TTLab at SemEval-2026 Task 10: Transformer-based Approaches for Psycholinguistic Conspiracy Detection in Social Media Discourse

Samuel Richer[1], **Mounika Marreddy**[1], **Alexander Mehler**[1]
[1]Goethe University, Frankfurt am Main, Germany
s1065895@stud.uni-frankfurt.de,
mmarredd@em.uni-frankfurt.de, mehler@em.uni-frankfurt.de

Abstract

Online platforms increasingly host conspiracy narratives that shape public debate, reduce trust in institutions, and contribute to polarization, highlighting the need for reliable automatic detection systems. In this paper, we participate in SemEval-2026 Task 10 (PsyCoMark), focusing on conspiracy detection in Reddit conversations using transformer-based models. We evaluate four approaches: raw text, structured psycholinguistic markers, a combined representation, and a stacking ensemble. Our results show that marker-based representations outperform text-only models, and that ensembling further improves robustness. These findings demonstrate the value of incorporating structured psychological cues for scalable conspiracy detection.

1 Introduction

As social media platforms continue to expand, conspiracy-related content spreads rapidly and reaches increasingly large audiences. While online discussions enable open exchange of ideas, they also provide space for narratives that promote distrust, hidden agendas, and oversimplified explanations of complex events. Over time, such narratives can shape public opinion, weaken trust in institutions, and intensify social and political polarization. Given the massive volume of online content, manual monitoring is impractical, making automatic detection systems essential.

However, detecting conspiracy discourse is not a straightforward text classification problem. Conspiracy narratives typically rely on subtle psychological cues, including the attribution of blame to powerful actors, the suggestion of hidden intentions, and the presentation of selective evidence. Such elements are often not captured by surface-level lexical features, thereby limiting the effectiveness of approaches that rely solely on raw text. Prior work on related misinformation tasks supports this observation. Giachanou et al. [4] show that psycholinguistic patterns provide stronger signals for distinguishing fake news spreaders from fact checkers than lexical features alone, while Giachanou et al. [3] demonstrate that integrating word embeddings with psycholinguistic characteristics improves the detection of conspiracy propagators compared to text-only baselines. In addition to user-level signals, Shahsavari [8] show that conspiracy narratives exhibit recurring structural patterns involving identifiable actors, motivations, and relationships, which can be computationally extracted to support detection. These findings highlight the need for more structured approaches that capture the underlying psychological mechanisms of such narratives.

SemEval 2026 Task 10 (PsyCoMark) [9] addresses this challenge by introducing psycholinguistic conspiracy markers in Reddit conversations. Beyond document-level labels,

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 727–734

the dataset provides span-level annotations for five psychological mechanisms: Actor, Action, Effect, Victim, and Evidence. This structured design enables models to move beyond traditional text classification and incorporate psychologically meaningful representations into the detection process.

In this work, we evaluate transformer-based models using four approaches: raw text fine-tuning, marker-only structured representations, a combined input format, and a stacking ensemble that integrates multiple transformer models via logistic regression. Our results show that marker-based representations outperform raw text alone, and that combining models through an ensemble achieves the best overall performance. These findings highlight the importance of structured psycholinguistic cues for improving conspiracy detection.

2 Dataset

The PsyCoMark dataset is designed to support research on detecting conspiracy related discourse in online discussions. It consists of more than 4,800 annotated Reddit submissions, covering over 4,100 unique posts collected from more than 190 subreddits.

Each instance is labeled at the document level as conspiracy, not conspiracy, or can't tell. The dataset includes 1,715 conspiracy instances, 2,263 not conspiracy instances, and 877 can't tell instances. In addition, span level annotations are provided for five psycholinguistic mechanisms: Actor, Action, Effect, Victim, and Evidence. These markers capture common narrative components of conspiracy discourse.

The data were collected between January and March 2025, while the original Reddit content spans from 2013 to 2023. Preprocessing included length filtering, markdown normalization, URL masking, and removal of quoted text. Annotations were performed by native English speaking crowdworkers, achieving moderate inter annotator agreement with a Krippendorff's alpha of 0.58.

3 Approaches

In this section, we describe the four approaches evaluated in our study. For each approach, we fine-tuned five transformer-based language models obtained from HuggingFace.

3.1 Models

We fine-tune five pretrained transformer models: BERT [2], DistilBERT [7], RoBERTa [5], XLNet [10], and ELECTRA [1]. These models differ in their pretraining objectives, including masked language modeling, permutation-based modeling, and replaced-token detection. This diversity allows us to assess whether performance improvements are consistent across architectural variations.

Detailed descriptions of the individual model architectures and hyperparameters are provided in the Appendix A and D

3.2 Text as Feature

As a baseline, each transformer model is fine tuned using only the raw Reddit comment as input. The task is formulated as a binary sequence classification problem in which the model predicts whether a submission contains conspiracy related content. Instances labeled as *can't tell* are excluded, as the inclusion of uncertain annotations would introduce noise into the binary classification setting. All models are trained independently under identical hyperparameter configurations. This setting evaluates the extent to which standard transformer models can detect conspiracy discourse using only surface level lexical and contextual information.

3.3 Markers as Feature

The PsyCoMark dataset provides span-level annotations for five psycholinguistic mecha-

nisms: Actor, Action, Effect, Victim, and Evidence. These markers capture recurring structural elements of conspiracy narratives, such as identifying a responsible agent, describing harmful actions, and highlighting affected groups.

Instead of using the raw comment, we reformulate each instance into a structured textual representation composed only of its extracted markers. Each marker type is explicitly labeled, and may contain zero, one, or multiple spans. This structured format serves as the sole input to the classifier.

By removing surface-level context and retaining only psychologically motivated components, this approach tests whether conspiracy detection can be driven primarily by narrative structure rather than lexical cues.

An example of this input format is shown in Appendix B

3.4 Text and Markers as features

The next step was to combine the two ideas from above. In this setting, the input consists of both the raw Reddit comment and the extracted psycholinguistic markers. A full example of the combined input format is shown in Appendix C

This hybrid representation allows the models to simultaneously access structured semantic cues from the markers and the full contextual information of the original comment. The markers highlight psychologically relevant elements while the raw text preserves linguistic and contextual nuances that may otherwise be lost.

Since gold marker annotations were not available for the official test set, the Marker and Both configurations are treated as oracle experiments that estimate an upper bound on performance under the assumption that psycholinguistic annotations are available.

3.5 Ensemble Method

To further improve robustness, we employ a stacking ensemble strategy. The five fine-tuned transformer models serve as base learners. Their prediction probabilities are used as input features to a logistic regression meta-learner.

Base models are trained on 80% of the dataset. Predictions generated on the remaining 20% are used to train and evaluate the meta-classifier using an additional internal split. This approach enables the meta-learner to capture complementary strengths across architectures and reduce variance.

Detailed training dynamics are provided in Appendix E

4 Results

Model performance is evaluated using Macro F1-score, following the official evaluation protocol of the shared task.

4.1 Quantitative Results

Table 1 reports performance across the three input representations, and Table 2 presents the ensemble results.

The Text-only setting yields the lowest performance across models. BERT achieves the strongest result in this category with an F1-score of 0.759, indicating that lexical and contextual information alone is insufficient for reliable conspiracy detection.

Using only structured psycholinguistic markers leads to a substantial improvement. In the Marker setting, BERT reaches an F1-score of 0.817, representing a gain of nearly six F1 points over the Text-only baseline. This demonstrates that narrative structure provides stronger predictive signals than surface-level lexical patterns.

The combined Text and Marker representation maintains consistently high performance. While BERT performs similarly in the Marker

Model	Approaches								
	Text Approach			Marker Approach			Both Approach		
	P	R	F1	P	R	F1	P	R	F1
distilbert	0.744	0.741	0.748	0.812	0.807	0.809	0.810	0.815	0.812
roberta	0.744	0.753	0.745	0.759	0.766	0.761	0.810	0.801	0.805
xlnet	0.728	0.727	0.727	0.779	0.788	0.782	0.811	0.825	0.806
bert	0.761	0.757	**0.759**	0.822	0.814	**0.817**	0.815	0.829	**0.816**
electra	0.733	0.740	0.735	0.776	0.784	0.778	0.806	0.792	0.797

Table 1: Performance comparison across approaches

Model	P	R	F1
Ensemble	0.848	0.835	**0.842**

Table 2: Performance of the ensemble method

(0.817) and Both (0.816) settings, other architectures benefit from the hybrid input. For example, XLNet improves from 0.782 to 0.806, and RoBERTa improves from 0.761 to 0.805 when raw text is incorporated. This suggests that lexical context provides complementary information, particularly for models that benefit from richer contextual cues.

Notably, the small performance gap between the Marker and Both settings indicates that structured psycholinguistic cues capture most of the predictive signal required for conspiracy detection in this dataset. This finding suggests that conspiracy discourse is characterized more by recurring narrative structure than by specific vocabulary.

The highest overall performance is achieved by the stacking ensemble, which attains an F1-score of 0.842. By integrating predictions from all five transformer models, the meta-learner captures complementary strengths across architectures and improves robustness beyond any individual model.

For the official leaderboard submission, we employed the text only configuration, as the test set did not include marker annotations.

4.2 Error Analysis

To better understand model behavior, we examine confusion matrices in Appendix F and also give examples for representative misclassifications in Table 3).

First, Text-only models rely heavily on lexical triggers such as *government*, *vaccine*, or *UFO*. Consequently, they sometimes misclassify neutral references to conspiracy topics as conspiratorial claims.

Second, the Marker-only setting exhibits structural overgeneralization. Instances containing an identifiable Actor performing an Action affecting a Victim are occasionally labeled as conspiracies even when the context is scientific or historical.

Third, the combined representation tends to increase false positives when emotionally charged vocabulary co-occurs with a conspiratorial narrative structure. In such cases, the interaction between lexical emphasis and structural cues may lead to overprediction.

Finally, politically expressive or activist language is frequently misclassified as conspiratorial, particularly when institutional actors are portrayed as responsible for negative outcomes.

Approach	Example	True	Pred	Reason
Text	"In the ultimate irony, these two evangelical Christians... are pretty well in line with most of the conspiracy theories we discuss on this board."	no	yes	The model sees the word "conspiracy" and ignores the conversational context
Marker	"[ACTOR: Lidar] [ACTION: penetrate] [EFFECT: revolutionized what we know about ancient Maya.]"	no	yes	The model mistakes a laser-scanning technology for a conspiratorial force
Both	"The five books of Moses... are symbolic and allegorical... attaining reality and its single force by purposeful, methodical inner changes... to understand the world around us behind the limited, superficial picture we perceive today."	no	yes	For the model, this looks structurally like a classic conspiracy reveal: Moses unveiling a truth.
Ensemble	"Australia is pushing further and further into full blown Authoritarianism... People need to wake up to what is happening."	no	yes	Model flags as a conspiracy, even if it's framed as political protest

Table 3: Wrong predictions by the different models

Overall, the Text-only, Marker-only, and Ensemble settings maintain relatively balanced false positive and false negative rates. In contrast, the combined representation shows a higher number of false positives, indicating increased sensitivity but reduced precision in ambiguous contexts.

5 Limitations and Future Work

Our marker based approach depends on gold span annotations, which may not be available in real world applications. Future research should therefore investigate end to end models that automatically predict markers while performing conspiracy detection. Additionally, the dataset is restricted to English Reddit posts, limiting generalizability across platforms and languages. The moderate inter annotator agreement further reflects the subjective and ambiguous nature of conspiracy discourse. Finally, while the ensemble enhances performance, its reliability depends on the availability of sufficient data. Specifically, robust training of the meta learner requires a sufficient number of outputs from the base learners, making the approach less suitable for low resource scenarios. Logistic regression was selected as the meta learner due to its simplicity and its lower risk of overfitting given the limited number of base model outputs. Future work should explore alternative meta learners to determine whether more complex models can further improve performance.

Conclusion

In this work, we evaluated transformer-based methods for psycholinguistic conspiracy detection in the PsyCoMark task. Structured marker representations clearly outperform text-only baselines, demonstrating that narrative structure provides strong predictive signals. While combining text and markers offers complemen-

tary benefits, the stacking ensemble achieves the best overall performance. Overall, our findings emphasize the value of incorporating structured psychological cues into automated conspiracy detection systems.

References

[1] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

[3] Anastasia Giachanou, Bilal Ghanem, and Paolo Rosso. 2023. Detection of conspiracy propagators using psycho-linguistic characteristics. *Journal of Information Science*, 49:3–17.

[4] Anastasia Giachanou, Bilal Ghanem, Esteban A. Ríssola, Paolo Rosso, Fabio Crestani, and Daniel Oberski. 2022. The impact of psycholinguistic patterns in discriminating between fake news spreaders and fact checkers. *Data Knowledge Engineering*, 138:101960.

[5] 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*.

[6] 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)*.

[7] 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*.

[8] Shadi Shahsavari. 2022. Automated conspiracy theory detection and narrative consensus tracking in social media.

[9] Marco Valentino, Leonardo Ranaldi, Giulia Pucci, Federico Ranaldi, and André Freitas. 2026. Semeval-2026 task 11: Disentangling content and formal reasoning in large language models. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

[10] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

A Model Architectures

This section provides brief descriptions of the pretrained transformer models used in our experiments.

BERT is a bidirectional transformer trained using masked language modeling and serves as a strong baseline for many classification tasks. [2]

DistilBERT is a compressed version of BERT obtained through knowledge distillation. It maintains competitive performance while being smaller and computationally more efficient. [7]

RoBERTa is an optimized reimplementation of BERT trained on larger corpora with improved pretraining strategies, often achieving stronger performance. [5]

XLNet is a permutation-based autoregressive transformer that captures bidirectional context without masked language modeling. [10]

ELECTRA uses a replaced-token detection objective, which improves sample efficiency during pretraining and leads to strong classification performance. [1]

B Input Format for Marker approach

In psychology there are "psycholinguistical markers" that can indicate whether a given text

is probably conspiracy-related or not. There are five of those:

Actor: Who is allegedly responsible for a malicious action or agenda?

Action: What is the actor doing or planning to do to cause negative outcomes?

Effect: What are the negative consequences of the actor's agenda?

Victim: Who is negatively affected by the actor's agenda?

Evidence: Which arguments or expressions does the writer use to support claims?

The dataset provides them extracted from the raw text. Therefore, instead of using the raw Reddit comment as input, we reformulate each instance into a structured textual representation of its markers, for example:

Original

```
"A great article on what's taking
place in Bolivia, referencing some
similar US backed coups in the
region as well as recounting some
of Bolivia's history and western
policy towards the country."
```

Reformulated into:

```
"[ACTOR] [US]
[ACTION] [coups]
[EFFECT] []
[EVIDENCE] [article]
[VICTIM] [Bolivia, "Bolivia's"]"
```

C Input Format for Text and Marker approach

This section shows an example of what the input looks like for the text and marker approach. Both markers and the raw text are concatenated into one string:

```
"[ACTOR] [US]
[ACTION] [coups]
[EFFECT] []
[EVIDENCE] [article]
[VICTIM] [Bolivia, "Bolivia's"]
[COMMENT] A great article on what's
taking place in Bolivia, referencing
some similar US backed coups in the
region as well as recounting some of
Bolivia's history and western policy
towards the country."
```

D Hyperparameters

Learning rate	2×10^{-5}
Weight decay	0.01
Batch size	16
Epochs	5
Test split	10%

Table 4: Training hyperparameters. For the ensemble method, models were trained for 10 epochs with a 20% test split.

E Training Details for Ensemble Method

We trained the five different language models at once with the same approach as in 3.4 and used their output as input for a Logistic Regression meta learner.

Instead of only five, the models were trained for ten epochs. Checkpoints were saved every 100 training steps, and only the best-performing checkpoint, based on validation F1-score, was retained for each model. Figure 1 illustrates the training dynamics of all models in terms of F1-score.

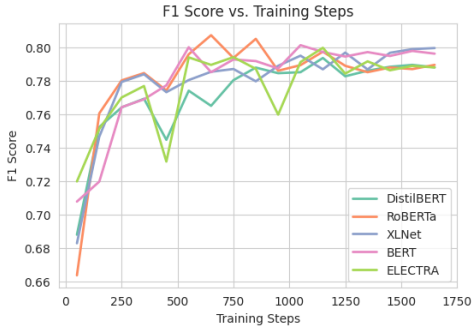

Figure 1: This plot shows the process how the five different models performed during training

After fine-tuning on 80% of the dataset (2,832 samples), the base models were used to generate predictions for the remaining 20% (708 samples). From this subset, 80% (566 samples) were used as training data for the logistic regression meta-learner, while the remaining 20% (141 samples) served as its evaluation set.

F Confusion Matrizes

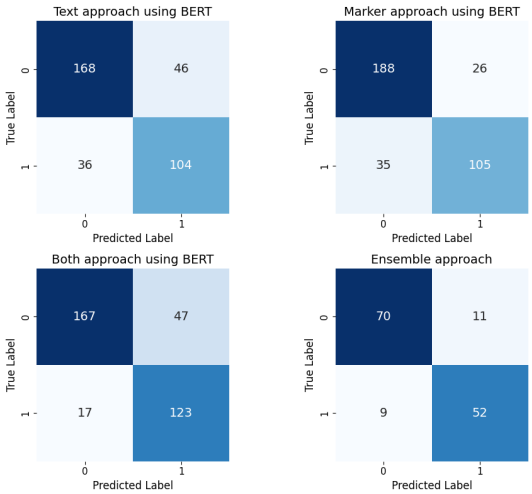

Figure 2: Confusion matrices for the different classification approaches. Labels 0 and 1 denote "No" and "Yes" respectively.