Unibuc-NLP at SemEval-2026 Task 10: Unmasking Conspiracies with Pre-Trained Language Models

Teodor-George Marchitan[1, 2], **Liviu P. Dinu**[1,2]
teodor.marchitan@s.unibuc.ro ldinu@fmi.unibuc.ro

[1]University of Bucharest, Faculty of Mathematics and Computer Science, [2]HLT Research Center

Abstract

The paper describes the system submitted by team **Unibuc-NLP** to SemEval-2026 Task 10 (PsyCoMark) Subtask 2: detecting whether a Reddit comment expresses a conspiracy belief. We investigate three modeling paradigms: **(A)** an embedding-and-classify pipeline using Jina-embeddings-v3, HateBERT and BERT-Sentiment with Optuna-tuned classical ML models, optionally enriched by 19 readability features from `textstat`; **(B)** end-to-end fine-tuning of encoder transformers (DeBERTa-v3-base, DistilBERT) with a compact 128-unit classifier head and multiple pooling strategies; and **(C)** parameter-efficient QLoRA fine-tuning of large decoder-only models (Mistral-7B-v0.3, Qwen3-0.6B). Our best system, DeBERTa-v3-base with a 128-dimensional classifier, achieves a weighted F1 of **0.74**, ranking **29/52** on the official leaderboard. Post-submission analysis further reveals that a weighted pooling strategy outperforms [CLS] on the official validation split by **+0.04**, achieving a weighted F1 of 0.78 (rank **8/52**), suggesting that conspiracy-relevant features are distributed across transformer layers rather than concentrated at the final output.

1 Introduction

The spread of conspiracy theories on social media poses significant challenges to public discourse and information integrity. SemEval-2026 (Ghosh et al., 2026) Task 10 (PsyCoMark) (Samory et al., 2026) addresses this with two subtasks grounded in psycholinguistic theory. We focus on **Subtask 2**: binary classification of whether a Reddit comment expresses a conspiracy belief. Unlike prior work restricted to single topics (e.g. COVID-19), PsyCoMark is topic-diverse, spanning subreddits such as `r/conspiracy`, `r/europe`, `r/harrypotter` and annotated using structural markers (*Actor*, *Victim*, *Action*, *Threat*, *Effect*, *Evidence*) that characterize conspiratorial thinking.

Our multi-pronged strategy compares three paradigms. **Approach A** extracts fixed representations from pre-trained models and trains lightweight classifiers, optionally augmented with handcrafted readability features. **Approach B** fine-tunes encoder transformers end-to-end with configurable pooling layers. **Approach C** applies QLoRA (Dettmers et al., 2023) to adapt large language models for classification.

Our best-performing system is DeBERTa-v3-base with weighted layer pooling (F1 0.78 on the official validation split) followed by DeBERTa with [CLS] pooling. Embedding-based classifiers with Optuna tuning (Jina + textstat, HateBERT + BERT-Sentiment + textstat) achieve competitive validation F1 (0.68 - 0.73). Error analysis reveals that the system struggles most with (a) non-conspiracy comments that discuss conspiracy-adjacent topics (UFOs, Epstein) and (b) subtle conspiracy beliefs in non-conspiracy subreddits. Short comments (< 50 words) show lower accuracy.

2 Background and Related Work

2.1 Conspiracy Detection

Prior work on conspiracy detection has largely focused on topic-specific datasets, such as COVID-19 misinformation (Mompelat et al., 2022) or Dravidian-language fake news (Subramanian et al., 2025). These settings allow models to exploit topical cues, which limits generalization. PsyCoMark takes a different stance: it is topic-agnostic and psycholinguistically grounded, requiring systems to detect conspiratorial *structure* (Actor, Victim, Action, Threat, Effect, Evidence) rather than surface vocabulary. This makes the task substantially harder and more realistic, since conspiratorial language can appear in any subreddit and non-conspiratorial text can contain the same surface markers.

A related line of work shows that linguistic complexity correlates with misinformation (Horne and

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1379–1384

Adali, 2017). This motivates our inclusion of 19 readability features via `textstat` (Flesch Reading Ease, Gunning Fog, SMOG Index, etc.) in Approach A: if conspiratorial text exhibits systematically different complexity profiles from non-conspiratorial text, these features should provide a complementary signal on top of semantic embeddings.

2.2 Models and Methods

Pre-trained transformers (Devlin et al., 2019; Liu et al., 2019; He et al., 2021) are the dominant paradigm for text classification. For this task, domain proximity to Reddit data is a natural desideratum. We include HateBERT (Caselli et al., 2021), further pre-trained on 1.5M Reddit comments from banned hateful communities, and Jina-embeddings-v3 (Günther et al., 2023), a general-purpose embedding model with task-specific LoRA adapters. Whether domain proximity outweighs general pre-training quality is an empirical question we directly test in Approach A.

For end-to-end fine-tuning (Approach B), the choice of pooling strategy determines how token-level representations are aggregated into a single sentence vector. Standard choices — [CLS] token extraction and mean pooling — treat all layers and positions uniformly. More expressive alternatives include Multi-Head Attention Pooling (MHAP) (India et al., 2019) and Global MHAP (Wang et al., 2020), originally developed for speaker verification, which compute per-head attention distributions over token states. A further dimension is cross-layer aggregation: Weighted Layer Pooling learns a soft combination across all transformer layers, motivated by the observation that different layers encode different levels of linguistic abstraction. Whether conspiracy-relevant features are concentrated at the final layer or distributed across intermediate ones is an open question that our pooling ablation (Section 4) directly addresses.

For large-model approaches (Approach C), QLoRA (Dettmers et al., 2023) enables fine-tuning under 4-bit quantization combined with LoRA adapters (Hu et al., 2021), making it feasible to adapt billion-parameter decoder models on a single GPU. Given the modest size of the PsyCoMark training set (∼4K samples), the risk of overfitting with such large models is non-trivial, which we examine in our results.

3 System Description

3.1 Dataset and Preprocessing

The PsyCoMark training set contains 4,361 Reddit comments, of which 4,316 remain after removing "Can't tell" entries and unavailable records (Table 1). The official validation set (100 samples) and test set (938 samples) are unlabeled. We create a local 80/20 validation split using stratified sampling (`seed=42`) for internal evaluation.

Split	Raw	Filtered
Train	4,361	4,316
Dev	100	100
Test	938	938

Table 1: Dataset statistics. "Filtered" excludes "Can't tell" entries and unretrievable comments.

3.2 Approach A: Embedding Extraction + Classical ML

We extract fixed-dimensional sentence embeddings from three models: Jina-embeddings-v3, HateBERT and BERT-Sentiment.

Optionally, we augment embeddings with **19 textstat features** (readability scores, word/syllable/sentence counts, complexity metrics), standardized with `StandardScaler`. Classifiers (Logistic Regression, Linear SVM and XGBoost) are tuned via Optuna with 100 trials each, using 5-fold stratified cross-validation with macro F1 as the objective. The best classifier across all three types is selected.

3.3 Approach B: End-to-End Fine-Tuned Transformers

The model architecture stacks a transformer backbone, a pooling layer and a compact MLP classifier. We experiment with two backbones (DeBERTa-v3-base, 86M parameters; DistilBERT-base-uncased, 66M parameters) under identical settings: classifier hidden dimensions , batch size 16, cross-entropy loss with label smoothing 0.1 and differential learning rates (backbone: 10^{-5}, classifier: 2×10^{-4}).

We compare seven pooling strategies: *Mean* (masked average of all token embeddings), *[CLS]* (first token's hidden state), *Attention* (single-head learnable attention over token states), *MHAP* (multi-head attention pooling splitting features across heads), *GlobalMHAP* (each head attends over the full feature vector), *Weighted Layer* (learned combination across transformer layers)

and *Weighted Layer + Attention*. Training uses cosine LR scheduling with 10% warmup and early stopping with patience of 7 epochs.

3.4 Approach C: QLoRA Fine-Tuning of Large LLMs

We fine-tune two decoder-only models (Qwen3-0.6B and Mistral-7B-v0.3) using 4-bit NF4 quantization and LoRA adapters applied to attention projection matrices. LoRA rank is 2 for Qwen3 ($\alpha = 8$) and 4 for Mistral ($\alpha = 16$). Since decoder models process tokens causally, we use left-padding and extract the last non-padding token's hidden state as the sentence representation. Gradient accumulation ensures an effective batch size of 16 across all models.

4 Results

4.1 Official Test Results

Table 2 reports our official Codabench submissions. DeBERTa-v3-base fine-tuned with a 128-dimensional head achieves the best performance among our submitted systems.

System	F1	Rank
DeBERTa [128]	**0.74**	**29/52**
HateBERT+BERT-Sent+textstat	**0.71**	**50/52**
Jina-v3 + textstat	**0.73**	**35/52**
DeBERTa [128] + Weighted Layer†	**0.78**	—

Table 2: Official test results on the SemEval-2026 Task 10 Subtask 2 leaderboard. †Weighted layer pooling was not submitted officially; the score is reported on the official validation split.

4.2 Ablation Study

Table 3 summarizes our internal ablations done on the official validation split. We report the weighted F1 for comparisons.

Effect of textstat features. Augmenting embeddings with readability features provides a modest improvement of $+0.02$ F1 for Jina-v3. The effect is smaller for the HateBERT and BERT-Sentiment combination, suggesting that surface-level complexity signals capture complementary information beyond what domain-specific embeddings encode.

Embedding model choice. HateBERT, pretrained on Reddit communities banned for hateful content, underperforms Jina-v3 and BERT-Sentiment under identical ML classifier settings. This challenges our hypothesis that domain proximity to the PsyCoMark data matters more than general embedding quality.

Backbone comparison. DeBERTa-v3-base outperforms DistilBERT by 0.01 under otherwise identical settings, confirming that disentangled attention provides a meaningful advantage for conspiracy detection, at the cost of $\sim$30% more parameters.

Pooling strategy. Table 4 compares the seven pooling strategies on DeBERTa with a fixed head. The submitted system uses [CLS] pooling. However, post-submission analysis on the validation split reveals that weighted layer pooling achieves the highest F1, outperforming [CLS] by $+0.04$. This suggests that conspiracy-relevant features are distributed across intermediate transformer layers rather than concentrated in the final [CLS] representation, and that learning a soft combination over layers is more informative than relying on a single fixed-depth output. Attention-based pooling methods (Attention, MHAP, GlobalMHAP) offer poor performance compared to [CLS] and weighted layer pooling.

Training technique ablations. Table 5 reports the effect of individual training design choices on DeBERTa [128] + [CLS]. Label smoothing and differential learning rates prove to be the most impactful regularization choices, while freezing the backbone before joint fine-tuning proves worse performance. Batch normalization has a small improvement, while adding an extra L2 normalization to the embeddings degrades model performance by a small margin.

QLoRA LLMs vs. encoder models. Despite having far more parameters, QLoRA-tuned models achieve lower performance than DeBERTa [128]. This highlights the efficiency of encoder models for binary classification on very small sized datasets.

5 Error Analysis

We perform error analysis on the validation set (77 samples with Yes/No labels) comparing our two best systems: DeBERTa with weighted layer pooling (F1 = 0.78) and DeBERTa with [CLS] pooling (F1 = 0.74).

Confusion matrix. Figure 1 shows the confusion matrix for DeBERTa weighted layer pooling model on the validation set. The model achieves 44

Comparison	Configuration	F1	Δ
Effect of textstat (Jina)	Jina-v3 (no textstat)	**0.71**	—
	Jina-v3 + textstat	**0.73**	**+0.02**
Effect of textstat (HateBERT)	HateBERT (no textstat)	**0.68**	—
	HateBERT + textstat	**0.68**	**+0.00**
Effect of textstat (BERT-Sentiment)	BERT-Sentiment (no textstat)	**0.70**	—
	BERT-Sentiment + textstat	**0.70**	**+0.00**
Embedding model (no textstat)	Jina-v3	**0.71**	—
	HateBERT	**0.68**	**-0.03**
	BERT-Sentiment	**0.70**	**-0.01**
Backbone (same head [128])	DeBERTa-v3-base	**0.75**	—
	DistilBERT-base	**0.74**	**-0.01**
QLoRA LLMs	Qwen3-0.6B	**0.18**	—
	Mistral-7B-v0.3	**0.51**	**+0.33**
Training techniques comparison	Jina-v3 textstat + XGBoost	**0.73**	—
	DeBERTa [128] (fine-tuned)	**0.75**	**+0.02**

Table 3: Ablation results on the official validation split. Δ is relative to the first row of each group.

Pooling Strategy	F1	Params
Mean	**0.73**	**86M**
[CLS]	**0.75**	**86M**
Attention	**0.65**	**+0.3M**
MHAP (4 heads)	**0.75**	**+0.5M**
GlobalMHAP (4 heads)	**0.65**	**+0.5M**
Weighted Layer	**0.78**	**+0.01M**
Weighted Layer + Attn	**0.72**	**+0.3M**

Table 4: Pooling strategy comparison on DeBERTa-v3-base [128] (official validation split).

Configuration	F1	Δ
baseline cross-entropy loss Label smoothing = 0.1 differential LR Batch normalization, no L2 normalization	**0.75**	—
Focal loss	**0.72**	**-0.03**
No label smoothing	**0.66**	**-0.09**
No differential LR	**0.66**	**-0.09**
Freeze backbone (5 ep)	**0.72**	**-0.03**
No Batch Normalization	**0.72**	**-0.03**
L2 normalization	**0.74**	**-0.01**

Table 5: Training technique ablations on DeBERTa-v3-base [128] + [CLS] (validation split).

true negatives, 17 true positives, 6 false positives, and 10 false negatives. The higher recall on the "No" class (0.88) compared to "Yes" (0.63) reflects both the class distribution and a tendency to predict "No" when uncertainty is high. [CLS] model produces substantially more false positives (10 vs. 6), suggesting that weighted layer pooling provides more discriminative representations for borderline positive cases.

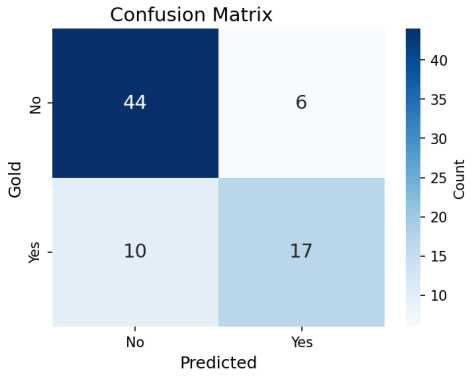

Figure 1: Confusion matrix for DeBERTa with weighted layer pooling on the validation set.

False positives. DeBERTa with weighted layer pooling produces 6 false positives. Common patterns include: (a) *factual discussions* of conspiracy-adjacent topics (e.g. UFOs, Epstein) that mention but do not endorse conspiratorial claims; (b) *short political news snippets* (e.g. arrests of politicians) that the model over-associates with conspiracy; and (c) comments from subreddits such as `r/wanttobelieve` or `r/ufo` where the model conflates topic identity with belief endorsement.

False negatives. Both models produce 10 false negatives, indicating this error type is harder to resolve through pooling choice alone. These cases frequently involve: (a) subtle or implicit conspiracy beliefs in non-conspiracy subreddits (e.g. `r/moderatepolitics`, `r/stupidpol`, `r/Bitcoin`) where conspiratorial framing is less

overt; (b) comments that convey conspiratorial themes without an explicit Actor-Victim-Action structure; and (c) short comments (under 50 words) where contextual cues are sparse.

Subreddit effects. Figure 2 shows per-subreddit F1 for subreddits with at least 3 validation samples. `r/Christianity` achieves perfect F1 (1.0, n=4), reflecting consistently unambiguous non-conspiratorial content. `r/conspiracy` scores 0.80 (n=3), and `r/stupidpol` scores 0.75 (n=4), the latter being a subreddit where conspiratorial framing appears alongside political commentary, making stance detection harder.

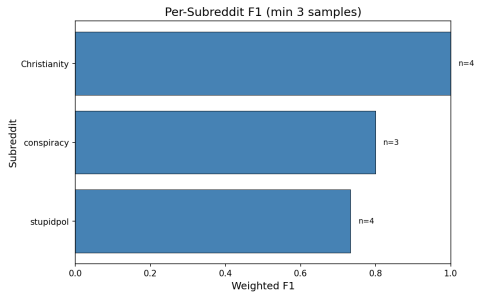

Figure 2: Per-subreddit weighted F1 for DeBERTa with weighted layer pooling (min. 3 samples per subreddit).

Text length effects. Figure 3 shows accuracy by text length bucket. Short comments (<50 words, n=29) achieve 0.76 accuracy, medium comments (50-150 words, n=45) achieve 0.80 and long comments (>150 words, n=3) achieve 1.0. The monotonic improvement with length supports the hypothesis that more context allows the model to identify conspiratorial stance more reliably, while brevity leaves predictions under-determined.

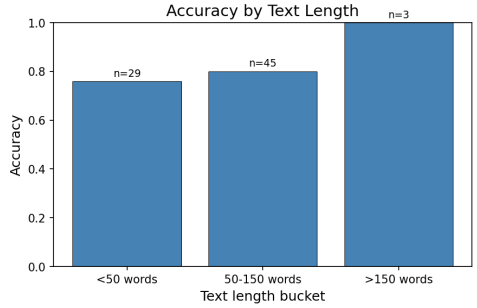

Figure 3: Accuracy by text length bucket for DeBERTa with weighted layer pooling on the validation set.

Example errors. **False Positive 1:** *"Although books have been written about the United States Air Force's war with UFOs, little's been written about the ongoing skirmishes and encounters between otherworldly craft known as USOs and the navies of the world."* - `r/wanttobelieve`, true label: No, predicted label: Yes. The comment describes a factual literary topic rather than endorsing a conspiracy belief. The model likely overweights UFO-related vocabulary and the subreddit's name.
False Negative 1: *"after the past months of protests, domestic terror charges, and even a death, the city of Atlanta has finally approved for the funding of the police training facility that critics call 'Cop City'... Do you agree with the city greenlighting this construction, are the protestors fears legitimate?"* - `r/moderatepolitics`, true label: Yes, predicted label: No. The comment frames state action in conspiratorial terms (protesters vs. state violence) but does so through a question rather than explicit endorsement, evading the model's detection.
False Positive 2: *"Jeffrey Epstein's famous little black book, full of hundreds of well-known elites [...] It has been said he is an intelligence asset known for his vast collection of honey-trap blackmail [...] WWG1WGA"* - `r/TruthLeaks`, true label: No, predicted label: Yes. Despite QAnon-associated phrasing, annotators labeled this as non-endorsement. The model conflates the presence of conspiracy-adjacent language with expressed belief, highlighting the difficulty of distinguishing *performance* from *belief*.

6 Conclusion

We presented a multi-strategy system for SemEval-2026 Task 10 Subtask 2, exploring three complementary paradigms for conspiracy detection in Reddit comments. Our best submitted system, DeBERTa-v3-base with [CLS] pooling and a compact classifier head, achieved **F1 = 0.74** (rank **29**). Post-submission analysis further shows that weighted layer pooling raises validation F1 to 0.78 (vs. 0.74 for [CLS]), confirming cross-layer aggregation as a productive direction. The fixed embedding with classical ML pipeline with jina-v3 embeddings augmented by textstat readability features provides a strong, lightweight alternative. QLoRA-tuned LLMs achieve very poor performance with a very strong sign of overfitting on the training set.

Error analysis on the validation set reveals three principal failure modes: (a) *factual discussions* of conspiracy-adjacent topics (UFOs, Epstein) that the model misclassifies as endorsements; (b) *im-*

plicit conspiratorial framing in general-interest subreddits, where conspiratorial stance is conveyed through presupposition or questioning rather than explicit claim; and (c) *short comments* (<50 words, accuracy 0.76) that lack sufficient context for reliable stance detection. The higher false negative rate on "Yes" (recall 0.63 vs. 0.88 for "No") underscores that conspiracy *belief* is harder to detect than its absence, particularly when expressed subtly.

Future directions include: **(a)** multi-task learning using Subtask 1 span annotations as auxiliary signal; **(b)** data augmentation through paraphrasing or back-translation to improve coverage of implicit conspiracy patterns; **(c)** prompt-based classification with instruction-tuned LLMs for few-shot generalization to unseen subreddits; and **(d)** addressing the low-context challenge for short comments (<50 words) through external context augmentation — such as retrieving the parent post or comment thread — or prompt-based methods that inject task-relevant priors, which may recover the stance signal absent from the comment itself.

Acknowledgments

This research is supported by the project "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization and Financial Instruments Program, 2021-2027, MySMIS no. 351416.

References

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Debanjan Ghosh, Kai North, Ekaterina Kochmar, Mamoru Komachi, and Marcos Zampieri, editors. 2026. *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics, San Diego, United States.

Michael Günther, Louis Milliken, Jonathan Geuter, Georgios Mastrapas, Bo Wang, and Han Xiao. 2023. Jina embeddings: A novel set of high-performance sentence embedding models. *Preprint*, arXiv:2307.11224.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. *Preprint*, arXiv:2006.03654.

Benjamin D. Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *CoRR*, abs/1703.09398.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Miquel India, Pooyan Safari, and Javier Hernando. 2019. Self multi-head attention for speaker recognition. *Preprint*, arXiv:1906.09890.

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. *Preprint*, arXiv:1907.11692.

Ludovic Mompelat, Zuoyu Tian, Amanda Kessler, Matthew Luettgen, Aaryana Rajanala, Sandra Kübler, and Michelle Seelig. 2022. How "loco" is the LOCO corpus? annotating the language of conspiracy theories. In *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) within LREC2022*, pages 111–119, Marseille, France. European Language Resources Association.

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)*.

Malliga Subramanian, Premjith B, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Balasubramanian Palani, and Bharathi Raja Chakravarthi. 2025. Overview of the shared task on fake news detection in Dravidian languages-DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 759–767, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Zhiming Wang, Kaisheng Yao, Xiaolong Li, and Shuo Fang. 2020. Multi-resolution multi-head attention in deep speaker embedding. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6464–6468.