LATE-iimas at SemEval-2026 Task 10: Conspiracy Detection via DeBERTa-v3 Ensemble and Weighted Loss Optimization

José Vázquez-Cerrillo[1], Helena Gómez-Adorno[2], Gemma Bel-Enguix[3],
[1]Posgrado en Ciencia e Ingeniería de la Computación,
[2]Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas,
[3]Instituto de Ingeniería,
Universidad Nacional Autónoma de México.
joccerrillo@ciencias.unam.mx, helena.gomez@iimas.unam.mx, gbele@iingen.unam.mx

Abstract

This paper describes the system developed by the LATE-iimas team for Task 10 of SemEval-2026: Psycomark, specifically for Subtask 2, which involves conspiracy detection. Our approach was based on fine-tuning the popular pre-trained language model `DeBERTa-v3-Large`. To address the challenges inherent in the provided dataset, such as class imbalance and the linguistic ambiguity of the "Can't tell" label, we implemented a 5-Fold Stratified Cross-Validation technique combined with a Weighted Cross-Entropy Loss function. The final system, which operates using an ensemble of the resulting models, achieved a Weighted F1-Score of 0.75, placing it in the top 10 of the ranking. The source code is available at https://github.com/PLN-disca-iimas/psycomark-semeval26

1 Introduction

The prevalence of conspiracy theories on social media is becoming increasingly common and poses a significant challenge to the dissemination of genuine content. According to an experiment conducted by Vosoughi et al. (2018), fake news spreads too quickly on social media, mainly because people are drawn to share information they consider novel without verifying its source. This is because sharing novel information can make them appear "interesting" to other users on social media, and thus perhaps gain more followers.

In Tripathi et al. (2025), it is noted that thanks to social networks, fake news spreads at an alarming rate, reaching millions of people in a very short time. This isn't limited to a single language or country; it occurs across diverse cultures and regions worldwide.

Zubiaga et al. (2018) state that within the distribution of fake news, there are different types of rumors:

1. Rumors that are spread at the moment of a breaking news story.

2. Rumors that persist for an extended period of time because there is no evidence to disprove them.

This last case is particularly noteworthy, as Lim et al. (2024) note that trying to dismantle a conspiracy rumor can end up being counterproductive; when scientific evidence shows it to be false, people who have believed in that rumor for a long time simply adopt new conspiracy theories.

Helping to detect and counteract conspiracy theories can also be beneficial for various vulnerable groups in society. Koo et al. (2025) note that LGBQ+ people of color living with HIV suffer from high levels of medical distrust due to historical racism and homophobia, but also because of the spread of conspiracy theories such as: "AIDS was created by the government to control the population of color". Based on this, we can say that the spread of conspiracy theories is a social phenomenon that can affect vulnerable groups due to a lack of information or real fear based on past historical events.

Singhania et al. (2017) argue that the problem of fake news is extremely complex, due to the interpretations that can be given to it according to different demographics; in other words, even for humans, it can be difficult to define whether a text includes conspiracy theories or not because the social and demographic context influences it.

SemEval-2026 Task 10 addresses the type of problems mentioned in this section by automatically detecting language with conspiratorial tendencies in comments on the social network Reddit (Samory et al., 2026). The original pipeline provided by the organizers used the `distilBERT` (Sanh et al., 2020) model. Our main strategy focused on maximizing the model's generalization

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2432–2437

capacity using an advanced Transformer architecture (DeBERTa-v3-Large) and minimizing training variance through ensembling techniques. Unlike approaches that simplify the problem to a binary classification, our system explicitly models uncertainty by training on three classes (Yes, No, and Can't tell), penalizing errors in the minority class more heavily. As we will demonstrate, explicitly handling the "Can't tell" class through class weights proved to be a highly effective strategy to improve overall performance on the Weighted F1-Score metric.

2 Related Work

Detecting conspiracy theories is a highly ambiguous task due to the inherent uncertainty often found within these online discussions. An important study highlighting this phenomenon is the work by Samory and Mitra (2018), who investigated various events that generated conspiracy theories on Reddit and observed how these events increased interaction within the "r/conspiracy" subreddit. They noted that while older members frequently generate the actual theories, new members typically contribute uncertainty, anxiety, and confusing arguments. Building directly upon this insight, our system explicitly embraces such ambiguity by optimizing for the "Can't tell" class, rather than treating uncertain posts as mere noise.

On the other hand, Batzdorfer et al. (2021) conducted a study using word-embeddings (Word2Vec) and temporal dynamics during the first months of the COVID-19 pandemic, discovering semantic relationships between the tweets of a group of 109 people identified as generators of conspiratorial content, in addition to finding relationships with the semantics used in previous conspiracy theories, such as those criticizing 5G connectivity or anti-vaccine movements. However, at an individual level, it is very difficult to define the behavior of Twitter users, since not all of their posts necessarily have to do with conspiracy theories.

Our work uses the DeBERTa model (He et al., 2021) (Decoding-enhanced BERT with disentangled attention). Its unraveling attention mechanism and improved mask decoding capture long-range syntactic and semantic dependencies more effectively than older models. We selected this specific architecture because handling the complex, context-heavy, and often sarcastic rhetoric of conspiracy theories requires a robust understanding of distant contextual cues within the text.

Furthermore, it is shown that DeBERTa-large (He et al., 2021) shows better results in six out of eight NLU tasks (Natural Language Understanding tasks) of GLUE (Wang et al., 2019), compared to BERT-large (Devlin et al., 2019), RoBERTa-large (Liu et al., 2019), XLNet-large (Yang et al., 2020) and ELECTRA-large (Clark et al., 2020). Furthermore, it is important to note that RoBERTa, XLNet, and ELECTRA are pre-trained with 160GB of training data, while DeBERTa is pre-trained with 78GB of training data. Therefore, DeBERTa obtained the highest average in the GLUE development set with less data.

During the last few years, models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) were the language model standards. In the work carried out by Akbari et al. (2025), it was mentioned that in some specific tasks, the use of models trained with particular data can achieve better results than general-purpose models: for example, CT-BERT (Müller et al., 2023) (which is based on BERT) outperformed BERT-Large in several tasks, such as Twitter Sentiment SemEval, because it was specifically trained with Twitter data. Another example is BERTweet (Nguyen et al., 2020) (also based on BERT), which produced better performance results than the previous state-of-the-art models on Tweet NLP tasks like: Part-of-speech tagging, Named-entity recognition, and text classification.

3 System Overview

As mentioned earlier, this task involves classifying Reddit comments based on the presence of conspiracy theories. The input data consists of comment identifiers that were rehydrated to obtain the unaltered original text from the social network. At first glance, the dataset presents the challenge of class imbalance: posts containing conspiracy, as well as posts where even annotators label them as uncertain ("Cannot be determined"), are less frequent than comments without conspiracy. The following is an overview of the system developed during the execution of this task.

3.1 Model Architecture

We used microsoft/deberta-v3-large as the base model. This model has approximately 435 million parameters. We added a linear classification layer on top of the token output (Classify to-

ken (CLS)) to project the latent representations into three classes: "No", "Yes", and "Can't tell".

3.2 Preprocessing

Since the original data was provided as Reddit IDs "redacted", the base pipeline implements a re-hydration script that connects to the Reddit API.

For Reddit comments:

- No data augmentation or text cleaning (such as stopwords removal) was applied in order to preserve the full semantic context and the completely original tone of speech.
- The texts were truncated to a maximum length of 512 tokens to optimize computational performance.

3.3 Weighted training

One of the elements that became fundamental within our system is the management of class imbalance. Because the "Can't tell" class was observed to be significantly in the minority (see Figure 1), leading standard models to ignore it, it was decided to implement a custom WeightedTrainer that modifies the standard loss function. We calculated class weights inversely proportional to their frequency in the training set according to the formula:

$$w_c = \frac{N}{C \times n_c} \tag{1}$$

Where N is the total number of samples available, C is the number of classes (in our case we have 3 classes) and n_c is the number of samples in the class c. These weights are integrated into the loss function (see Equation 2), forcing the model to pay more attention to the difficult classes during backpropagation.

To address the class imbalance, particularly the under-representation of the "Can't tell" class, we replaced the standard objective function with a Weighted Cross-Entropy Loss. For a given observation, the loss is defined as:

$$\mathcal{L} = -w_c \log(p_c) \tag{2}$$

where:

- c denotes the true class label of the instance.
- $p_c \in [0, 1]$ represents the model's predicted probability for the true class c.
- $-\log(p_c)$ acts as the core penalty mechanism. It approaches 0 when the model is confident and correct ($p_c \to 1$); but grows exponentially large when the model is confidently incorrect ($p_c \to 0$).
- w_c is the class-specific scaling weight from Equation 1.

While standard cross-entropy implicitly assumes $w_c = 1$ for all classes, this formulation dynamically scales the error penalty based on inverse class frequencies. By assigning a proportionally larger w_c to the minority class, the model incurs a significantly higher loss when misclassifying it. This heavily penalizes the network's tendency to safely default to the majority class ("No"), forcing it to adjust its representations to accurately detect the minority instances.

3.4 Ensemble

To avoid overfitting to a specific validation partition, we combined the official training and development sets. Once this unified set was obtained, we applied Stratified 5-Fold Cross-Validation as follows:

- We trained 5 independent instances of the model, each using 80% of the data for training and 20% for validation.
- For the final inference, we employed a soft-voting ensemble strategy. Specifically, we extracted the raw logits from the 5 independent models, applied a Softmax function to convert them into normalized probability distributions, and then averaged these probabilities across all folds before applying Argmax to determine the final class label.

This strategy effectively helped to reduce the variance of the predictions and improve generalization.

4 Experimental Setup

The following shows the configuration used during the experimentation carried out by the complete pipeline.

4.1 Data

The data used was exclusively the one provided by the SemEval (Train) organizers and rehydrated. No external corpora were used. Figure 1 shows us

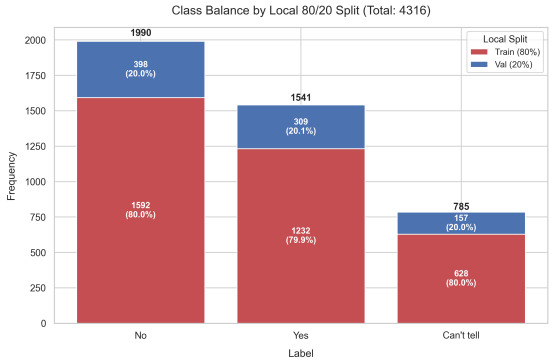

Figure 1: Class Balance by Local 80/20 Split

the distribution of the three different classes, which helps us appreciate the significant minority in the "Can't tell" class compared to the other two classes, being approximately 2.53 times smaller than the "No" class and approximately 1.96 times smaller than the "Yes" class. For this reason, the Weighted Cross-Entropy Loss was chosen to mitigate class imbalance (see Subsection 3.3).

4.2 Hyperparameters

The training was performed on an NVIDIA RTX A5000 GPU (24GB VRAM). The final hyperparameter configuration of the model, determined after several experiments, was as follows:

- **Optimizer:** AdamW with weight decay of 0.01.
- **Learning rate:** 3e-6. A low rate was crucial to stabilizing the fine-tuning of the Large model.
- **Batch size effective:** 32 (Achieved by *per_device_train_batch_size*=8 and *gradient_accumulation_steps*=4).
- **Epochs:** 7, with *early_stopping_patience=4*.
- **Warmup:** Ratio of 0.086 (approximately the first steps of training) to avoid initial divergence.
- **Precision:** FP16 (Mixed Precision) to reduce memory usage and speed up computing.
- **Gradient Checkpointing:** Enabled to allow training of the Large model in 24GB of VRAM.

5 Results

5.1 Quantitative Results

Given the hyperparameter configuration shown in the previous subsection, our system achieved a Weighted F1-Score of 0.75 in the official test set, placing it in tenth place in the competition, as shown in Table 1. To achieve this result, we performed experiments with different models and configurations, as can be seen in Table 2.

Table 1: Official performance comparison sent by the task organizers (considering only the "Yes" and "No" labels in the calculation of the F1-score metric).

Position	Participant	F1-Score
1	NJUST_KMG	0.89
2	mdok-style	0.78
10	**LATE-iimas**	**0.75**
27	Jia	0.37
28	GUNLP	0.34

Table 2: Performance comparison during the development phase in local validation.

Model Configuration	F1-Score (3-Classes)	F1-Score (2-Classes)
SVM + TF-IDF	0.5613	0.6577
Logistic Regression	0.5708	0.6644
Random Forest	0.5775	0.6771
Naive Bayes Multinomial	0.4871	0.6470
DistilBERT-Base-Uncased	0.5488	0.7234
DeBERTa-v3-Base	0.5516	0.7626
DeBERTa-Large (Single)	0.5632	0.7581
DeBERTa-Large + Weighted Loss	0.5799	0.6897
DeBERTa-Large + Weighted Loss + 5-Fold CV	**0.5913**	**0.7233**

It is important to address the variance between our local validation scores, peaking at 0.5913 Weighted F1, and the official test score of 0.75. During our development phase, we calculated the Weighted F1-Score across all three classes to accurately reflect the model's ability to handle ambiguity. However, an analysis of the official evaluation script provided by the organizers revealed that instances labeled as "Can't tell" are filtered out prior to calculating the final Weighted F1-Score.

Consequently, standard models that suffer from class collapse, effectively acting as binary classifiers, receive artificially inflated scores under the official 2-class metric. Conversely, our Weighted

Loss model is heavily penalized locally in the 2-class evaluation, as its cautious predictions on hard binary instances count strictly as false negatives. Despite this metric misalignment, our soft-voting ensemble successfully smoothed out individual model uncertainties, bridging the gap between a robust, nuanced 3-class system locally and a highly competitive 0.75 Weighted F1 on the binary-focused official test set.

We hypothesize that explicitly training on the "Can't tell" class acts as a regularizer for the latent space. By forcing the model to map ambiguous rhetoric to a distinct, penalized category, it prevents the decision boundary between pure "Yes" and "No" instances from becoming overly rigid or biased by sensationalist vocabulary. Given the inherently vague and shifting nature of conspiracy narratives, this implies that acknowledging and modeling ambiguity is a more effective moderation strategy than forcing a strict binary classification on complex discourse.

5.2 Ablation Analysis

Our ablation study (see Table 2) demonstrates the critical impact of both the weighted loss function and the ensembling strategy. Standard fine-tuning of `DeBERTa-Large` yielded a sub-optimal Weighted F1 of 0.5632, largely due to class collapse, where the minority "Can't tell" instances were ignored in favor of the majority classes.

Integrating the Weighted Cross-Entropy Loss mitigated this issue, improving the single-model performance to 0.5799 by explicitly penalizing misclassifications of the minority class and increasing its recall. Finally, the soft-voting ensemble mechanism provided the most substantial performance gain. By smoothing individual model variances and refining the decision boundaries across all three classes, the ensemble system jumped to 0.5913 in local validation, proving its robustness before evaluation on the official test set.

5.3 Error Analysis

To understand the model's limitations, we manually inspected a random sample of misclassified instances generated during our local 5-Fold validation. A qualitative analysis of these errors reveals that the ensemble struggles with the nuanced nature of conspiratorial discourse, primarily in three scenarios:

- **Sarcasm and Quotations (False Negatives):** The model struggles to separate the author's stance from quoted conspiratorial text. For example, a journalistic comment reporting that *"Fighters and coaches... have propagated the popular conspiracy theory... It has hit overdrive with QAnon"* was misclassified as **"No"** instead of the true label **"Yes"** (made by the annotators).

- **Missing External Context:** Comments dependent on external links or parent threads force the model into incorrect assumptions. A comment discussing a leaked geopolitical document *"Egypt's President... ordered the production of up to 40,000 rockets..."* was labeled **"Can't tell"** by annotators, but the model confidently misclassified it as **"No"**.

- **Sensationalist Language (False Positives):** Alarmist vocabulary easily triggers false positives. A text citing a biological study using terms like *"bombshell scientific article"* and *"killed thousands"* was predicted as **"Yes"**, despite the true label being **"No"**.

Future work could incorporate more data, more metadata, or a temporal analysis of users who frequently post conspiracy content, to provide the necessary basis for disambiguation on this extremely difficult topic.

6 Conclusion

In this paper, we described the LATE-iimas system for SemEval-2026 Task 10, achieving a competitive Weighted F1-Score of 0.75 (10th place in the official ranking). Our approach highlights that while advanced pre-trained language models like `DeBERTa-v3-Large` provide strong contextual representations, they are highly susceptible to the severe class imbalance typical of conspiratorial discourse.

We demonstrated that explicitly modeling uncertainty—rather than simplifying the problem to binary classification—through a Weighted Cross-Entropy Loss, combined with a variance-reducing soft-voting ensemble, is a highly effective methodology for this domain. Future work will explore targeted data augmentation using Large Language Models (LLMs) to synthetically balance the training distribution, as well as the integration of thread-level metadata to further disambiguate context-dependent claims.

Acknowledgments

This paper was supported by project PAPIIT IG400325. José Vázquez-Cerrillo (CVU-2158157) thanks the SECIHTI graduate degree scholarship program. The authors also thank Adrian Durán Chavesti, Ricardo Villareal, and Rita Rodriguez of the Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS) for their support with the computational resources used to run the experiments.

References

Rohullah Akbari, Daniel Schroeder, Petra Filkukova, and Johannes Langguth. 2025. Monitoring digital wildfires: a large-scale dataset of covid-19 conspiracy tweets created via fast nlp inference using the graphcore ipu. In *2025 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 241–250.

Veronika Batzdorfer, Holger Steinmetz, Marco Biella, and Meysam Alizadeh. 2021. Conspiracy theories on twitter: emerging motifs and temporal dynamics during the covid-19 pandemic. *International Journal of Data Science and Analytics*, 13:315 – 333.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

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

Gyo Hyun Koo, Soojeong Kim, Zhi Lin, Thomas J Johnson, Sungwon Jung, and Salih Hürdoğan. 2025. HIV conspiracy beliefs among black LGBQ+ people: The roles of public health sources, social media, algorithmic assistants, and religious leaders. *Health Commun.*, 40(12):2730–2744.

Dongwoo Lim, Fujio Toriumi, and Mikihito Tanaka. 2024. Revealing patterns in artificial earthquake misinformation: Detecting stubborn conspiracy adherents through social media analysis in japan. In *2024 IEEE International Conference on Big Data (BigData)*, pages 2983–2988.

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.

Martin Müller, Marcel Salathé, and Per E. Kummervold. 2023. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *Frontiers in Artificial Intelligence*, Volume 6 - 2023.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. *Preprint*, arXiv:2005.10200.

Mattia Samory and Tanushree Mitra. 2018. Conspiracies online: User discussions in a conspiracy community following dramatic events. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

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. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Sneha Singhania, Nigel Fernandez, and Shrisha Rao. 2017. *3HAN: A Deep Neural Network for Fake News Detection*, page 572–581. Springer International Publishing.

Abhishek Tripathi, Achyutesh Dixit, Vanam Narsimha, Madhan Punati, Sree Lakshmi Shashank, Kathem Venkata Maha Siva, and Subhashish Tiwari. 2025. Automated fake news classification with nlp and deep learning techniques. In *2025 International Conference on Intelligent Communication Networks and Computational Techniques (ICICNCT)*, pages 1–5.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. *Preprint*, arXiv:1804.07461.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. Xlnet: Generalized autoregressive pretraining for language understanding. *Preprint*, arXiv:1906.08237.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Comput. Surv.*, 51(2).