Team Macaroni at SemEval-2026 Task 10: PsyCoMark: Psycholinguistic Conspiracy Marker Extraction and Detection

Rofaïda Rabehi, Nicolai Plenk, Sung-Jin Miriam Han
University of Tübingen
macaroni@mujistan.email

Abstract

This paper describes our submission to SemEval-2026 Task 10: PsyCoMark, which addresses span-level conspiracy marker extraction and document-level conspiracy detection. For conspiracy marker extraction (subtask 1), we fine-tune several pretrained transformer encoders and analyse their behaviour under different training configurations. For conspiracy detection (subtask 2), we develop a hybrid system that combines ModernBERT-large with surface-level linguistic features.

Our results show that straightforward fine-tuning of strong pretrained models is more effective than more complex pipelines and that additional handcrafted features do not yield consistent improvements. On the official test set, we rank 18th in conspiracy marker extraction (overlap-based macro F1 = 0.16) and 20th in conspiracy detection (macro F1 = 0.76).

1 Introduction

Conspiracy theories can shape public opinion, erode trust in institutions, and even motivate harmful real-world actions, making it crucial to understand how they are expressed and spread online. Detecting conspiratorial content early and understanding the language patterns that signal it can help mitigate the spread of misinformation. PsyCoMark (Samory et al., 2026) addresses this by focusing on both extracting markers—specific words or phrases that reveal conspiratorial framing—and detecting whether an entire text conveys conspiracy beliefs.

Our approach to this task relies on fine-tuning pretrained transformer models and systematically comparing different architectures and training strategies. For conspiracy marker extraction (subtask 1), we evaluate various encoder architectures to extract conspiracy markers. For conspiracy detection (subtask 2), we use the encoder-only transformer ModernBERT-large (Warner et al., 2024) and explore the integration of additional linguistic and readability features. We further employ stratified cross-validation, focal loss to address class imbalance, and an ensemble of fold-specific models with tuned decision thresholds. We also experiment with generative models and larger auxiliary systems, but these do not lead to improvements and are therefore not part of our final submission.

To support reproducibility, we released our code on GitHub.

2 Background

2.1 Dataset

The dataset (Samory et al., 2025) consists of 4361 submission statements from Reddit, which are top-level comments made by discussion starters that usually accompany media submissions and explain their relevance to the Subreddit. Each submission statement is labelled for conspiracy content as `Yes`, `No`, or `Can't tell`, and is also annotated for zero or more of the psycholinguistic markers introduced in the conspiracy marker extraction subtask.

2.2 Conspiracy marker extraction

This subtask requires identifying spans of text which contain conspiracy markers based on evolutionary psychology. Each Reddit comment may have zero or several overlapping spans. The conspiracy markers are, as can be seen in Table 1:

- actor: the entities responsible for a malicious act or agenda,
- action: the malicious act an actor is carrying out or planning,
- effect: the negative outcomes of the action or agenda,
- victim: the entities harmed by the action or agenda,

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 338–342

- evidence: reasoning or claims backing the conspiracy theory.

Text	NYC Mandates Vaccine in order to participate in "society". My Body, My Rules, My choice. Segregation is wrong. Take back NYC and USA from forced experimental injections on the people! End GMO Humans!
Actor	NYC
Action	NYC Mandates Vaccine in order to participate in "society".
Effect	forced experimental injections on the people! End GMO Humans!
Victim	the people!
Evidence	My Body, My Rules, My choice.

Table 1: Examplary visualisation of a dataset entry showing a conspiratorial comment annotated with all five markers. Only the relevant text and markers are shown; full entries include additional fields.

Results are evaluated using an overlap-based macro F1-score for each marker.

2.3 Conspiracy detection

This subtask frames conspiracy detection as a binary classification task: given a Reddit comment, the system must determine whether the text is conspiratorial or not. While the evaluation is binary, the underlying dataset annotates each comment with one of three labels—`Yes`, `No`, or `Can't tell`—introducing an ambiguity that systems must account for at training and inference time. Participants were permitted, but not required, to leverage the psycholinguistic markers identified in the first subtask as additional input features. Performance is measured using a macro-averaged F1-score.

2.4 Related work

Computational research on conspiracy theories has explored both their linguistic properties and automatic detection methods. Giachanou et al. (2023) model the psychological and stylistic profiles of conspiracy propagators at the author level using psycholinguistic and other surface features. Their results show that users who consistently post conspiratorial content display characteristic patterns in their language. This approach offers insights into the psychological aspects of conspiratorial thinking.

More recent work has shifted towards contextual language models. Zrnić (2024) utilises transformer-based models for document-level conspiracy classification as well as token-level extraction of narrative elements in Telegram messages. Fine-tuned BERT and RoBERTa perform well on the classification task, while RoBERTa proves more effective for span detection.

The PsyCoMark shared task builds on these two strands of research by combining psycholinguistic aspects with sequence labeling and text classification. It focuses on identifying conspiracy markers in the text and on predicting whether a comment expresses conspiratorial thinking. Our approach is similar to the transformer-based work of Zrnić (2024). We fine-tune pretrained encoder models for both classification and span-level prediction. In addition, and in contrast to earlier approaches, we test various models and whether handcrafted linguistic features prove useful.

3 System overview

3.1 Conspiracy marker extraction

We first used the starter pack (Samory, 2025) provided by the SemEval Team due to the somewhat long training time varying between 30 to 90 minutes as we did not have immediate access to the BwUniCluster3.0 [1]. In the startercode we experimented by updating the epochs, learning rate, batch size and tokeniser. We used the following models: `microsoft/deberta-v3-base`, `microsoft/deberta-v3-large` (He et al., 2021), `FacebookAI/roberta-base` (Liu et al., 2019), `FacebookAI/roberta-large`, `google/flan-t5-base` (Chung et al., 2024), `distilbert/distilbert-base-cased` (Sanh et al., 2019), and `distilbert/distilbert-base-uncased`.

We fine-tuned all models and evaluated their performance comparatively, examining the factors underlying varying results and discussing potential improvements. The original binary token labelling scheme was retained throughout.

[1]The joint high-performance computing system of the Baden-Württemberg universities.

3.2 Conspiracy detection

We developed a hybrid classification system combining a pretrained transformer backbone with handcrafted linguistic features. We used `ModernBERT-large` (Warner et al., 2024) as our encoder, extracting the `[CLS]` token representation.

3.2.1 Linguistic Features

Alongside the contextual embeddings, we computed eight surface-level and readability features per input using the `textstat` (Ward, 2022) library: Flesch Reading Ease (Flesch, 1948); Dale–Chall Readability Score (Dale and Chall, 1948); Type-Token Ratio (TTR); sentence count; syllable count; character count; ratio of exclamation and question marks to total characters; and ratio of upper-case characters. These features were normalised per fold using min-max scaling fitted exclusively on the training split to prevent data leakage.

3.2.2 Architecture

The linguistic features were projected through a small MLP ($8 \rightarrow 32$, with batch normalisation, ReLU activation, and dropout at $p = 0.2$), then concatenated with the `[CLS]` embedding. The fused representation was passed through a two-layer classification head ($d_{\text{hidden}} + 32 \rightarrow 128 \rightarrow 3$) with ReLU and dropout.

4 Experimental setup

4.1 Conspiracy marker extraction

All training and inference were conducted on the bwUniCluster 3.0, using 16 CPU cores and 2 GPUs, with runtimes ranging from one to two hours. Starter scripts for both training and inference were provided by the shared task organisers. We experimented with a range of encoder models of varying sizes, as well as one generative model, to investigate whether model scale correlates with performance. The learning rate was varied between 1.5×10^{-5} and 6×10^{-5}, and the number of training epochs between 2 and 15.

4.1.1 Training

Due to the higher complexity of this task, we decided to first fine-tune the models based on the starter script and to change the epochs, weight decay, learning rate and batch size. For larger models we used 8 epochs with a weight decay of 0.01 and learning rate of 3×10^{-5}.

4.2 Conspiracy detection

4.2.1 Training

We trained using 5-fold stratified cross-validation (Pedregosa et al., 2011) for 6 epochs per fold. To address class imbalance, we employed Focal Loss (Lin et al., 2017) with $\gamma = 2.0$, which down-weights well-classified examples and encourages the model to focus on harder instances. Optimisation was performed with AdamW (Loshchilov and Hutter, 2019) at a learning rate of 1.5×10^{-5} with a linear decay schedule. The best checkpoint per fold was selected based on macro F1 on the held-out validation split. All training was conducted on Google Colab using an NVIDIA A100 GPU.

4.2.2 Inference

At test time, all five fold checkpoints were loaded and their softmax probability distributions averaged to produce an ensemble prediction. To account for the inherent ambiguity of the three-class setup—particularly the difficulty of the `Can't tell` category—we applied tuned decision thresholds: a sample was classified as `Yes` if $P(\text{Yes}) \geq 0.35$; otherwise, $P(\text{Can't tell})$ was penalised by a factor of 0.9 before comparing against $P(\text{No})$.

5 Results

5.1 Conspiracy marker extraction

Our system achieved 18$^{\text{th}}$ place on the leaderboard; both during the development track and the test track with an overlap-based macro F1-score of 0.16 using the above mentioned method. We also tried to use a generative model, namely `Flan-T5-base` to see if a generative model would improve our performance, however we could not see an improvement. Shortening the training data and making our own dev set during the development track was not successful either. Table 2 shows that Actor consistently achieved the highest F1 across both phases (0.330 dev, 0.300 test), though its recall dropped from 0.346 to 0.259 while precision improved. Victim showed the greatest improvement from development to test (+0.065, from 0.144 to 0.209). Effect was the weakest marker on the test set (F1 = 0.072), followed by Action (0.120). As mentioned earlier, we used encoder models of varying sizes and they all performed the same.

5.2 Conspiracy detection

Using the `ModernBERT-large` hybrid system described in Section 3, our system achieved a macro

	DEV			TEST		
Marker	**F1**	**P**	**R**	**F1**	**P**	**R**
Action	**0.143**	**0.152**	**0.135**	0.120	0.123	0.117
Actor	**0.330**	0.315	**0.346**	0.300	**0.357**	0.259
Effect	**0.110**	**0.108**	**0.113**	0.072	0.073	0.071
Evidence	0.092	0.103	0.082	**0.102**	**0.112**	**0.093**
Victim	0.144	0.170	0.125	**0.209**	**0.247**	**0.181**
Macro F1	**0.164**	—	—	0.161	—	—
Aggregate	**0.190**	**0.197**	**0.184**	0.173	0.190	0.159

Table 2: Per-marker precision (P), recall (R), and F1-score on the development and test sets for conspiracy marker extraction. For each marker, the higher value between development and test is shown in bold.

F1-score of 0.76, placing 20th on the shared task leaderboard. During the development phase, we explored several extensions that were ultimately discarded. First, we experimented with incorporating `Gemma-3-27B-IT`(Gemma Team, 2025) as an auxiliary model; second, we investigated a more comprehensive feature set including NLTK-based (Bird and Loper, 2004) sentiment analysis and a hand-curated dictionary of conspiracy-related keywords inspired by Lenoir (2024). However, neither of these additions yielded improvements on the test set, so they were excluded from the final submission.

Table 3 summarises development- and test-phase performance across all model variants explored. As a first step, we replaced the shared task baseline (DeBERTa-base) with DeBERTaV3-Large, yielding a substantial gain in macro F1 (+0.09) and raising the `Yes`-class F1 from 0.500 to 0.638. We subsequently replaced the encoder with `ModernBERT-large`. While it appears that `DeBERTaV3-large` scores better than our final submission, this is due to the difference in datasets between the development- and test-phases. We also investigated an ensemble approach incorporating `Gemma-3-27B-IT`, which we found conceptually promising; however, it scored well below the baseline even after several iterations and was therefore discarded.

6 Conclusion

Perhaps the most interesting finding from our experiments is that more is not necessarily better: neither larger models, generative architectures, nor richer feature sets consistently outperformed straightforward fine-tuning of pretrained encoders. Handcrafted linguistic features, sentiment signals, and a 27B-parameter generative ensemble all failed to improve over a well-tuned ModernBERT-large, suggesting that strong pretrained representations already capture the stylistic and semantic cues relevant to conspiratorial language.

				F1 per class	
Phase	**Model**	**Macro F1**	**Acc.**	**No**	**Yes**
DEV	Baseline (DeBERTa-base)	0.678	0.688	0.774	0.500
	DeBERTaV3-Large	0.770	0.779	0.841	0.638
TEST	ModernBERT-large (initial)	0.757	0.758	0.784	0.724
	ModernBERT-large (final)	**0.761**	**0.764**	**0.800**	**0.714**
	Gemma-3-27B-IT (ensemble)†	0.387	0.531	0.691	0.032

Table 3: Development- and test-phase results for conspiracy detection. The submitted system is shown in bold; † denotes a discarded configuration.

Equally surprising are the disparities at the marker level. Action, despite having more training instances (3,601) than Victim (2,378), scores considerably lower on the test set (F1 = 0.120 vs. 0.209). Actor, with a comparable instance count (3,639), achieves the highest F1 of all markers (0.300). This suggests that some marker types are inherently harder to delineate actions in conspiracy narratives tend to be diffuse and woven throughout a sentence, whereas actors and victims are typically expressed as discrete noun phrases. Investigating these structural differences, and whether marker-aware architectures or span-boundary objectives can address them, is a promising direction for future work.

7 Acknowledgments

We would like to thank our lecturer Çağrı Çöltekin for being a helpful guide, and the team at SemEval for giving us this challenge. The authors acknowledge support by the state of Baden-Württemberg through bwHPC.

References

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2024. Scaling instruction-finetuned language

models. *Journal of Machine Learning Research*, 25:1–53.

Edgar Dale and Jeanne S. Chall. 1948. A formula for predicting readability. *Educational Research Bulletin*, 27:1–20, 37–54.

Rudolf Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Gemma Team. 2025. Gemma 3 technical report. Technical report, Google DeepMind.

Anastasia Giachanou, Bilal Ghanem, and Paolo Rosso. 2023. Detection of conspiracy propagators using psycho-linguistic characteristics. *J. Inf. Sci.*, 49(1):3–17.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Damien Lenoir. 2024. Authoritative and epistemic stance in the construction of conspiracy theories: A case study. *ELAD-SILDA*, (9).

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Mattia Samory. 2025. SemEval-2026 Task 10 Starter Pack. `https://github.com/hide-ous/semeval26_task10_starter_pack`. Scripts to facilitate participation in SemEval-2026 Task 10: PsyCoMark.

Mattia Samory, Felix Soldner, and Veronika Batzdorfer. 2025. PsyCoMark – Psycholinguistic Conspiracy Marker Dataset.

Mattia Samory, Felix Soldner, and Veronika Batzdorfer. 2026. SemEval-2026 task 10: PsyCoMark – psycholinguistic conspiracy marker extraction and detection. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

A. Ward. 2022. Textstat.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.

Leon Zrnić. 2024. Conspiracy theory detection using transformers with multi-task and multilingual approaches. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024)*, volume 3740 of *CEUR Workshop Proceedings*, pages 3012–3020. CEUR-WS.org.