zhangpeng at SemEval-2026 Task 10: PsyCoMark - A Transformer-based Approach for Psycholinguistic Conspiracy Marker Extraction and Detection

Peng Zhang[1,2], **Gehao Lu**[1,2]
[1]School of Information Science and Engineering, Yunnan University
[2]Yunnan Province Smart Tourism Engineering Research Center, Yunnan University
Kunming 650500, Yunnan, China
[1]zpp1219@gmail.com, [2]glu@ynu.edu.cn

Abstract

We describe our system for SemEval-2026 Task 10 on psycholinguistic conspiracy marker extraction and conspiracy detection from English texts (Ghosh et al., 2026). The shared task consists of two subtasks: (1) extracting conspiracy-related markers—actor, action, effect, victim, and evidence—evaluated using an overlap-based macro F1-score, and (2) detecting conspiracy content as a binary text classification problem evaluated using macro-averaged F1-score. Our approach relies on fine-tuning pre-trained transformer encoders, including multilingual DistilBERT variants and DeBERTa-v3, without using external corpora or data augmentation techniques. Experimental results show that our best models achieve a macro-F1 score of 0.1476 for Subtask 1 and a Weighted-F1 score of 0.7267 for Subtask 2. These results show that simple fine-tuning of pre-trained models provides a strong baseline for both marker extraction and conspiracy detection.

1 Introduction

Conspiracy theories have become increasingly prevalent in online discourse, influencing public opinion, social trust, and decision-making processes (Douglas et al., 2017). Understanding how conspiracy beliefs are expressed in language is therefore an important problem for natural language processing (NLP) and computational social science. SemEval-2026 Task 10: Psycholinguistic Conspiracy Marker Extraction and Detection (PsyCoMark) (Samory et al., 2026) aims to advance research in this direction by introducing a shared task that focuses on identifying the psycholinguistic structure underlying conspiracy narratives rather than treating conspiracy detection as a purely topical classification problem.

The task consists of two complementary subtasks. Subtask 1 requires extracting spans of text that express psycholinguistic conspiracy markers grounded in evolutionary psychology, including actor, action, effect, victim, and evidence. Each document may contain zero or multiple overlapping spans, making the task a challenging sequence labeling problem. Subtask 2 focuses on conspiracy detection, where systems classify Reddit comments as conspiracy-related or not. Both subtasks use English data collected exclusively from Reddit, capturing diverse and complex conspiracy phenomena in real-world online discourse.

To address these challenges, we adopt a unified transformer-based fine-tuning framework for both subtasks. Our approach leverages multilingual and English pre-trained language models, including DistilBERT and DeBERTa variants, and fine-tunes them using only the provided training data without external corpora or data augmentation. We formulate Subtask 1 as a token-level sequence labeling problem and Subtask 2 as a binary text classification task. This unified modeling paradigm enables a simple and reproducible system while maintaining competitive performance across both tasks.

Our contributions can be summarized as follows: (1) we present a unified transformer-based framework that jointly addresses marker extraction and conspiracy detection; (2) we provide empirical evaluation of multiple pre-trained language models for psycholinguistic conspiracy analysis; and (3) we release our implementation to support reproducibility.[1]

2 Related Work

Recent progress in natural language processing (NLP) has been largely driven by pre-trained transformer models, which provide contextualized representations that significantly improve performance across a wide range of tasks. Our work is closely related to two major research directions: sequence

[1]https://github.com/zpp1219/SemEval-2026-Task-10

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 755–760

labeling for span-level information extraction and text classification for document-level prediction.

2.1 Sequence Labeling and Named Entity Recognition

Sequence labeling (Akhundov et al., 2018) is a fundamental NLP paradigm that aims to identify and classify spans of text into predefined semantic categories. Named Entity Recognition (NER) (Mohit, 2014) is one of the most widely studied sequence labeling tasks, traditionally addressed using feature-based statistical models such as Conditional Random Fields (CRFs) (Zheng et al., 2015). More recently, neural approaches based on recurrent networks and transformer encoders have achieved substantial improvements.

Pre-trained transformer models, such as BERT, have become the dominant architecture for sequence labeling due to their ability to capture contextual dependencies and long-range interactions. These models have been successfully applied beyond classical NER to tasks such as event extraction, argument mining, and opinion mining, typically formulated as token-level classification problems. Inspired by these approaches, conspiracy marker extraction in PsyCoMark can be naturally modeled as a sequence labeling task that identifies psycholinguistic roles within text.

2.2 Text Classification

Text classification (Kowsari et al., 2019) is another core NLP task with a long research history. Earlier approaches relied on bag-of-words representations combined with linear classifiers, while recent work predominantly adopts fine-tuned pre-trained language models. Transformer-based architectures have demonstrated strong performance in applications such as sentiment analysis, topic classification, misinformation detection, and stance classification.

Recent studies have also explored the use of multiple pre-trained models to improve robustness and performance, as different architectures and pre-training objectives may capture complementary linguistic features. For example, models such as DistilBERT (Sanh et al., 2019) and DeBERTa (He et al., 2020) introduce architectural and training improvements that enhance efficiency and representation quality. These contextualized representations are particularly important for detecting conspiracy-related content, where meaning is often implicit and depends on complex semantic relationships. Consequently, transformer-based classifiers provide a strong baseline for conspiracy detection tasks.

Our work builds on these advances by applying transformer-based sequence labeling and classification models to psycholinguistic conspiracy analysis in the PsyCoMark shared task.

3 Methodology

3.1 Task Formulation

We adopt a unified transformer-based framework to address both subtasks of PsyCoMark.

For Subtask 1, we formulate the problem as a token-level sequence labeling task. Given an input sequence, the model predicts a label for each token indicating whether it belongs to one of the five psycholinguistic marker categories: Actor, Action, Effect, Victim, or Evidence. The predicted token labels are post-processed to reconstruct character-level spans according to the official evaluation format.

For Subtask 2, we treat the task as a binary text classification problem. Given a Reddit comment, the model predicts whether the text expresses a conspiracy belief.

3.2 Model Architecture

We employ pre-trained transformer encoders as backbone models for both subtasks due to their strong contextual representation ability.

For Subtask 1, we fine-tune multilingual transformer models, including distilbert-base-multilingual-cased and lxyuan/distilbert-base-multilingual-cased-sentiments-student, with a token-level classification head for sequence labeling.

For Subtask 2, we fine-tune several transformer encoders independently, including DistilBERT and DeBERTa-v3, each followed by a classification layer. The models are trained separately without parameter sharing, and their predictions are used to analyze performance across different architectures.

3.3 Implementation Details

3.3.1 Subtask 1: Conspiracy Marker Extraction

We formulate Subtask 1 as a token-level sequence labeling task and implement a pipeline consisting of data preprocessing, model fine-tuning, and span reconstruction. The original annotations are provided as character-level spans corresponding to five

psycholinguistic conspiracy markers, including Actor, Action, Effect, Victim, and Evidence. These annotations are first converted into token-level labels using a BIO tagging scheme(Marquez et al., 2005), where each token is assigned a label indicating whether it belongs to a specific marker category or to the non-entity class.

During preprocessing, input texts are tokenized using the tokenizer associated with each pre-trained model, and character offsets are aligned with token boundaries. We fine-tune two multilingual transformer encoders, distilbert-base-multilingual-cased and lxyuan/distilbert-base-multilingual-cased-sentiments-student, for token classification. A linear classification layer is added on top of each encoder to predict token labels, and the models are optimized using cross-entropy loss.

After prediction, token-level outputs are converted back into character-level spans by merging consecutive tokens with the same predicted label. The reconstructed spans are then mapped to their corresponding marker categories and aggregated for each document. Finally, the extracted spans are formatted according to the official task specification and stored in JSONL format for evaluation.

3.3.2 Subtask 2: Conspiracy Detection

Subtask 2 is formulated as a binary text classification problem. Each Reddit comment is treated as an input sequence and assigned a label indicating whether it expresses conspiracy-related content. The input texts are tokenized using model-specific tokenizers and truncated or padded to a fixed maximum sequence length.

We fine-tune three pre-trained transformer encoders independently on the annotated training data, including distilbert-base-multilingual-cased, lxyuan/distilbert-base-multilingual-cased-sentiments-student, and microsoft/deberta-v3-base. A classification layer is added on top of each encoder to predict binary labels, and the models are optimized using cross-entropy loss. The models are trained separately without parameter sharing, allowing different architectures to capture diverse contextual representations.

During inference, each fine-tuned model produces predictions for the input text, and the final outputs are generated according to the official evaluation format. This framework provides a simple and reproducible approach for conspiracy detection while maintaining strong performance across different model architectures.

In addition, Table 6 summarizes the datasets used in our experiments. We rely exclusively on the official PsyCoMark dataset provided by the shared task organizers and do not incorporate any external corpora or additional training resources. This design ensures a fair comparison with other participating systems and highlights the effectiveness of simple fine-tuning strategies.

4 Results and Analysis

4.1 Training Dataset Analysis

Table 1 presents the statistics of text length in the training set, while Table 2 shows the label distribution for Subtask 2. Instances labeled as "Can't tell" are excluded from the evaluation.

Figure 1 further illustrates the distribution of text lengths across different classes. We observe that the length distributions of conspiracy and non-conspiracy texts largely overlap, suggesting that text length alone is not a strong discriminative feature. This indicates that effective classification requires models to capture deeper semantic and contextual information rather than relying on superficial features.

Table 1: Text length statistics for Subtask 2 training data.

Metric	Value
Count	3531
Mean	80.06
Std	39.88
Min	24
25%	48
50% (Median)	69
75%	104
Max	276

Table 2: Label distribution for Subtask 2 training data.

Label	Count
Yes	1541
No	1990
Can't tell	785

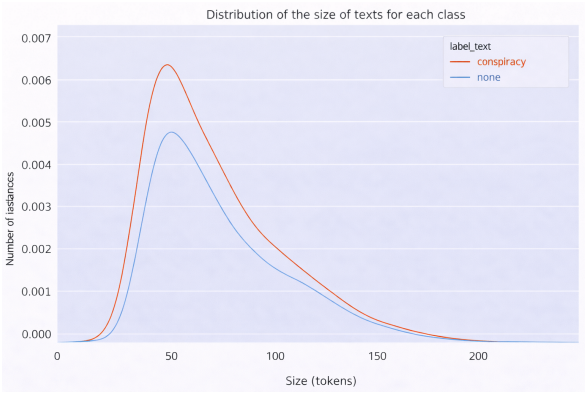

Figure 1: Distribution of the size of texts for each class.

4.2 Experimentation Configuration

To ensure comprehensive evaluation and improve model performance, we experiment with several pre-trained transformer models, including DistilBERT and DeBERTa variants. For each model, we explore two different hyperparameter settings, as shown in Table 3.

Table 3: Experimentation configuration hyperparameters for the four different transformer pre-train models.

model	Hyperparameter	Values
distilbert-base-multilingual-cased	|batch_size|	16
	|learning_rate|	2e-5
lxyuan/distilbert-base-multilingual-cased-sentiments-student	|num_epochs|	10
	|weight_decay|	0.01
microsoft/deberta-v3-base	|num_train_epochs|	10
	|learning_rate|	2e-5
	|per_device_train_batch_size|	32
	|warmup_steps|	100
	|weight_decay|	0.01
	|logging_steps|	100
	|save_total_limit|	1

4.3 Development Dataset Result

Table 4 presents the official results of SemEval-2026 Task 10: Psycholinguistic Conspiracy Marker Extraction and Detection for both Subtask 1 (Conspiracy Marker Extraction) and Subtask 2 (Conspiracy Detection) on the English development set. The best-performing system for each evaluation metric is highlighted in bold.

4.4 Test Dataset Result

Table 5 presents the official results of SemEval-2026 Task 10: Psycholinguistic Conspiracy Marker Extraction and Detection for both Subtask 1 (Conspiracy Marker Extraction) and Subtask 2 (Conspiracy Detection) on the English test set. The best-performing system for each evaluation metric is highlighted in bold.

4.5 Data Distribution Analysis

We conduct a preliminary analysis of the training data to better understand its characteristics. As shown in Table 1, approximately 75% of the texts contain fewer than 100 words, which informs the choice of maximum sequence length and model configuration.

For Subtask 1, the task can be formulated as a sequence labeling problem (Nguyen and Guo, 2007) that aims to identify spans corresponding to five psycholinguistic conspiracy markers, namely Actor, Action, Effect, Victim, and Evidence. Each annotation includes a start index, end index, marker type, and text span. The distribution of marker types in the training set is imbalanced (see Table 6). Actor markers are the most frequent (6,416 instances), followed by Action (4,841 instances), while Victim markers are the least frequent (3,315 instances). The frequencies of Effect and Evidence are relatively similar. Figure 2 provides a visual representation of this distribution, highlighting the imbalance among different marker categories. This imbalance may contribute to the lower performance of models on less frequent categories.

For Subtask 2, the task is formulated as a binary text classification problem (Kumari and Srivastava, 2017). Although the dataset originally includes three labels (Yes, No, and Can't tell), instances labeled as Can't tell are excluded. The label distribution in the English training set is relatively balanced, with approximately 43.6% conspiracy-related comments and 56.4% non-conspiracy comments.

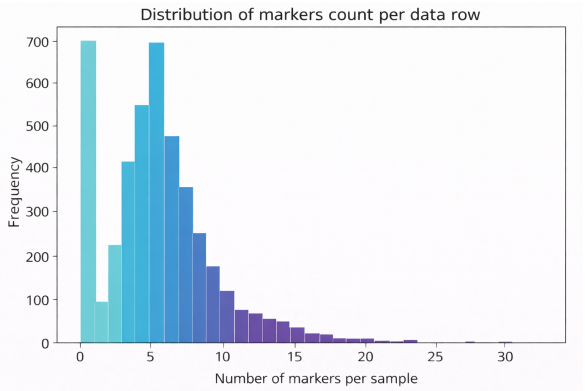

Figure 2: The markers quantity distribution of training markers data are analyzed.

Table 4: The development dataset experiment situation detailed results are described.

Subtask	Language	Main Technologies	F1-Score (Agg/Micro)	F1-Score (Macro)
1	English	**distilbert-base-multilingual-cased**	**0.1913**	**0.1762**
1	English	lxyuan/distilbert-base-multilingual-cased-sentiments-student	0.1748	0.1602
Subtask	**Language**	**Main Technologies**	**Weighted F1 Score**	**Accuracy**
2	English	distilbert-base-multilingual-cased	0.7000	0.7013
2	English	lxyuan/distilbert-base-multilingual-cased-sentiments-student	0.6766	0.6753
2	English	**microsoft/deberta-v3-base**	**0.7582**	**0.7662**

Table 5: The test dataset experiment situation detailed results are described.

Subtask	Language	Main Technologies	F1-Score(Agg/Micro)	F1-Score (Macro)
1	English	distilbert-base-multilingual-cased	0.1585	0.1462
1	English	**lxyuan/distilbert-base-multilingual-cased-sentiments-student**	**0.1595**	**0.1476**
Subtask	**Language**	**Main Technologies**	**Weighted F1 Score**	**Accuracy**
2	English	distilbert-base-multilingual-cased	0.6800	0.6822
2	English	lxyuan/distilbert-base-multilingual-cased-sentiments-student	0.6380	0.6408
2	English	**microsoft/deberta-v3-base**	**0.7267**	**0.7313**

Table 6: The overall quantity statistics of each type in markers for Subtask 1: Conspiracy Marker Extraction training set are described.

Markers Type	Count
Actor	6416
Action	4841
Effect	3739
Evidence	3654
Victim	3315

5 Conclusion

Our system adopts a fine-tuning framework based on pre-trained transformer models to address both psycholinguistic conspiracy marker extraction and conspiracy detection. We evaluate several transformer encoders, including distilbert-base-uncased, distilbert-base-multilingual-cased, lxyuan/distilbert-base-multilingual-cased-sentiments-student, and microsoft/deberta-v3-base, using consistent hyperparameter settings. The models are trained solely on the official PsyCoMark dataset without external corpora or data augmentation.

Experimental results show that simple fine-tuning of pre-trained language models provides a strong baseline for both subtasks. For Subtask 1, the sequence labeling framework effectively captures psycholinguistic conspiracy markers, while for Subtask 2, transformer-based classifiers achieve competitive performance in detecting conspiracy-related content. These findings demonstrate the effectiveness of contextualized representations for modeling psycholinguistic structures in online discourse.

6 Limitations and Future Work

Despite achieving competitive performance, our approach has several limitations. First, our system relies solely on fine-tuning pre-trained transformer models without incorporating task-specific architectures or external knowledge resources. While this design ensures simplicity and reproducibility, it may limit the model's ability to capture complex psycholinguistic structures and implicit conspiracy cues, particularly for the span extraction task in Subtask 1.

Second, the training data exhibits label imbalance across different psycholinguistic marker categories, which may negatively affect the extraction performance for low-frequency markers such as Victim and Evidence. More advanced strategies, such as data augmentation (Shorten et al., 2021) or class-balanced learning (Chen et al., 2021), could potentially improve performance.

Finally, our experiments are conducted only on English data, and the generalizability of the proposed approach to other languages remains unexplored. Future work may investigate cross-lingual transfer methods (Lample and Conneau, 2019), multilingual modeling, and joint learning strategies that leverage the interaction between marker extraction and conspiracy detection.

Acknowledgments

We are very grateful for the assistance and discussions provided by SemEval-2026 Task 10: Psycholinguistic Conspiracy Marker Extraction and Detection leaders and organizers.

Table 7: Use dataset supported by Semeval-2026 Task 10: Psycholinguistic Conspiracy Marker Extraction and Detection, on Subtask 1: Conspiracy Marker Extraction and Subtask 2: Conspiracy Detection. The style is based on raw data.

Dataset Input	**Description**	**Use or Not**
PsyCoMark official dataset	datasets for English language	yes
other dataset	use external or additional corpora	no

References

Adnan Akhundov, Dietrich Trautmann, and Georg Groh. 2018. Sequence labeling: A practical approach. *arXiv preprint arXiv:1808.03926*.

Zhi Chen, Jiang Duan, Li Kang, and Guoping Qiu. 2021. Class-imbalanced deep learning via a class-balanced ensemble. *IEEE transactions on neural networks and learning systems*, 33(10):5626–5640.

Karen M Douglas, Robbie M Sutton, and Aleksandra Cichocka. 2017. The psychology of conspiracy theories. *Current directions in psychological science*, 26(6):538–542.

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 Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. 2019. Text classification algorithms: A survey. *Information*, 10(4):150.

Roshan Kumari and Saurabh Kr Srivastava. 2017. Machine learning: A review on binary classification. *International Journal of Computer Applications*, 160(7).

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Lluis Marquez, Pere Comas, Jesús Giménez, and Neus Catala. 2005. Semantic role labeling as sequential tagging. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 193–196.

Behrang Mohit. 2014. Named entity recognition. In *Natural language processing of semitic languages*, pages 221–245. Springer.

Nam Nguyen and Yunsong Guo. 2007. Comparisons of sequence labeling algorithms and extensions. In *Proceedings of the 24th international conference on Machine learning*, pages 681–688.

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*.

Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. 2021. Text data augmentation for deep learning. *Journal of big Data*, 8(1):101.

Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. 2015. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1529–1537.