CSECU-DSG at SemEval-2026 Task 10: Fine-Tuning DeBERTa Transformer Model for Conspiracy Detection

Debashish Chakraborty, Sumaiya Tabassum, Sabrina Ibnath, and Abu Nowshed Chy
Department of Computer Science and Engineering
University of Chittagong, Chattogram-4331, Bangladesh
{debashish.csecu, sumaiya.tabassum.csecu, sabrina.ibnath.cu}@gmail.com
and nowshed@cu.ac.bd

Abstract

Conspiracy detection aims to determine whether a social media post expresses belief in conspiracy theories. This task is essential for understanding harmful online discourse and mitigating the spread of misinformation. However, detecting conspiracy beliefs is challenging due to subtle psycholinguistic cues and the strong contextual dependency of such claims. To address these challenges, SemEval-2026 Task 10 introduced a shared task named PsyCoMark. In this paper, we describe our approach to Subtask 2, which focuses on detecting conspiracy beliefs. We propose a transformer-based classification approach using a fine-tuned DeBERTa-v3-base model to detect conspiracy beliefs in Reddit comments. Each post is processed as a single input sequence. To address class imbalance and improve generalization, we employ class-weighted cross-entropy loss with label smoothing during training. Our approach achieves competitive performance, ranked ninth among participating teams. The findings demonstrate that fine-tuned transformer models effectively capture contextual and psycholinguistic patterns in conspiracy-related discourse and achieve competitive performance compared to other systems.

1 Introduction

The rapid advancement of technologies and social platforms increases the growth of user-generated content, making it easier for people to share their ideas and opinions. However, it also has significant disadvantages, including the proliferation of conspiracy theories and misinformation. A conspiracy theory is a belief that some influential or controlling organization or group is secretly responsible for a notable event or phenomenon. These kinds of narratives can shape people's opinions, create mistrust of authority and widen social inequalities. They can also dilute factual information, confuse the public, and ultimately erode trust in verified truths. Therefore, identifying and detecting conspiracy-related statements has become an important yet challenging problem in the field of natural language processing.

Sentence	Conspiracy
We discuss how different world events are secretly controlled by Illuminati.	Yes
This post simply reports the official election results without any further claims.	No

Table 1: Example of sub-task 2

In an effort to fully capture the deeper psycholinguistic structure of conspiratorial narratives, (Ghosh et al., 2026) proposed a shared task at SemEval-2026 Task 10, PsyCoMark: Psycholinguistic Conspiracy Marker Extraction and Detection. The task is divided into two subtasks. Subtask 1 is to extract psycholinguistic markers of conspiracy thinking (e.g., Actors, Victims, Effects), and sub-task 2 is to detect whether a Reddit comment expresses a conspiracy belief. We employ a transformer-based model in sub-task 2 of this task because of its strong ability to capture contextual dependencies and subtle semantic patterns, both of which are critical for detecting subtle psycholinguistic signals in conspiracy-related discourse.

2 Related Work

This task represents a progression from prior well-defined approaches to the identification and detection of conspiracies. (Rosenbusch et al., 2021) proposed an introduction to supervised machine learning methods in psychology, describing prediction and pattern detection approaches such as linear regression, ridge regression, decision trees, and random forest, along with annotated R code

*The first three authors have equal contributions.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1122–1127

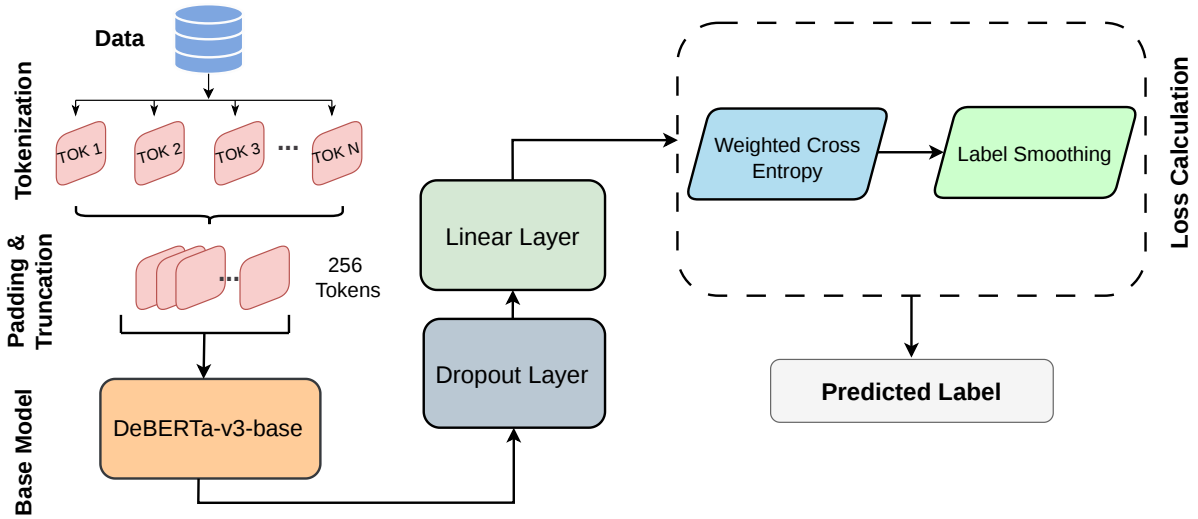

Figure 1: Overview of our proposed framework.

for model implementation, evaluation, and tuning. (Batzdorfer et al., 2022) used word embeddings to separate conspiracy-related tweets from non-conspiracy related tweets and applied time analysis that includes STL decomposition, autocorrelation, and generalized additive mixed model to study the trends and narrative patterns of COVID-19 conspiracy theories on Twitter. (Marini and Jezek, 2024) created the Complotto corpus, a 40000-token bilingual English-Italian dataset of conspiracy-oriented Telegram comments, and used the sketch engine platform to identify statistically significant lexical and semantic markets of conspiracy discourse, including evidentiality, epistemic modality, debunking vocabulary and conceptual metaphors such as *"Institutions are abusers"*.

3 Proposed Framework

In this section, we describe our proposed conspiracy detection framework as shown in Figure 1. Our goal is to classify a given rehydrated reddit post into one of three predefined conspiracy categories.

In our framework, each post is treated as a single input sequence and encoded using a transformer-based architecture. We employ the DeBERTa-v3-base (He et al., 2021, 2023) model as the backbone encoder to capture rich contextual representations of the input text. The tokenized sequence is fed into the transformer encoder, which produces contextualized embeddings for all tokens. The representation corresponding to the special classification token is then passed to a task-specific linear layer to generate prediction scores over all conspiracy categories.

3.1 Fine-tuned Transformer Model

We fine-tune the DeBERTa-v3-base transformer architecture for sequence classification as shown in Figure 2. Each rehydrated Reddit post is treated as a single input sequence and tokenized using the fast DeBERTa tokenizer provided with the model, with truncation and padding applied to a maximum sequence length of 256 tokens. The tokenized inputs are processed by DeBERTa's Disentangled Attention encoder to produce contextualized token representations. The contextual embedding of the special [CLS] token from the final encoder layer is passed through a dropout layer followed by a linear classification head to produce logits over the conspiracy label categories. During training, we apply label smoothing with a factor of 0.1 and optimize a class-weighted cross-entropy loss to mitigate the effects of class imbalance.We describe the model architecture in the subsequent sections.

3.1.1 DeBERTa

DeBERTa (He et al., 2021) refines upon BERT (Devlin et al., 2019) through the use of disentangled attention and an enhanced mask decoder. This study leveraged a meliorated DeBERTa model, known as DeBERTa-V3-base (He et al., 2023) which further optimizes the architecture through Electra-style pretraining. In DeBERTa-V3, the ELECTRA model's mask language modeling (MLM) has been substituted with a replaced token detection (RTD) approach. To ascertain if an input token is authentic or if it has has been substituted with a generator, the model is trained in the role of a discriminator. Moreover, it employs the gradient-segregated

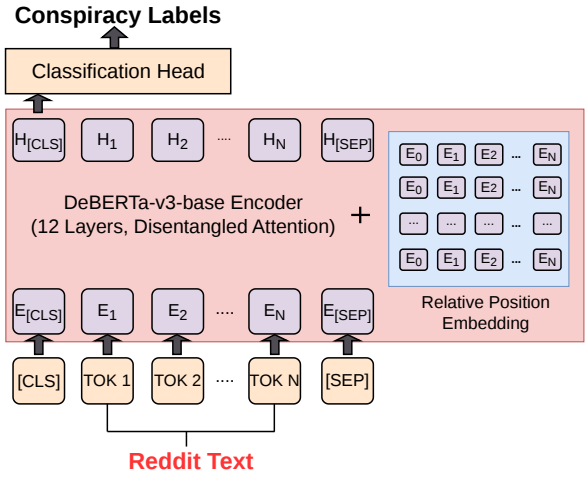

Figure 2: DeBERTa-v3-base model based classifier.

embedding sharing (GSES) approach, enabling embedding sharing between generators and discriminators. Consequently, due to this one-way sharing, the generator distributes its embeddings to the discriminator, even though the discriminator can only backpropagate the representations. With the modifications to the component mentioned above, the DeBERTa model has demonstrated considerable enhancements across numerous downstream tasks. Driven by this, we aim to retrieve the contextual representations of each input post through utilizing Huggingface's (Wolf et al., 2020) execution of the microsoft/deberta-v3-base. The DeBERTa-v3-base model comprises 12 transformer blocks, a hidden size of 768, and it has only 86M backbone parameters with a vocabulary containing 128K tokens and 98M parameters in the Embedding layer.

3.1.2 Training Strategy and Optimization

The model is fine-tuned in a supervised setting using a stratified train–validation split to preserve the label distribution. Performance is monitored on the validation set using the macro-averaged F_1 score, which is well-suited for imbalanced multi-class classification. Training is performed using the AdamW optimizer with a learning rate of 1×10^{-5} and weight decay of 0.01, along with a linear warmup over the initial training steps. Early stopping is applied to select the best-performing model and mitigate overfitting.

3.1.3 Loss Function and Class Imbalance Handling

Given the skewed distribution of conspiracy categories in the dataset, we incorporate class imbalance mitigation directly into the learning objective. Class weights are computed from the training data using inverse frequency weighting and incorporated into the Cross-Entropy loss function. This cost-sensitive learning approach penalizes misclassification of minority classes more heavily, encouraging the model to learn more discriminative representations for underrepresented conspiracy categories. In addition, label smoothing with a factor of 0.1 is applied during training to prevent overconfidence and improve generalization in multi-class classification.

4 Experiment and Evaluation

4.1 Dataset Description

Category	Count
Unique submission statements	$>$ 4,100
Total annotated entries	$\approx$ 4,800
Subreddits	$>$ 190
Conspiracy = Yes	1,715
Conspiracy = No	2,263
Conspiracy = Can't tell	877

Table 2: Statistics of the PsyCoMark dataset.

The organizers of the Psycholinguistic Conspiracy Marker Extraction and Detection shared task (Task 10 at SemEval-2026) provided an English benchmark dataset, PsyCoMark (Samory et al., 2025), to evaluate the performance of the participating systems. The dataset statistics are summarized in Table 2. The dataset comprises over 4,100 unique Reddit submission statements collected between March 2013 and December 2023 from more than 190 subreddits. Each submission statement is annotated with a three-way conspiracy label: *Yes*, *No*, or *Can't tell*. The organizers also provided a development set of 100 instances (*No: 50, Yes: 27, Can't tell: 23*) and a test set of 938 instances. The dataset was constructed by oversampling comments likely to contain conspiratorial content (e.g., from `r/conspiracy`), and therefore is not representative of the overall prevalence of conspiracy theories on Reddit.

4.2 Experimental Settings

We now describe the details of our experimental and hyper-parameter settings, along with the fine-tuning strategy that we have employed to design our proposed conspiracy detection system. We finetune a state-of-the-art Huggingface transformer

model named `microsoft/deberta-v3-base` to implement our system. We use a CUDA-enabled GPU and set a manual seed of 42 to generate reproducible results.

Parameter	Optimal Value
Max sequence length	256
Train batch size	8
Validation/test batch size	16
Number of epochs	4
Learning rate	1e-5
Warmup ratio	0.1
Weight decay	0.01
Label smoothing	0.1
Validation split	10% (stratified)
Early stopping patience	2

Table 3: Model settings for our proposed system.

We split the provided training data into training and validation sets with a 90/10 ratio using a stratified split. We address class imbalance by computing class weights with the `compute_class_weight` function and optimizing a weighted Cross-Entropy loss within a customized `Trainer`. During training, we monitor the macro-averaged F1-score on the validation set. The optimal hyper-parameter settings are summarized in Table 3. We use the default settings for the other parameters.

4.3 Results and Analysis

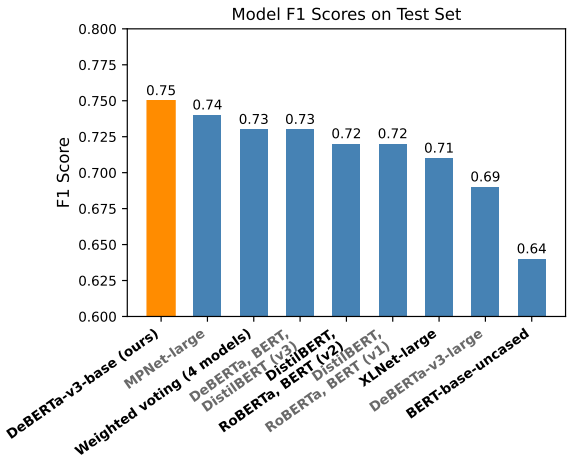

Figure 3: Performance of models on the test set.

In this section, we analyze the performance of our proposed system. The organizers of the shared task considered the macro-averaged F1-score as the primary evaluation measure. We evaluated multiple transformer-based models individually as well as using ensemble strategies. Following the competition's conclusion, the organizers updated the evaluation metric to account for missing predictions and non-committal labels across the entire test set. Under this revised official scoring, Figure 3 shows the comparative performance of all the models.

Team Name	F1 Score	Rank
CSECU-DSG (ours)	0.75	9th
Performance Comparison		
NJUST_KMG	0.89	1st
mdok-style	0.78	2nd
dangphuduy	0.78	3rd

Table 4: Comparative performance of our proposed method along with top-performing participants' methods (F1 score; Higher is better).

A majority voting ensemble of DistilBERT (Sanh et al., 2020), RoBERTa (Liu et al., 2019) (excluding the Can't tell class), and BERT-base-uncased (Koroteev, 2021) (including Can't tell) achieved an F1-score of 0.72 (V1), while including the Can't tell class for all three models slightly improved performance to 0.72 (V2). Another majority-voting ensemble over DeBERTa, BERT, and DistilBERT (V3) yielded an F1-score of 0.73, and a weighted-voting ensemble combining DistilBERT, BERT, RoBERTa, and DeBERTa achieved 0.73. Among individual models, MPNet-large (Song et al., 2020) and XLNet-large (Yang et al., 2019) obtained F1-scores of 0.74 and 0.71, respectively. Overall, DeBERTa-v3-base achieved the best performance with an F1-score of 0.75, demonstrating the robustness of the proposed system for conspiracy detection.

4.4 Comparison and Analysis of Top-Ranked Systems

Table 4 compares our system with top-performing submissions on the official test set. Our CSECU-DSG system achieved an F1-score of 0.75, ranking ninth overall. The top-ranked system, NJUST_KMG, obtained an F1-score of 0.89, while mdok-style, ranked second, scored 0.78 and dangphuduy, ranked third, also scored 0.78. These results indicate that our DeBERTa-v3-base–based approach is competitive within the shared task.

While the top-ranked systems achieved high performance through complex multi-stage pipelines, large-scale models, and external feature integration, our approach, CSECU_DSG, is distinguished by its architectural focus on optimizing a single, efficient transformer backbone for robust sequence classification. The primary differences between our system and the top three performers are summarized below:

- **Model Scale and Efficiency vs. Large-Scale LLMs:** The 1st and 2nd place systems (NJUST_KMG and mdok-style) heavily relied on massive language models, such as ChatGPT5.2, Qwen2.5, and Qwen3-32B. In contrast, our approach utilizes DeBERTa-v3-base, a significantly more compact model with 86M backbone parameters[cite: 4]. Despite its smaller size, our system captures essential contextual and psycholinguistic patterns through DeBERTa's disentangled attention mechanism and replaced token detection (RTD) pre-training, offering a more computationally efficient solution for conspiracy detection.

- **Supervised Fine-Tuning vs. Semi-Supervised and Two-Stage Pipelines:** Our framework employs a direct, supervised fine-tuning strategy on the provided dataset. This differs significantly from the 1st place system (NJUST_KMG), which used a two-stage approach involving large-model filtering to remove unrelated samples, followed by retrieval-enhanced custom prompting. It also contrasts with the 2nd place system (mdok-style), which utilized self-training, an iterative semi-supervised technique where the model acts as its own teacher to label unlabeled development and test data.

- **Focus on Label Imbalance vs. Multi-Modal and Emotion Fusion:** To handle the skewed distribution of conspiracy categories, we integrated class-weighted cross-entropy loss with label smoothing directly into our learning objective. While the 3rd place system (dangphuduy) also addressed the task through fine-tuning, they distinguished their approach by incorporating emotion-aware representations via a dynamic gating mechanism that fused semantic data with scores from a fixed emotion feature extractor. Our approach remains grounded in purely semantic and contextual features, demonstrating that refined optimization of these signals can still yield highly competitive results.

- **Data Strategy:** Our system focuses on maximizing the utility of the original data through stratified splits and early stopping to mitigate overfitting. This is distinct from the top systems that relied on extensive data augmentation, such as the retrieval and reranking used by NJUST_KMG or the obfuscation and anonymization techniques adapted from machine-generated text detection by mdok-style.

While the top-ranked systems leveraged the scale of LLMs and auxiliary features like emotional patterns or retrieval-augmented contexts, CSECU_DSG demonstrates that a well-optimized, base-sized transformer model, when combined with strategic loss functions and label smoothing, can achieve a rank in the top percentile with greater architectural simplicity.

5 Conclusion and Future Directions

In this study, we proposed a transformer-based framework for detecting conspiracy-related content in social media posts, leveraging the DeBERTa-v3-base model to capture rich contextual representations. For future work, we aim to explore data augmentation, long-document modeling, multi-task learning, and hybrid systems combining transformer-based classifiers with large language models to better capture ambiguity and contextual reasoning in conspiracy-related discourse.

References

Veronika Batzdorfer, Holger Steinmetz, Marco Biella, and Meysam Alizadeh. 2022. Conspiracy theories on twitter: emerging motifs and temporal dynamics during the covid-19 pandemic. *International Journal of Data Science and Analytics*, 13.

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.

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, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

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

M. V. Koroteev. 2021. Bert: A review of applications in natural language processing and understanding. *Preprint*, arXiv:2103.11943.

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.

Costanza Marini and Elisabetta Jezek. 2024. What to annotate: Retrieving lexical markers of conspiracy discourse from an Italian-English corpus of telegram data. In *Proceedings of the 20th Joint ACL - ISO Workshop on Interoperable Semantic Annotation @ LREC-COLING 2024*, pages 47–52, Torino, Italia. ELRA and ICCL.

Hannes Rosenbusch, Felix Soldner, Anthony Evans, and Marcel Zeelenberg. 2021. Supervised machine learning methods in psychology: A practical introduction with annotated r code. *Social and Personality Psychology Compass*, 15.

Mattia Samory, Felix Soldner, and Veronika Batzdorfer. 2025. PsyCoMark - Psycholinguistic Conspiracy Marker Dataset. [Data set]. Version 0.0.2.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.