653/v1/N18-1175, new Orleans, Louisiana. Association for Computational Linguistics.
[18] K. Singh, N. Inoue, F. Sultana Mim, S. Naitoh, K. Inui, IRAC: A domain-specific annotated corpus of implicit reasoning in arguments, Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022) (2022) 4674–4683. Marseille, France. European Language Resources Association.
[19] E. Sviridova, E. Cabrio, S. Villata, Mining implicit arguments for reasoning: A survey, Argument & Computation 17 (2025). doi:10.1177/19462174251344764.
[20] A. Hunter, Understanding enthymemes in deductive argumentation using semantic distance measures, Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22) 36 (2022) 5729–5736. doi:10.1609/aaai.v36i5.20515.
[21] X. Feng, A. Hunter, Making implicit premises explicit in logical understanding of enthymemes, arXiv preprint arXiv:2603.06114 (2025).
[22] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, Advances in Neural Information Processing Systems 35 (NeurIPS 2022) (2022) 24824–24837.
[23] DeepSeek-AI, DeepSeek-V3.2: Pushing the frontier of open large language models, arXiv preprint arXiv:2512.02556 (2025).
[24] S. Saha, P. Yadav, L. Bauer, M. Bansal, ExplaGraphs: An explanation graph generation task for structured commonsense reasoning, Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2021) 7716–7740. doi:10.18653/v1/2021.emnlp-main.609, online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
[25] M. Pastor, N. Oostdijk, A resource for enthymeme detection in controversial political discourse, 2026. URL: https://arxiv.org/abs/2606.12186. doi:10.48550/arXiv.2606.12186. arXiv:2606.12186, submitted to Language Resources and Evaluation.
[26] 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.
[27] J. C. Peterson, R. M. Battleday, T. L. Griffiths, O. Zamir, Human uncertainty makes classification more robust (2019) 9617–9626.
[28] A. Prasad, S. Saha, X. Zhou, M. Bansal, ReCEval: Evaluating reasoning chains via correctness and informativeness (2023) 10066–10086. doi:10.18653/v1/2023.emnlp-main.622.
[29] 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.

A. Proposition Unpacking Example

Consider the following tweet:

> *"I'm **pleased** to announce over 20,000 refugees from Syria have been resettled in the UK since 2015."*

From this text two explicit propositions should be unpacked:

- Proposition 1 (explicit): Over 20,000 Syrian refugees have been resettled in the UK since 2015.
- Proposition 2 **(explicit, proposition unpacking):** The resettlement of refugees is a desirable outcome.

The loaded verb *pleased* transforms a bare factual announcement into a two-segment argument: the speaker's emotional stance unpacks into an evaluative claim, providing the second anchor.