AAA at SemEval-2026 Task 10: Psycholinguistic Conspiracy Marker Extraction and Detection

Xintong Pan
University of Tübingen
xintong.pan@student.uni-tuebingen.de

Abstract

This article presents our study on task 10: Psycholinguistic conspiracy marker extraction and detection(Ghosh et al., 2026), which includes token-level extraction tasks and sentence-level conspiracy detection tasks. Focusing on conspiracy theory texts on social media, this paper proposes a classification method that combines semantic encoding with large language model reasoning and generation. Semantic features are extracted using DeBERTa-v3, and explanatory reasoning text is generated through ConspEmoLLM-v2. The two are then combined for classification, thereby enhancing the model's ability to recognize implicit conspiratorial logic. For the extraction subtask, this paper provides systematic comparison results of several mainstream pre-trained models, mainly conducting baseline model comparisons and performance analysis.

1 Introduction

With the rapid development of social media platforms, conspiracy theory content has spread widely. Task 10 focuses on identifying and analyzing the psycholinguistic markers of conspiracy theories, combining the insights of psychology and natural language processing. The task aims to reveal the words related to conspiracy in the daily text and clearly model the structural components behind the conspiracy narrative.

The task consists of two subtasks.

Conspiracy marker extraction. This subtask involves conspiracy marker extraction, which is formulated as a token-level sequence mark problem. Given an input text, the model needs to assign semantic labels to a single token, including Actor, Action, Effect, Evidence and Victim.

Conspiracy detection. This subtask is the sentence-level binary classification task. Given a text sample, the model determines whether the content expresses conspiracy and outputs binary predictions of Yes or No.

Unlike ordinary misinformation, conspiracy theories often do not directly present false facts but instead construct narrative frameworks through suggestive language, causal inferences, emotionally reinforced expressions, and skepticism toward authoritative institutions (Zelalem and Guest, 2021).

In recent years, text classification methods based on pretrained language models have achieved significant results in various natural language processing tasks. However, in tasks like detecting and extracting conspiracy theories that require handling implicit expressions, merely capturing lexical and semantic information may not be sufficient to reveal the underlying conspiratorial logic in the text. Looking at prior work (Pustet et al., 2024), relying solely on keywords or a single model cannot overcome contextual bias, highlighting the potential of large language models in handling more complex language structures. (Weller et al., 2025) show that although encoder-only models generally outperform decoder-only models of comparable parameter size in classification tasks, the latter have unique advantages in generating explanatory content and capturing complex reasoning patterns.

Meanwhile, ConspEmoLLM introduces emotional features and large language model fine-tuning to propose a conspiracy theory detection method that integrates emotion and semantics, demonstrating the importance of incorporating emotional information to improve model performance (Liu et al., 2024). The latest ConspEmoLLM-v2 further extends this approach on standard datasets, proving the robustness and stability of emotion-driven large language models in handling conspiracy text with altered emotional content (Liu et al., 2025).

Based on this context, this paper focuses on two subtasks, conspiracy classification and extraction. In particular, we focus on the design of models that incorporate enhanced reasoning in the classification task. We use a framework that combines semantic

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1988–1991

encoding with explanation-generating reasoning, using explanatory text generated by large language models as auxiliary features to enhance conspiracy detection capability.

2 Dataset

The dataset used in this experiment is the PsyCoMark dataset. This dataset is constructed based on comments from the Reddit platform and contains over 4,100 individual texts and approximately 4,800 annotations, covering more than 190 different subreddits.

Each instance in the dataset corresponds to a comment. The comment mode includes sentence-level tags for conspiracy detection and token-level span comments for conspiracy tag extraction.

For the detection subtask, each instance is annotated with one of three labels: Yes, No or Can't tell. The label Yes indicates that the text clearly expresses the conspiracy narrative, while the No indicates non-conspiracy content. When the semantic content of the text is ambiguous and the annotator cannot confidently determine whether it constitutes a conspiracy theory, the sample will be labelled by Can't tell.

For the extraction subtask, each tag is annotated as a character-level span, and one of the five predefined types is assigned: Actor, Action, Victim, Effect, Evidence. In addition, the dataset contains instances with empty tag sets. Such situations mainly occur in two situations. When the conspiracy label is No, the missing mark naturally reflects the lack of recognizable conspiracy components and constitutes effective training data for detecting subtasks. In contrast, instances marked as Can't tell reflect the uncertainty of the annotator and may adversely affect the training of the two subtasks. Therefore, in the subsequent experiments, these samples were not used.

3 Experiments

3.1 Data augmentation

Given the limited size of the dataset, we applied synonym replacement based on WordNet to generate augmented samples, following the Easy Data Augmentation (EDA) framework of text classification (Wei and Zou, 2019). This approach increased the size of the dataset by approximately 2,000 instances. However, despite the increased training data, this augmentation strategy did not improve the performance on the test set.

3.2 Conspiracy detection

Before building the reasoning-enhanced framework, we conduct systematic comparison experiments on several mainstream pre-trained language models to select the model most suitable as the backbone semantic encoder. This section will provide a detailed introduction to the model selection process, experimental setup, and analysis conclusions. In this experiment, we selected four pre-trained models widely used in text classification tasks as candidate encoders: DistilBERT (Sanh et al., 2019), BERT-base (Devlin et al., 2019), RoBERTa-base (Liu et al., 2019), and DeBERTa-v3-base (He et al., 2021). The experimental settings are shown in Table 1.

Parameter	value
Number of epochs	15
batch size	16
Learning rate	2×10^{-4}
Optimizer	adam

Table 1: Training Hyperparameters

Table 2 shows the evaluation results of these four encoder models on the test set. The F1 score is the standard used officially in the competition. In addition, we also report the scores for accuracy, recall, and precision. Overall, all models achieved good scores, with performance roughly consistent and F1 scores ranging from 0.745 to 0.755. Among them, DeBERTa achieved the highest score, so DeBERTa was chosen as the semantic encoder model for subsequent experiments.

(Wei et al., 2022) proposed that explicitly generating reasoning steps can significantly improve performance on complex inference tasks. (Kojima et al., 2022) further pointed out that even under zero-shot conditions, appropriate prompting can elicit the model's reasoning capabilities. These studies indicate that large language models has the ability for implicit reasoning. Meanwhile, explicit reasoning generation helps externalize this knowledge structure, which is suitable for our experiment on conspiracy judgment.

Based on the above studies, we use DeBERTa-v3 as the base encoder model and leverage the large language model ConspEmoLLM-v2, adapted to the context of conspiracy narratives, to generate explanatory reasoning texts for input texts. Unlike traditional approaches that generate text vectors solely using the encoder, this experiment combines

Model	F1 score	Precision	Recall
BERT	0.753	0.745	0.755
DeBERTa	0.755	0.752	0.753
RoBERTa	0.745	0.745	0.747
DistilBERT	0.747	0.743	0.747

Table 2: Comparative Performance of Pretrained Encoders

the semantic vectors generated by the encoder with the reasoning text vectors produced by the LLM, retaining a lightweight supervised classifier while enhancing reasoning capabilities, ensuring the stability and efficiency of the model training process.

The experimental results in Table 3 show that after using reasoning representations, the model achieves a consistent improvement in F1 score compared to systems that only use semantic encoders, validating the effectiveness of the reasoning-enhanced strategy in conspiracy detection tasks.

Metric	Score
Accuracy	0.845
F1 score	0.849
Precision	0.853
Recall	0.845

Table 3: Performance Metrics of Hybrid Model

3.3 Conspiracy marker extraction

Unlike classification tasks, the extraction subtask aims to identify components in the text related to conspiracy narratives, including Actors, Actions, Victims, Effects, and Evidence. Given that the main innovation of our study focuses on the reasoning-enhanced classification framework, the extraction task part did not introduce additional structural improvements, but instead used mainstream pre-trained language models for sequence labeling modeling.

In the experiments on the extraction subtask, the performance of various models varied significantly across different element categories. Overall, all models achieved relatively high accuracy, but the F1 scores were significantly lower than the accuracy, indicating a class imbalance problem in the data. A large number of unannotated tokens inflate the accuracy metric, while recognizing entity categories remains challenging.

The results show that the performance differences among different encoder models on the extraction task are relatively limited, with overall stable performance. In contrast, the performance improvement brought by introducing the reasoning-enhanced module in the classification task is more significant, further confirming the importance of explicit reasoning information for recognizing conspiracy narratives.

4 Conclusion

This project proposes a hybrid classification framework enhanced by large language model reasoning. This approach effectively combines semantic encoding capabilities with explicit reasoning abilities, thereby improving the model's ability to identify implicit features of conspiracy narratives. In the classification task, experimental results show that after introducing the reasoning enhancement module, the model achieves stable improvements in metrics such as accuracy and F1 score. For the extraction subtask, this study provides comparisons using various mainstream pre-trained models, establishing a reproducible baseline for future research.

Although this study has achieved certain results, there are still several limitations. The quality of generated reasoning text depends on the capabilities of the large language model itself, and any biases in the generated content may interfere with the final representations.

Future research will focus on extending reasoning enhancement to multilingual scenarios to verify the generalizability of the method.

References

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Category	Model	Accuracy	F1	Precision	Recall
Action	RoBERTa	0.936	0.368	0.480	0.298
	DeBERTa	0.938	0.372	0.491	0.299
	BERT	0.934	0.383	0.480	0.319
	DistilBERT	0.931	0.346	0.443	0.287
Actor	RoBERTa	0.923	0.642	0.711	0.586
	DeBERTa	0.927	0.666	0.740	0.606
	BERT	0.915	0.629	0.658	0.602
	DistilBERT	0.921	0.649	0.691	0.611
Effect	RoBERTa	0.944	0.371	0.512	0.291
	DeBERTa	0.941	0.377	0.453	0.322
	BERT	0.941	0.375	0.481	0.307
	DistilBERT	0.939	0.353	0.453	0.290
Evidence	RoBERTa	0.915	0.321	0.495	0.238
	DeBERTa	0.907	0.312	0.428	0.246
	BERT	0.915	0.333	0.512	0.247
	DistilBERT	0.913	0.323	0.486	0.242
Victim	RoBERTa	0.969	0.407	0.497	0.344
	DeBERTa	0.968	0.388	0.470	0.330
	BERT	0.966	0.361	0.443	0.304
	DistilBERT	0.969	0.343	0.506	0.259

Table 4: Performance Comparison of Pretrained Encoders on the Extraction Subtask

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.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Wei Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *2021 International Conference on Learning Representations*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *ICML 2022 Workshop on Knowledge Retrieval and Language Models*.

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*, abs/1907.11692.

Zhiwei Liu, Boyang Liu, Paul Thompson, Kailai Yang, and Sophia Ananiadou. 2024. *ConspEmoLLM: Conspiracy Theory Detection Using an Emotion-Based Large Language Model*. IOS Press.

Zhiwei Liu, Paul Thompson, Jiaqi Rong, and Sophia Ananiadou. 2025. *ConspEmoLLM-v2: A Robust and Stable Model to Detect Sentiment-Transformed Conspiracy Theories*. IOS Press.

Milena Pustet, Elisabeth Steffen, and Helena Mihaljevic. 2024. Detection of conspiracy theories beyond keyword bias in german-language telegram using large language models. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, page 13–27. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *NeurIPS 2019 Workshop on Energy Efficient Machine Learning and Cognitive Computing*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Orion Weller, Kathryn Ricci, Marc Marone, Antoine Chaffin, Dawn Lawrie, and Benjamin Van Durme. 2025. Seq vs seq: An open suite of paired encoders and decoders. *arXiv:2507.11412*.

Zecharias Zelalem and Peter Guest. 2021. Why facebook keeps failing in ethiopia. Accessed: 2026-02-27.