UMUTeam at SemEval-2026 Task 10: Transformer Ensembles for Conspiratorial Span Extraction and Detection

Jorge Gómez-Navalón, Ronghao Pan, Tomás Bernal-Beltrán, José Antonio García-Díaz, Rafael Valencia-García
Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100 Murcia, Spain
{jorge.gomezn,ronghao.pan,tomas.bernalb,joseantonio.garcia8,valencia}@um.es

Abstract

Conspiracy theories pose significant societal risks and require reliable automated detection methods. In this paper, we present our system for SemEval 2026 Task 10, addressing both conspiracy detection and psycholinguistic marker extraction. We leverage multiple pretrained transformer architectures and ensemble strategies to model conspiratorial discourse at both document and token levels. For classification, our ensemble achieves a weighted F1-score of 0.7688, indicating effective performance in distinguishing conspiratorial statements. For marker extraction, we formulate the task as a BIOES sequence labeling problem and enhance predictions through ensemble and specialist models. Our results highlight both the effectiveness of transformer-based approaches and the challenges of fine-grained conspiracy marker extraction.

1 Introduction

New technologies have changed not only the way we communicate but also the way we obtain information, becoming the main source of information. This, added to the ease with which social networks allow any type of message to be posted, has led to increased social concern about detecting hate speech. Among all types, there is one that arises in response to uncertainty and attempts to undermine authorities and knowledge on the subject: conspiracy messages (Grusauskaite et al., 2022). To this end, they reinforce the theory by citing selective information, elaborating on details, raising doubts, and shifting the burden of proof (Kou et al., 2017). Recently, we have been confronted with major conspiracy theories, such as those that emerged during COVID-19, which undermined institutional trust and support for regulations (Pummerer et al., 2020). Or the ability to reduce confidence in scientific solutions, such as those related to climate change (Bolsen et al., 2022).

This shows us the great power that conspiracy theories have over society. That is why different strategies have been applied to reduce their impact, some of which have been through education (O'Mahony et al., 2023) or inoculation of small, easily refuted conspiracies (Compton et al., 2021). Despite these efforts, most interventions were ineffective in significantly changing beliefs in only 40% of cases and without experiments demonstrating that these changes endure (O'Mahony et al., 2023).

However, Artificial Intelligence (AI) has emerged as a promising approach for addressing harmful online content, particularly with the rise of Large Language Models (LLMs). These models have demonstrated strong performance across various natural language processing tasks, such as hate speech detection (García-Díaz et al., 2023) and conspiracy classification (Moffitt et al., 2021). Nevertheless, their application to conspiratorial discourse presents unique challenges that remain insufficiently explored.

The existing ambiguity in belonging to a conspiracy is one of the major problems, where models often fail to classify texts that are critiques or sarcasms (Diab et al., 2024). Interpretability also presents a great challenge, as models are often black boxes that provide only a single output (Papageorgiou et al., 2024a).

The PsyCoMark shared task (SemEval 2026) combines psychology with NLP to study how conspiracy theories are expressed in conversations (Samory et al., 2026). The task requires not only document-level classification but also fine-grained extraction of psycholinguistic markers. It is divided into two subtasks: (1) **Subtask 1: Conspiracy marker extraction**. In each sentence, you have to identify, if they exist, 5 different types of psycholinguistic markers: Actor, Action, Victim, Threat, and Evidence; and (2) **Subtask 2: Conspiracy detection (classification)**. Detect if a phrase is

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 483–489

conspiratorial or not.

To address these complementary objectives, we propose a unified framework that combines document-level detection with span-level extraction. For Subtask 1 (extraction), we formulate marker identification as a token-level sequence labeling problem using the BIOES tagging scheme, where B, I, E, and S denote the beginning, inside, end, and single-token spans, respectively, and O denotes tokens outside any marker span, enabling precise span reconstruction. To enhance consistency and reduce boundary noise, we incorporate a token-set voting ensemble strategy and explore specialist models for high-variance marker types. For Subtask 2 (classification), we fine-tune multiple pretrained transformer models using stratified cross-validation and ensemble strategies to improve robustness and threshold calibration.

By jointly modeling global detection and local discourse structure, our approach aims to improve both predictive reliability and interpretability. This dual perspective allows us not only to determine whether a text is conspiratorial, but also to identify how conspiratorial reasoning is linguistically constructed.

2 Background

Automatic detection of conspiratorial discourse has gained increasing attention in recent years (Papageorgiou et al., 2024b), particularly in the context of misinformation and online radicalization. Its significant impact on society, as evidenced during the COVID-19 pandemic, has motivated the development of diverse computational approaches to mitigate the spread of conspiracy narratives.

Early studies explored traditional machine learning methods, such as random forest classifiers applied to COVID-19-related tweets, demonstrating competitive performance in supervised classification settings (Gerts et al., 2021). With the rise of transformer-based architectures, domain-adapted models such as COVID-BERT were proposed to improve contextual modeling in pandemic-related misinformation detection (Peskine et al., 2021). More recently, large language models (LLMs) have been investigated for conspiracy detection, including approaches that incorporate emotional signals to enhance discriminative power (Liu et al., 2024). Subsequent improvements have focused on robustness against stylistic variation by generating paraphrased versions of input sentences (Liu et al., 2025).

While most prior work formulates conspiracy detection as a document-level classification task, research on fine-grained span-level modeling remains limited. Span extraction techniques, commonly used in tasks such as named entity recognition (Yu et al., 2022) and argument mining (Kawarada et al., 2024), offer a promising alternative for capturing the internal structure of discourse. Methods based on structured term extraction and interaction modeling have demonstrated the feasibility of identifying semantically relevant components within text (Xu et al., 2021). Additionally, BIO-style tagging schemes have consistently demonstrated strong empirical performance in structured sequence labeling tasks (Zeng et al., 2024).

In contrast to generic span extraction tasks, the PsyCoMark framework is grounded in psychological theories of conspiratorial reasoning (Baele, 2019), which conceptualize conspiracy narratives as being constructed through recurrent discourse components. Specifically, the task defines five psycholinguistic marker types: Actor, Action, Victim, Threat, and Evidence, which represent structural elements of conspiratorial narratives. Modeling these markers computationally enables a structured and interpretable representation of conspiratorial reasoning, effectively bridging psychological theory and sequence-based NLP modeling.

3 System overview

We build our system using multiple pretrained transformer architectures fine-tuned for both subtasks, specifically DeBERTa-v3-large (He et al., 2021), RoBERTa-large (Liu et al., 2019), EuroBERT-610M (Boizard et al., 2025), and ModernBERT-large (Warner et al., 2024). These models allow us to capture contextual representations at both token and document levels, which are necessary for marker extraction and conspiracy detection. We use these architectures as independent components in an ensemble framework for both subtasks.

Figure 1 illustrates the overall architecture of our system for the conspiracy marker extraction subtask. Given an input sentence, we identify spans corresponding to five psycholinguistic markers: Actor, Action, Victim, Threat, and Evidence. We formulate this task as a token-level sequence labeling problem.

We first tokenize the input text using the tok-

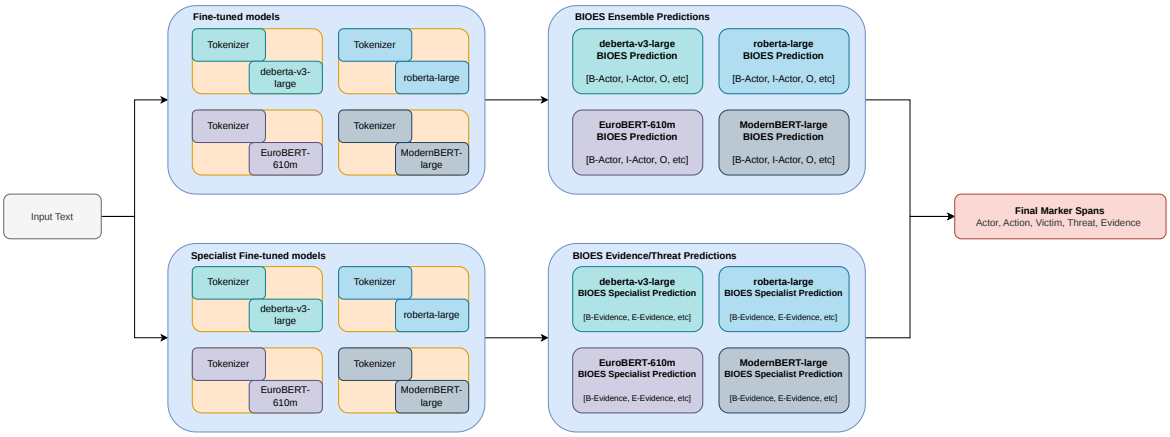

Figure 1: System architecture for Subtask 1.

enizer associated with each transformer model. We then feed the tokenized sequence into each fine-tuned encoder, which predicts BIOES labels for every token. These labels indicate whether a token belongs to a marker span and specify its corresponding marker type.

After obtaining token-level predictions, we reconstruct character-level spans by grouping contiguous labeled tokens. Since different models may produce slightly different span boundaries, we train multiple models using cross-validation and collect predictions from each fold. We aggregate span candidates using token-set voting across models and folds. Each predicted span is represented as a set of character-aligned tokens, and spans are retained only when they are supported by at least 3 predictions. When multiple overlapping candidates are available, we keep the span with the highest support.

To further refine span extraction, we incorporate specialist models based on the same transformer architectures as the general ensemble. These models are trained to focus on the most challenging marker types, namely Evidence and Threat, which exhibit higher semantic variability and less stable span boundaries.

The specialist models generate additional span candidates, which are compared against those produced by the general ensemble. Only spans with sufficient overlap are retained, improving precision and reducing false positives. This targeted modeling approach allows the system to better handle marker types that showed lower performance during development.

We compare these predictions with the ensemble output and retain only spans with sufficient overlap, improving precision and reducing false positives.

Figure 2 illustrates the architecture of our system for the conspiracy detection subtask. Given an input sentence, we predict whether the input contains conspiratorial discourse. We formulate this task as a binary classification problem.

We tokenize each input sentence and process it using the same set of transformer models adapted for sequence classification. Each model produces a probability indicating whether the input contains conspiratorial content. We train these models using cross-validation and collect predictions from each fold.

Final predictions are obtained through a weighted average of model probabilities. Weights were manually tuned on the development set by maximizing weighted F1-score, assigning higher importance to models with stronger individual performance. This approach leverages the complementary strengths of different transformer architectures and improves prediction robustness.

4 Experimental setup

We use the PsyCoMark dataset provided by the SemEval 2026 Task 10 organizers. Each instance corresponds to a Reddit submission statement and includes a conspiracy label (Yes, No, or Can't tell) and, when applicable, character-level annotations corresponding to five marker types: Actor, Action, Victim, Threat, and Evidence.

For Subtask 1 (marker extraction), we use all available instances, since marker annotations may appear regardless of the conspiracy label. For Subtask 2 (conspiracy detection), we formulate the task as a binary classification problem. Therefore, we remove all instances labeled as Can't tell, as these

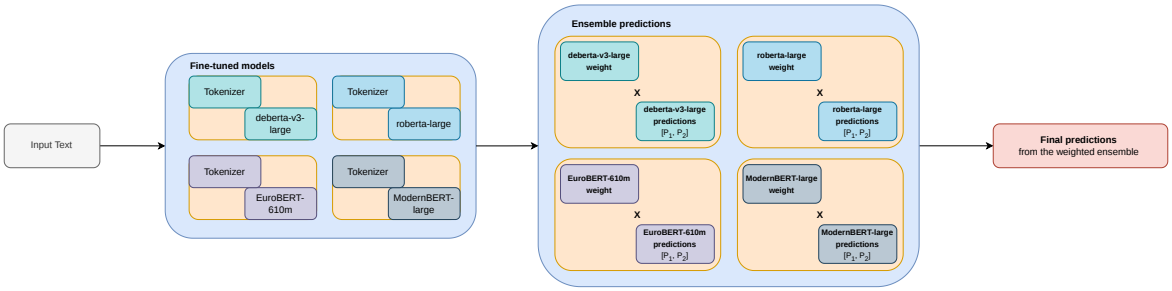

Figure 2: System architecture for Subtask 2.

Table 1: Dataset distribution with number of examples and the percentage they represent

Dataset	Total	Yes	No	Can't tell
Training	4316	1541 (35.7%)	1990 (46.1%)	785 (18.2%)

Table 2: Subtask 2 results on the official test set.

Metric	Score
Accuracy	0.7700
F1-score (weighted)	0.7688
F1-score (No)	0.7949
F1-score (Yes)	0.7382

do not provide a definitive ground-truth label. This preprocessing step reduces the effective training dataset by 18.2%, as shown in Table 1.

Both subtasks use the same four pretrained transformer architectures: DeBERTa-v3-large, RoBERTa-large, EuroBERT-610M, and ModernBERT-large. We fine-tune each model separately for token classification in Subtask 1 and sequence classification in Subtask 2.

For both subtasks, we trained each architecture using cross-validation and aggregated predictions across folds at inference time. The final system therefore combines diversity across both model architectures and training splits.

We train all models for 6 epochs with a batch size of 8 and a maximum sequence length of 512 tokens. For DeBERTa, we use a learning rate of 9.93e-06, a weight decay of 0.05 and a warmup ratio of 0.06. RoBERTa has a learning rate of 8e-06, weight decay of 0.01 and a warmup ratio of 0.08. EuroBERT has a fixed learning rate of 8e-06, a weight decay of 0.05 and a warmup ratio of 0.08. Finally, ModernBERT has a learning rate of 8e-06, a weight decay of 0.01 and a warmup ratio of 0.08.

5 Results

In this section, we present and analyze the results obtained by our system on the official PsyCoMark test set. During development, we evaluated multiple model configurations and ensemble strategies, and the results reported below correspond to our final submission.

5.1 Subtask 2: Conspiracy Detection

Table 2 shows the performance of our system on the conspiracy detection subtask. Our approach achieved a weighted F1-score of 0.7688 and an overall accuracy of 0.7700, indicating effective performance in distinguishing conspiratorial from non-conspiratorial statements.

The higher F1-score for the *No* class indicates that the system more reliably identified non-conspiratorial statements. In contrast, conspiratorial statements proved more challenging, likely due to their linguistic diversity, indirect framing, and implicit rhetorical strategies. These results suggest that transformer-based models can effectively capture global semantic patterns associated with conspiratorial discourse.

The ensemble strategy was intended to improve robustness by combining complementary representations from different architectures, as combining multiple pretrained architectures allowed the system to leverage complementary contextual representations and reduce model-specific biases. Cross-validation further improved stability by allowing the ensemble to capture diverse decision boundaries across folds.

5.2 Subtask 1: Marker Extraction

Table 3 presents the token-level performance of our system for conspiracy marker extraction. Our approach achieved an aggregate token-level F1-score of 0.2551 and a macro F1-score of 0.2411.

Performance varied across marker types. The

Table 3: Subtask 1 token-level results on the official test set.

Metric	Score
F1 Aggregate	0.2551
Precision Aggregate	0.2584
Recall Aggregate	0.2519
Macro F1	0.2411
F1 Actor	0.3991
F1 Victim	0.2654
F1 Threat	0.1891
F1 Action	0.1783
F1 Evidence	0.1738

Table 4: Subtask 2 results comparison.

System	Acc.	Weighted F1
DeBERTa-v3-large	0.7403	0.7273
Ensemble	0.7700	0.7688

Table 5: Subtask 1 results comparison.

System	Macro F1	Aggregate F1
DeBERTa-v3-large	0.1053	0.1542
Ensemble	0.1636	0.1311
Ensemble + Specialist	0.2411	0.2551

system achieved the highest performance for the *Actor* marker, suggesting that entities involved in conspiratorial narratives are more explicitly expressed and easier to detect. In contrast, markers such as *Action* and *Evidence* obtained lower scores, likely due to their broader semantic variability and less clearly defined boundaries.

Overall, the relatively lower performance compared to classification highlights the increased complexity of fine-grained span extraction. This task requires precise boundary detection and semantic interpretation at the token level. Nevertheless, the use of BIOES tagging combined with transformer-based encoders and ensemble strategies enabled the system to identify meaningful components of conspiratorial discourse.

In particular, the BIOES tagging scheme facilitated precise span boundary modeling, while specialist models helped improve consistency for markers with higher semantic variability.

Our system ranked 5th out of 30 teams in Subtask 1 and 13th out of 52 teams in Subtask 2.

5.3 Ablation and Error Analysis

A brief manual inspection of the predictions revealed that most errors in Subtask 1 correspond to boundary mismatches, where the model captures only part of a multi-token span and confusion between semantically similar markers such as Action and Evidence. In many cases, the model correctly identified the presence of a marker but failed to capture its full span. This suggests that the model struggles to consistently identify complete span boundaries when markers are expressed through longer or syntactically complex constructions.

For Subtask 2, the system struggled particularly with implicit or sarcastic statements, where conspiratorial intent is expressed indirectly or through rhetorical questions. These cases often led to misclassification, highlighting the difficulty of modeling subtle discourse cues.

As shown in Table 4, the ensemble clearly outperforms the best single model, confirming the benefit of combining multiple transformer architectures through ensemble strategies.

Table 5 shows that the ensemble significantly improves over the best single model in Subtask 1. Furthermore, incorporating specialist models leads to additional gains, particularly in terms of aggregate F1, confirming their effectiveness for handling more challenging marker types. These improvements are consistent with the observed error patterns, as specialist models help reduce boundary inconsistencies and confusion between semantically overlapping markers.

Overall, these results highlight the importance of combining ensemble strategies with targeted modeling approaches to effectively handle both global classification and fine-grained span extraction.

6 Conclusion

In this paper, we presented our system for SemEval 2026 Task 10, addressing both conspiracy detection and psycholinguistic marker extraction. We leveraged multiple pretrained transformer architectures and ensemble strategies to model conspiratorial discourse at both document and token levels. For conspiracy detection, our approach achieved competitive performance by combining complementary representations from different transformer models. For marker extraction, we formulated the task as a sequence labeling problem using a BIOES tagging scheme and enhanced predictions through ensemble and specialist models.

Our results highlight the effectiveness of transformer-based models for detecting conspiratorial content, while also confirming the increased difficulty of fine-grained marker extraction. Model-

ing psycholinguistic markers requires precise span identification and deeper semantic understanding.

In future work, we plan to explore joint modeling approaches that integrate classification and marker extraction within a unified architecture. Additionally, incorporating linguistic and discourse-aware features may further improve the detection of implicit conspiratorial reasoning.

Acknowledgments

This work is part of the research project LaTe4PoliticES (PID2022-138099OB-I00) funded by MICIU/AEI/10.13039/501100011033 and the European Regional Development Fund (ERDF/EU - FEDER/UE)-a way of making Europe.

References

Stephane J Baele. 2019. Conspiratorial narratives in violent political actors' language. *Journal of language and social psychology*, 38(5-6):706–734.

Nicolas Boizard, Hippolyte Gisserot-Boukhlef, Duarte M. Alves, André Martins, Ayoub Hammal, Caio Corro, Céline Hudelot, Emmanuel Malherbe, Etienne Malaboeuf, Fanny Jourdan, Gabriel Hautreux, João Alves, Kevin El-Haddad, Manuel Faysse, Maxime Peyrard, Nuno M. Guerreiro, Patrick Fernandes, Ricardo Rei, and Pierre Colombo. 2025. Eurobert: Scaling multilingual encoders for european languages.

Toby Bolsen, Risa Palm, and Justin T Kingsland. 2022. Effects of conspiracy rhetoric on views about the consequences of climate change and support for direct carbon capture. *Environmental Communication*, 16(2):209–224.

J. Compton, S. Linden, J. Cook, and Melisa Basol. 2021. Inoculation theory in the post-truth era: Extant findings and new frontiers for contested science, misinformation, and conspiracy theories. *Social and Personality Psychology Compass*, 15.

Ahmad Diab, Rr Nefriana, and Yu-Ru Lin. 2024. Classifying conspiratorial narratives at scale: False alarms and erroneous connections. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 340–353.

José Antonio García-Díaz, Ronghao Pan, and Rafael Valencia-García. 2023. Leveraging zero and few-shot learning for enhanced model generality in hate speech detection in spanish and english. *Mathematics*, 11(24):5004.

Dax Gerts, Courtney D Shelley, Nidhi Parikh, Travis Pitts, Chrysm Watson Ross, Geoffrey Fairchild, Nidia Yadria Vaquera Chavez, and Ashlynn R Daughton. 2021. "thought i'd share first" and other conspiracy theory tweets from the covid-19 infodemic: Exploratory study. *JMIR public health and surveillance*, 7(4):e26527.

Kamile Grusauskaite, Jaron Harambam, and Stef Aupers. 2022. Picturing opaque power: How conspiracy theorists construct oppositional videos on youtube. *Social Media+ Society*, 8(2):20563051221089568.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Masayuki Kawarada, T. Hirao, Wataru Uchida, and Masaaki Nagata. 2024. Argument mining as a text-to-text generation task. pages 2002–2014.

Yubo Kou, Xinning Gui, Yunan Chen, and Kathleen H. Pine. 2017. Conspiracy talk on social media. *Proceedings of the ACM on Human-Computer Interaction*, 1:1 – 21.

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*.

Zhiwei Liu, Boyang Liu, Paul Thompson, Kailai Yang, and Sophia Ananiadou. 2024. Conspemollm: Conspiracy theory detection using an emotion-based large language model. *arXiv preprint arXiv:2403.06765*.

Zhiwei Liu, Paul Thompson, Jiaqi Rong, and Sophia Ananiadou. 2025. Conspemollm-v2: A robust and stable model to detect sentiment-transformed conspiracy theories. *arXiv preprint arXiv:2505.14917*.

J. D. Moffitt, Catherine King, and K. Carley. 2021. Hunting conspiracy theories during the covid-19 pandemic. *Social Media + Society*, 7.

Cian O'Mahony, M. Brassil, G. Murphy, and Conor Linehan. 2023. The efficacy of interventions in reducing belief in conspiracy theories: A systematic review. *PLOS ONE*, 18.

Eleftheria Papageorgiou, Christos Chronis, Iraklis Varlamis, and Yassine Himeur. 2024a. A survey on the use of large language models (llms) in fake news. *Future Internet*, 16:298.

Eleftheria Papageorgiou, Christos Chronis, Iraklis Varlamis, and Yassine Himeur. 2024b. A survey on the use of large language models (llms) in fake news. *Future Internet*, 16(8):298.

Youri Peskine, Giulio Alfarano, Ismail Harrando, Paolo Papotti, and Raphael Troncy. 2021. Detecting covid-19-related conspiracy theories in tweets. In *MediaEval*.

Lotte Pummerer, Robert Böhm, Lau Lilleholt, Kevin Winter, Ingo Zettler, and K. Sassenberg. 2020. Conspiracy theories and their societal effects during the covid-19 pandemic. *Social Psychological and Personality Science*, 13:49 – 59.

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)*.

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.

Lu Xu, Yew Ken Chia, and Lidong Bing. 2021. Learning span-level interactions for aspect sentiment triplet extraction. *ArXiv*, abs/2107.12214.

Jie Yu, Bin Ji, Shasha Li, Jun Ma, Huijun Liu, and Hao Xu. 2022. S-ner: A concise and efficient span-based model for named entity recognition. *Sensors (Basel, Switzerland)*, 22.

Zhengqiao Zeng, Zhongyuan Han, Jingyan Ye, Yaozu Tan, Haojie Cao, Zengyao Li, and Runjin Huang. 2024. A conspiracy theory text detection method based on roberta and xlm-roberta models. pages 3002–3006.