YNU-HPCC at SemEval-2025 Task 10: A Two-Stage Approach to Solving Multi-Label and Multi-Class Role Classification Based on DeBERTa

Ning Li, You Zhang*, Jin Wang, Dan Xu, and Xuejie Zhang
School of Information Science and Engineering
Yunnan University
Kunming, China
Contact: lining@stu.ynu.edu.cn, yzhang0202@ynu.edu.cn

Abstract

A two-stage role classification model based on DeBERTa is proposed for the Entity Framework task in SemEval 2025 Task 10. The task is confronted with challenges such as multi-labeling, multi-category, and category imbalance, particularly in the semantic overlap and data sparsity of fine-grained roles. Existing methods primarily rely on rules, traditional machine learning, or deep learning, but the accurate classification of fine-grained roles is difficult to achieve. To address this, the proposed model integrates the deep semantic representation of the DeBERTa pre-trained language model through two sub-models: main role classification and sub-role classification, and utilizes Focal Loss to optimize the category imbalance issue. Experimental results indicate that the model achieves an accuracy of 75.32% in predicting the main role, while the exact matching rate for the sub-role is 8.94%. This is mainly limited by the strict matching standard and semantic overlap of fine-grained roles in the multi-label task. Compared to the baseline's sub-role exact matching rate of 3.83%, the proposed model significantly improves this metric. The model ultimately ranked 23rd on the leaderboard. The code of this paper is available at: https://github.com/jiyuaner/YNU-HPCC-at-SemEval-2025-Task10.

1 Introduction

In subtask 1, given a news article and all named entity mentions (NEs) in that article, each mention is required to be assigned one or more role tags (Piskorski et al., 2025; Stefanovitch et al., 2025). Two levels of characters are defined: one for the main characters (protagonist, villain, innocent), and the other for the fine-grained characters. The evaluation criterion of the task is primarily the exact match rate, which measures the consistency between the main role and the fine-grained role of the evaluation prediction and the gold standard.

Semantic overlap (Kumar and Toshniwal, 2024) and ambiguity between fine-grained roles may occur (Peng et al., 2019), making it easy for models to be confused when differentiating. Entity roles are often determined based on the information in the context in which they are located. In news texts, the context in which entities appear may involve various metaphors, sarcasm, or indirect expressions, requiring the system to deeply understand and capture the details of the context. In real data, some roles (especially fine-grained roles) may have small sample sizes, leading to overfitting of common categories and under-identification of rare categories during model training.

In the past, rule-based methods (Grishman, 1996), traditional machine learning methods, or deep learning methods such as BERT (Devlin et al., 2019) and its variant DeBERTa (He et al., 2021), have often been used to solve similar entity character annotation tasks.

The two-stage classification model based on DeBERTa proposed in this paper utilizes the deep semantic representation of the pre-trained language model, combines the entity context, adopts Focal Loss (Lin et al., 2018) to address the category imbalance problem, and employs Optuna (Akiba et al., 2019) to fine-tune hyperparameters for handling entity role labeling tasks with multiple labels and categories.

Although the fine-grained role matching accuracy still has room for improvement, the model demonstrates effectiveness in main role recognition, with an accuracy rate of 75.32%.

The rest of the paper is organized as follows: Section 2 details the two-stage DeBERTa-based role classification model, including the main role and sub-role classification sub-models, and Focal Loss optimization for addressing class imbalance. Section 3 introduces the experimental setup, results analysis, and case study, covering dataset specifications, evaluation metrics, comparisons with base-

*Corresponding author.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1993–1999

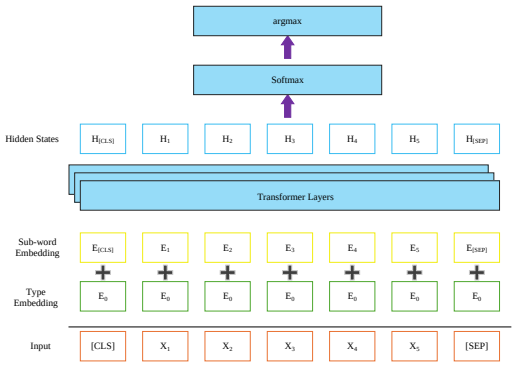

Figure 1: The structure of the system

line methods, implementation details, and a discussion of the limitations and challenges encountered in both experiments and real-world scenarios. Section 4 concludes the paper by summarizing the experimental outcomes and key findings.

2 DeBERTa-based Two-Stage Role Classification Model

In this paper, a two-stage classification model (Ruder, 2017) based on DeBERTa (Decoded Enhanced BERT with a Decoupled Attention Mechanism) is proposed to handle the complex task of character labeling of named entities in text. The overall architecture of the model is divided into two main sub-models: main role classification and sub-role classification. Each sub-model is designed to predict character labels at different granularities. The main role refers to a more macro category (e.g., protagonist, villain, and innocent), while the sub-role is a more granular category that describes the specific characteristics of an entity. Additionally, Focal Loss is employed to address the category imbalance issue.

2.1 Main Role Classification Sub-model

The main role classification sub-model is responsible for identifying the main role of an entity. Its primary task is to determine whether the entity belongs to the Protagonist, Antagonist, or Innocent category. The sub-model is based on the DeBERTa architecture and utilizes the powerful language representation capabilities learned from pre-training on a large-scale corpus. Through decoding enhancement and decoupling attention mechanisms (Zhang et al., 2021), deep contextual information in text can be effectively captured, and rich semantic representations can be generated by DeBERTa. The input entity mentions are processed by the DeBERTa tokenizer, which encodes the text sequence into a continuous contextual representation. This representation is then fed into the model. Based on these representations, logits values are calculated and output for each category, reflecting the probability distribution of an entity belonging to the Protagonist, Antagonist, or Innocent. The sub-model addresses the category imbalance issue (Buda et al., 2018) through the optimization of the Focal Loss function, improving recognition performance on rare categories.

2.2 Sub-role Classification Sub-model

The sub-role classification sub-model further refines the role labels of entities based on the main role classification. This sub-model is also based on the DeBERTa architecture but is optimized for more sub-role categories, such as *Guardian*, *Tyrant*, *Victim*, etc. The goal of sub-role classification is to identify the specific sub-roles of an entity within the main role category to which it belongs, providing a more granular role label for each entity.

2.3 Focal Loss Layer for Category Imbalance

The Focal Loss layer is integrated into the main role classification and sub-role classification models to address the category imbalance issue (Johnson and Khoshgoftaar, 2019). On unbalanced datasets, traditional loss functions (such as binary cross-entropy) (Zhang and Sabuncu, 2018) are often unable to effectively distinguish between different classes, resulting in excessive learning from samples of common categories and insufficient learning from rare classes during training. By introducing a tuning factor, Focal Loss reduces the focus on easy-to-classify samples and increases the learning of hard-to-classify samples, improving the model's performance on difficult-to-classify low-frequency classes. The formula for Focal Loss is as follows:

$$FL(p_t) = -\alpha(1-p_t)^{\gamma}\log(p_t) \quad (1)$$

where p_t is the predicted probability of the model for the correct class. α is a weight factor that adjusts the importance of different categories. γ is a focused parameter that reduces the relative loss to a correctly classified sample.

The loss function is applied to both the main role and sub-role classification models to ensure that more attention is given to hard-to-classify samples during training. This improves model performance

Type	Learning Rate	Focal Loss α	Focal Loss γ	Batch Size	Epochs
Value	4.86e-05	0.5202	2.9722	8	2

Table 1: Experimental results at each stage of model optimization

Method	Exact Match Ratio	micro P	micro R	micro F1	Main Role Acc.
Baseline	0.0383	0.0468	0.0415	0.0440	0.2581
Ours	0.0894	0.1149	0.1019	0.1080	0.7532

Table 2: Comparison of final results between our method and the baseline

on category-imbalanced datasets. By adopting a two-stage classification architecture and incorporating Focal Loss, the model effectively handles multi-label and multi-class role classification tasks, especially in cases of category imbalance. Experimental results demonstrate that this method achieves good results in both main and sub-role classification tasks, particularly in identifying minority classes (Byrd and Lipton, 2019).

3 Experimental Details

Datasets. The dataset is sourced from SemEval 2025 Task 10, Subtask 1 (role recognition task) and is only studied for the English part. The dataset contains the following: Training: Consists of multiple English-language news articles, each stored in plain text, with the start and end of entity mentions. Development: Serves as the basis for training the final model, which is ultimately used for predicting on the test set. Test: The final test dataset to be submitted for evaluation, the label of which will not be published until submitted. The number of entity mentions, the number of articles, and the distribution of tags for each role (e.g., main role and fine-grained sub-role) in the dataset bring challenges of multi-label, multi-category, and category imbalance to this task.

Evaluation Metrics. To fully evaluate the model's performance on entity character annotation tasks, the following metrics were used: Exact Match Ratio: Measure the accuracy of the label (fine-grained subpersona) predicted by the model versus the gold standard (Tsoumakas and Katakis, 2007). Micro Precision / Recall / F1: In multi-label, multi-category scenarios, micro-average precision, recall, and F1 scores are calculated from the prediction results of all sub-character samples (Sokolova and Lapalme, 2009). Main Role Accuracy: Specifically measures the accuracy of the model in predicting the main characters (Protagonist, Antagonist, Innocent). In this task, the official evaluation metrics are mainly based on the accuracy of fine-grained roles, etc., and the accuracy of the main roles is evaluated to reflect the system's ability to capture information about core characters.

Baselines. The official dev set provides two baseline prediction methods: random prediction and majority voting prediction. In random prediction, a role from the lists of main roles and sub-roles is chosen randomly. In majority voting prediction, the most frequently occurring main and sub-roles in the training data are selected as the predicted output. On the final test set, the baseline achieved a sub-role exact match rate of only 0.0383, and the main role accuracy was only 0.2851.

Implementation Details. In the final submission, the microsoft/deberta-base PLM was used as the text feature extractor and encoder, with its built-in tokenizer (DeBERTa Tokenizer) applied for text tokenization. SpaCy (Albade and Salisbury, 2022) was utilized for basic text cleaning and punctuation processing. For each entity mention, context of 100 characters before and after its start and end positions in the original text was extracted to ensure sufficient semantic information. The maximum token length was limited to 128 to meet the model's input requirements, and samples exceeding the length limit were checked.

Parameters Fine-tuning. In the Dev set, the Optuna tuning tool is used in this paper, and the optimal hyperparameter combination is shown in Table 1.

Result. Experimental results show that an accuracy of about 75.32% is achieved in the prediction of the main role, verifying the effectiveness of the proposed two-stage classification method in capturing core role information. Compared to the baseline's Main Role Acc of 25.81%, our method shows a significant improvement of approximately 49.51 percentage points. However, the exact match rate for sub-roles is low, at only 8.94%, mainly due to the strict matching requirements of multi-tag and

Method	Exact Match Ratio	micro P	micro R	micro F1	Main Role Acc.
Zero-shot	0.0085	0.0085	0.0075	0.008	0.1617
Similarity	0.0255	0.0298	0.0264	0.028	0.3915
Simple Description	0.034	0.0468	0.0415	0.044	0.3915
Bert	0.034	0.0468	0.0415	0.044	0.7362

Table 3: Experimental results at each stage of model optimization

α	γ	Exact Match Ratio	micro P	micro R	micro F1	Main Role Acc.
1	2	0.0979	0.1319	0.117	0.124	0.8255
1	1	0.0809	0.1064	0.0943	0.1	0.8255
1	3	0.0298	0.0681	0.0604	0.064	0.8255
0.5	1.5	0.034	0.0723	0.0642	0.068	0.8255
0.4	1.5	0.1149	0.166	0.1472	0.156	0.8255

Table 4: Parametric sensitivity experimental results for focal loss

multi-category tasks and the ambiguity between fine-grained roles. The low micro-average metric further indicates that fine-grained persona prediction remains challenging, particularly in terms of handling persona segmentation. Comparison with the benchmark model shows that, while the proposed method demonstrates clear advantages in predicting main roles, significant improvement is still needed for fine-grained role matching.

However, after the competition ended, the model was further optimized, resulting in significant performance improvements.

Firstly, an incremental approach was adopted to explore effective solutions for the role recognition task, and the experimental results at each stage are shown in Table 3:

- **Zero-shot learning baseline**: An unsupervised zero-shot learning method (Xian et al., 2020; Yin et al., 2019) was used, which performed poorly on the test set. The exact match rate was only 0.85%, and the main role accuracy was 16.17%. This confirmed the limitations of the unsupervised paradigm for this task.

- **Semantic enhancement method**: After introducing a BERT-based semantic similarity matching, the model performance improved significantly. The exact match rate increased to 2.55%, and the main role accuracy reached 39.15%. However, the micro F1 score remained below 3%, indicating that relying solely on surface-level semantic matching failed to capture the deeper role relationships.

- **Description enhancement strategy**: By adding role description information, the precision of sub-role prediction increased to 3.4%, and the F1 score reached 4.4%. However, the main role accuracy showed no significant change, suggesting that description information has a specific enhancement effect on fine-grained classification.

- **End-to-end classification model**: A BERT-based sequence classification model was built, with a 128-token length truncation strategy. MultiLabelBinarizer was used for label vectorization. Binary cross-entropy loss (BCEWithLogitsLoss) and the AdamW optimizer (learning rate 2e-5) were used, and the model was trained for 3 epochs with a batch size of 32. This approach led to an exact match rate of 0.34% and a main role accuracy of 73.62%, confirming the effectiveness of the end-to-end deep learning method for this task.

Subsequently, an end-to-end Transformer architecture based on DeBERTa-v3 was adopted, with dynamic context window extraction (extract_entity_context) employed to enable context-aware modeling. The model jointly learns the main role (forced constraint as Antagonist) and sub-role classification tasks, and a probabilistic filtering mechanism is introduced to constrain the sub-role candidate space. The loss function uses the improved binary cross-entropy Focal Loss (Equation 2), and the parameter sensitivity experiment results are shown in Table 4.

$$L = -\frac{1}{C}\sum_{c=1}^{C}\alpha_c(1-p_c)^{\gamma}y_c\log(p_c) \qquad (2)$$

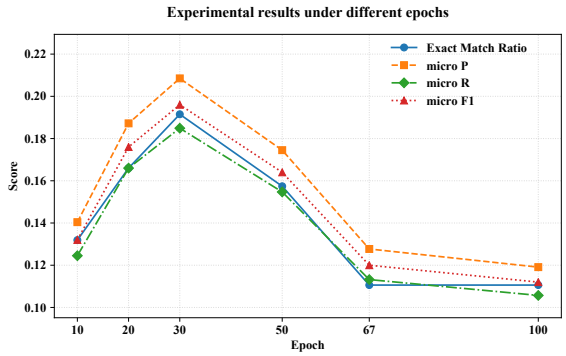

Figure 2: Experimental results under different epochs

The experiment shows that when α = 0.4 and γ = 1.5, the sub-role exact match rate reaches the optimal value of 11.49%. This parameter combination effectively alleviates the class imbalance issue by reducing the weight of the majority class ($\alpha <$ 1) and increasing the attention on difficult samples ($\gamma >$ 1).A contrastive learning framework based on Sentence-BERT (Reimers and Gurevych, 2019) was constructed, with the first 500 characters of the text directly extracted to generate embedding vectors (Wang et al., 2022). Sub-role classification was performed using a fully connected network (with the main role constrained to Antagonist). Standard cross-entropy loss (Equation 3) was used, and the training dynamics are shown in Figure 2.

$$L = -\sum_{c=1}^{C} y_c \log(\text{softmax}(z_c)) \qquad (3)$$

It can be observed that the sub-role exact match rate reaches its highest value of 0.1915 at epoch 30. As shown in the table, with the increase in training epochs, the model's performance on the minor categories first improved and then declined. The Exact Match Ratio and micro F1 peaked at epoch 30 (19.15% and 19.6%, respectively), and then gradually decreased due to overfitting (dropping to 11.06% and 11.2% at epoch 100).

Discussion. The overall exact match rate and other metrics for fine-grained role prediction are low, reflecting issues such as semantic overlap, data sparsity, and class imbalance between fine-grained roles in multi-label, multi-class scenarios. The model still requires further optimization.

Future work may explore the introduction of additional data augmentation strategies, deeper model architectures, and more advanced balancing techniques to further improve sub-role identification accuracy and the overall performance of the model.

Case Analysis. This study investigates the dynamic learning mechanisms and limitations of semantic models in complex moral categorization through a text-based entity classification task analyzing the role of *Washington* in the text fragment (EN_UA_DEV_100012.txt).

Zero-shot learning misclassifies the entity as *Protagonist/Virtuous* due to reliance on generalized narrative patterns, exposing how pre-trained knowledge obscures contextually critical semantics. Similarity matching and rule-based models generate biased labels like *Forgotten/Exploited* through shallow lexical associations (e.g., passive voice, resource-related conflict terms), confirming the failure of heuristic methods in decoding power-dynamic behaviors. While the BERT baseline captures the *Antagonist* primary category via pre-trained semantic understanding, its misattribution of systemic corruption to an individualized *Conspirator* reflects pre-trained models' inability to disentangle institutional power alienation mechanisms.

Parameter sensitivity tests (α/γ adjustments causing *Tyrant/Instigator* misclassifications) confirm that loss function design must balance explicit conflict terms and morally ambiguous representations to prevent attention mechanisms from fragmenting compound semantic features.

Experiments show that models need over 30 epochs to overcome initial stereotypical categorizations like *Foreign Adversary,* gradually forming stable correlations between *Corrupt* and implicit textual features (e.g., policy-embedded benefits, metaphors of power abuse). This hysteresis highlights moral categorization's reliance on deep contextual interdependencies.

The case underscores two challenges in entity classification: mitigating narrative biases while refining perception of power-ethics interactions. Future improvements via domain-specific knowledge graphs and adaptive attention could enhance models' ability to decode complex institutional semantics like systemic corruption.

4 Conclusions

In this paper, a DeBERTa-based two-stage role classification model is proposed for the role recognition task in SemEval 2025 Task 10, Subtask 1. The model successfully addresses the multi-label, multi-class role classification problem through a main

role classification layer and a fine-grained sub-role classification layer, and Focal Loss is used to optimize the class imbalance issue. This two-stage structure allows the model to not only accurately predict the main roles in news texts but also further identify fine-grained sub-roles, improving the precision of role categorization. Experimental results show that the model outperforms the baseline methods in all evaluation metrics. Main role accuracy reached 75.32%, but the exact match rate for sub-roles were low, at 0.0894, mainly due to the strict matching criteria in the multi-label task and semantic overlap in fine-grained roles.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. 61966038 and 62266051. We would like to thank the anonymous reviewers for their constructive comments.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. *arXiv preprint arXiv:1907.10902*.

James Von Albade and Joseph Peter Salisbury. 2022. Social media event detection using spacy named entity recognition and spectral embeddings. *World Congress on Electrical Engineering and Computer Systems and Science*.

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106:249–259.

Jonathon Byrd and Zachary C. Lipton. 2019. What is the effect of importance weighting in deep learning? *arXiv preprint arXiv:1812.03372*.

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

Ralph Grishman. 1996. Message understanding conference-6: A brief history. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-1996)*, volume 1, pages 466–470.

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

Justin M. Johnson and Taghi M. Khoshgoftaar. 2019. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54.

Ashish Kumar and Durga Toshniwal. 2024. Modeling text-label alignment for hierarchical text classification. In *Machine Learning and Knowledge Discovery in Databases. Research Track*, pages 163–179.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2018. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*.

Hao Peng, Jianxin Li, Senzhang Wang, Lihong Wang, Qiran Gong, Renyu Yang, Bo Li, S Yu Philip, and Lifang He. 2019. Hierarchical taxonomy-aware and attentional graph capsule rcnns for large-scale multi-label text classification. *IEEE Transactions on Knowledge and Data Engineering*, 33(6):2505–2519.

Jakub Piskorski, Tarek Mahmoud, Nikolaos Nikolaidis, Ricardo Campos, Alípio Jorge, Dimitar Dimitrov, Purificação Silvano, Roman Yangarber, Shivam Sharma, Tanmoy Chakraborty, Nuno Ricardo Guimarães, Elisa Sartori, Nicolas Stefanovitch, Zhuohan Xie, Preslav Nakov, and Giovanni Da San Martino. 2025. SemEval-2025 task 10: Multilingual characterization and extraction of narratives from online news. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. *arXiv preprint arXiv:1908.10084*.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437.

Nicolas Stefanovitch, Tarek Mahmoud, Nikolaos Nikolaidis, Jorge Alípio, Ricardo Campos, Dimitar Dimitrov, Purificação Silvano, Shivam Sharma, Roman Yangarber, Nuno Guimarães, Elisa Sartori, Ana Filipa Pacheco, Cecília Ortiz, Cláudia Couto, Glória Reis de Oliveira, Ari Gonçalves, Ivan Koychev, Ivo Moravski, Nicolo Faggiani, Sopho Kharazi, Bonka Kotseva, Ion Androutsopoulos, John Pavlopoulos, Gayatri Oke, Kanupriya Pathak, Dhairya Suman, Sohini Mazumdar, Tanmoy Chakraborty, Zhuohan Xie, Denis Kvachev, Irina Gatsuk, Ksenia Semenova, Matilda Villanen, Aamos Waher, Daria Lyakhnovich, Giovanni Da San Martino, Preslav Nakov, and Jakub Piskorski. 2025. Multilingual Characterization and Extraction of Narratives from Online News: Annotation Guidelines. Technical Report JRC141322, European Commission Joint Research Centre.

Grigorios Tsoumakas and Ioannis Katakis. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3):1–13.

Jin Wang, You Zhang, Liang Chih Yu, and Xuejie Zhang. 2022. Contextual sentiment embeddings via bi-directional GRU language model. *Knowledge-Based Systems*, 235:107663.

Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. 2020. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *arXiv preprint arXiv:1707.00600*.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint arXiv:1909.00161*.

You Zhang, Jin Wang, and Xuejie Zhang. 2021. Learning sentiment sentence representation with multiview attention model. *Information Sciences*, 571:459–474.

Zhilu Zhang and Mert Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. *arXiv preprint arXiv:1805.07836*.