# NCLTeam at SemEval-2025 Task 10: Enhancing Multilingual, multi-class, and Multi-Label Document Classification via Contrastive Learning Augmented Cascaded UNet and Embedding based Approaches

**Shu Li, George Snape, Huizhi Liang**
Newcastle University
Newcastle Upon Tyne, England
{c1041562, c0022646,huizhi.liang}@newcastle.ac.uk

## Abstract

The SemEval-2025 Task 10 Subtask 2 presents a multi-class multi-label text classification challenge(Piskorski et al., 2025). The task requires systems to classify documents simultaneously across three categories: Climate Change (CC), Ukraine-Russia War (URW), and Others. Several challenges were identified, including the distinct characteristics of climate change and warfare topics, category imbalance, insufficient training samples, and distributional differences across development and test sets. To address these challenges, we implemented two approaches. The first approach applies a Contrastive learning augmented Cascaded UNet model (CCU), which employs a cascaded architecture to explicitly model the label taxonomy. This model incorporates a UNet-style architecture to classify embeddings extracted by the base text encoder, with specialized pathways for different categories. We addressed data insufficiency through contrastive learning and mitigated data imbalance using an asymmetric loss function. The second approach implemented a Bi-Sequential Trees with Embeddings, Sentiment and Topics (BST-EST). In this approach, transformer encoder models were applied to extract word embeddings, then we applied classical machine learning based classifiers such as Random Forest and XGBoost. In the experiments, the CCU model achieves a higher F1 (samples) score (0.345) on the test set.

## 1 Introduction

SemEval-2025 Task 10 Subtask 2 introduces a multilingual, hierarchical, and multilabel document classification challenge. Our approaches integrate contrastive learning, hierarchical pathway modeling, and domain adaptation techniques to address these challenges. We also identified several data-specific issues, including insufficient training samples, significant class imbalance, and distribution shifts between development and test sets. Our methodology systematically addresses these challenges through strategic neural architecture design and feature engineering.

In the context of processing text data across multiple languages, existing research generally follows two approaches, pre-training model on massive multilingual corpora to enable cross-lingual transfer (Zhuang et al., 2021), or distilling multilingual knowledge into monolingual language models to optimize the computational efficiency(Reimers and Gurevych, 2020). In our work, we applied a pre-trained monolingual encoder model, fine-tuned with a multilingual dataset as our base encoder.

For hierarchical multi-label classification with limited samples, prior research such as the hierarchical verbalizer model employs prompt-based learning to incorporate label hierarchy knowledge into the model(Ji et al., 2023). In our approach, we explicitly defined the hierarchical data structure by organizing the dataset into topic-based sub-categories, and designing a corresponding cascaded model architecture specifically tailored to this dataset taxonomy. This architectural framework enforces a hierarchical prediction flow. Ensuring that classification results adhere to the inherent structure of the classification task. The Proposed BST-EST model utilizes the prompt templates with masked language models to leverage pre-trained knowledge for few-shot learning, dynamically capturing the label-text interaction.

To mitigate the significant differences between the CC and URW topics, we implement a Gradient Reversal Layer (GRL) for domain adaptation. The GRL adversarially aligns feature representations from different domains by reversing gradients during backpropagation. This technique encourages the feature extractor to produce domain-invariant features, enabling better knowledge transfer between the structurally similar but topically distinct CC and URW narratives.(Ganin and Lempitsky, 2015).



*Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 418–423

During inference, attention masks were applied to ensure the model predictions adhere to the natural hierarchy of labels, forcing classifications to respect the narrative and sub-narrative relationships in the label taxonomy. In detail, the parent category probabilities effectively act as an attention mechanism that gates the flow of information to specialized pathways. This ensures that the narrative and sub-narrative predictions are only evaluated for samples belonging to the correct parent category by the masked attention mechanism.

A gradient inverse layer was implemented to achieve domain adaption by learning domain-invariant features across CC and URW. Contrastive learning was applied to address the insufficient data (Ye et al., 2021) (Li et al., 2024). Our augmentation pipeline combines contextual word substitution and back translation to mitigate the influence of the data imbalance. During the experimentation, we observed a significant limitation in the cascaded model. When trained on both the narratives and sub-narratives classification tasks simultaneously, the model would effectively learn one task while performing poorly on the other. This phenomenon, which we identified through gradient flow visualization, appeared to be caused by gradient vanishing in one of the task pathway. By implementing asymmetric loss, we successfully addressed this problem, enabling the model to learn both tasks effectively. By combining these techniques, our approach provides a framework that achieves competitive performance for multilingual, hierarchical, multi-label text classification in low-resource scenarios.

## 2 Methodology

To address the multilingual, multi-class, multi-label documentation classification task, we applied a BST-EST model, and a CCU model to test the classification performance.

### 2.1 CCU model

Our primary approach integrates several methodological components including text data augmentation, contrastive learning, a cascaded UNet architecture, asymmetric loss functions, and an attention mask mechanism during inference.

#### 2.1.1 Data Augmentation

Initially, to reduce the distributional discrepancies between development and test sets. We apply two

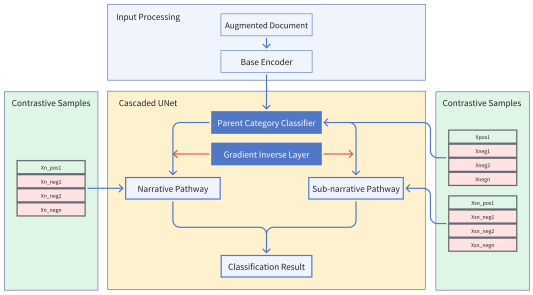


Figure 1: The CCU model architecture.

augmentation strategies, back translation and synonym replacement. These augmentation techniques expand the training corpus while maintaining label consistency. It is particularly critical for class imbalance. The dataset was split into sub-categories. The dataset was organized into a three level hierarchical structure, parent categories for 'CC', 'URW', and 'Others' at top level, followed by narratives and subnarratives were then split by topic at lower levels. The hierarchical data organization directly informed our model architecture, each pathways designed to correspond to each level of the dataset hierarchy. This design ensuring that the classification logits follow the label taxonomy.

#### 2.1.2 Model Architecture

The hierarchical model architecture was designed to adapt the dataset taxonomy. The model leverages the pre-trained all-MiniLM-L12-v2 transformer encoder model (Wang et al., 2020), with cascaded UNet for text classification. We selected MiniLM to enable rapid experimentation which benefits from its performance and computational efficiency. We chose to adapt the UNet architecture, traditionally used for image segmentation, for hierarchical text classification because of its structural advantages. The UNet encoder-decoder design with skip connections allows information to flow across different resolution levels, which we repurpose for hierarchical label prediction. In our text classification context, the UNet cascaded structure enables feature extraction at different levels, while the skip connections facilitate information sharing between domain classification and narrative and sub-narrative pathways. This architecture efficiently models our label taxonomy through its natural hierarchical processing.

Based on the parent category classification, the model activates one of two specialized domain pathways: the CC pathway or the URW pathway. The

skip connections facilitate information sharing between the domain classification and the narrative and sub-narrative pathways, allowing later classification decisions to leverage earlier domain-specific features while maintaining hierarchical consistency. This architecture efficiently models our label taxonomy through a hierarchical pathway processing flow.

Given our limited training data and the need to learn discriminative features across multiple languages and domains, we employed contrastive learning as a data-efficient training strategy. This approach maximizes the utilization of available samples by learning from both positive and negative pairs, effectively expanding our training samples without requiring additional labeled data. The implementation of contrastive learning improves the ability of the model to learn features from limited training data. A cosine embedding loss was implemented, which maximizes similarity between positive pairs while pushing negative pairs apart in the embedding space. Each pathway defined different sample strategies based on the sub-divided datasets.

The multi-class training objective combines three components: Cross-Entropy loss was implemented to classify the three parent category. To address severe class imbalance in our hierarchical classification task, we implemented asymmetric loss. During experimentation, we observed that the model struggled to learn rare classes, leading to gradient vanishing issues. Asymmetric loss assigns different penalties to false positives and false negatives, providing stronger learning signals for underrepresented classes and stabilizing the training process. During experimentation, we observed gradient vanishing through gradient flow visualization. The asymmetry loss was implemented to classify the sub-groups of narratives and sub-narratives to handle class imbalance(Ben-Baruch et al., 2021). Contrastive learning and cosine embedding loss were implemented to adapt to the small dataset. The final loss is computed as the sum of all loss values.

During inference, parent category probabilities thresholded at 0.5 dynamically activate subcategory pathways by applying an attention mask, ensuring classification results adhere to the hierarchical labels.

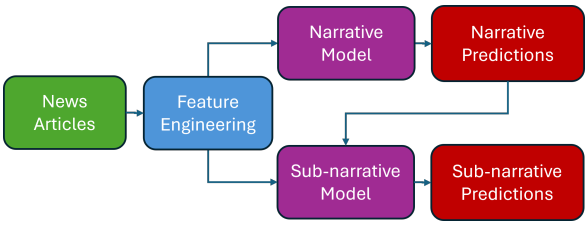


Figure 2: The BST-EST Model Architecture Diagram

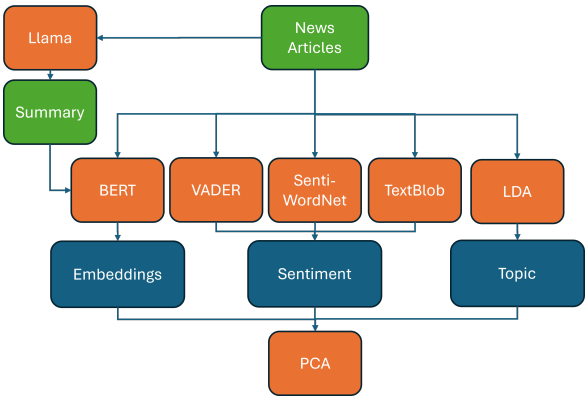


Figure 3: The BST-EST Feature Engineering pipeline

## 2.2 BST-EST model

In our second approach, we develop a hierarchical classification system that first transforms news articles into numerical representations using multiple feature-engineering techniques. These include BERT and ModernBERT embeddings for semantic encoding, BART-generated summaries for narrative abstraction, and complementary features from sentiment analysis (VADER, TextBlob, SentiWordNet) and LDA topic modeling. These representations are then processed by two machine learning models—ranging from logistic regression to XGBoost—with the first predicting broad narratives and the second using these predicting sub-narratives. Model selection and hyperparameter optimisation are performed systematically using Optuna's TPE sampler with k-fold cross-validation. The diagrams for the model architecture and the feature engineering process outline this visually.

Initially, the dataset needed cleaning. This included removal of newline characters and excessive whitespace to ensure that the language models could generate the most accurate inferences. To capture the rich context of the news articles and improve model performance, several advanced feature engineering techniques were applied. First, BERT embeddings were used to convert text into dense vector representations. The BERT and ModernBERT models were leveraged to generate embed-

dings by processing the input articles. Each article was tokenized and passed through the BERT model, with the embedding extracted from the special CLS token. These embeddings serve as powerful representations that capture both syntactic and semantic features of the text, allowing the models to better understand and process the articles.

Next, BART Summaries were used to extract narrative summaries from the articles. The BART model was employed to summarise each article by generating a concise version that captured the main narratives. These summaries were then passed through BERT to generate embeddings, further improving the model's ability to process the content. This step served as an abstraction layer that could help focus on the most relevant parts of the article when classifying narratives.

In addition to embeddings, several additional features were engineered using common sentiment analysis techniques and topic modeling(Liang et al., 2020). VADER Sentiment Analysis was applied to each article to generate sentiment scores(Hutto and Gilbert, 2014), including negative, neutral, positive and compound scores. This feature helped capture the overall emotional tone of the articles. TextBlob was also used to assess sentiment polarity and subjectivity, offering complementary information on the emotional stance and subjective nature of the text. Furthermore, SentiWordNet(Esuli et al.), a lexical resource for sentiment analysis, was employed to calculate the positive and negative sentiment scores based on word-level analysis. For further context, Latent Dirichlet Allocation (LDA) was also applied as a topic modeling technique(Blei et al., 2003). LDA helped extract the latent topics present in the articles by using a count vectorizer to convert the text into a bag-of-words format. This was followed by LDA to assign a distribution of topics to each article, which could be useful for understanding the broader themes or issues being discussed. The sentiment scores from Valence Aware Dictionary and sEntiment Reasoner (VADER) and TextBlob, as well as the topic distributions from LDA, were concatenated with the BERT embeddings to form a single feature vector for each article.

Optuna's Tree-structured Parzen Estimator (TPE) was used for the optimisation of Macro F1 for both models (Watanabe, 2023). The optimisation process incorporated k-fold cross-validation to ensure generalisability across datasets. it involved hyperparameter-tuning for a range of machine learning models, from simpler approaches such as logistics regression, decision trees and SVC to more advanced ensembles like Random For, GBM, XGBoost and LightGBM. This then saw exploring a wider range of the successful architectures. In addition to hyperparamter-tuning, the process also evaluated different which of the embeddings methods were best, the number of principal components and simultaneously with the hyperparameters. With the relatively small data sample, this was possible.

## 3 Experiment

The training set consists of 1699 samples distributed across five languages (English, Russian, Hindi, Portuguese, and Bulgarian). The taxonomy was designed to analyze and compare propaganda narratives in two conflict domains: the Russia-Ukraine war and climate change. Both categories employ similar broadcasting techniques that attempt to redirect responsibility, challenge opposing credibility, heighten concerns about negative outcomes, and divert attention. These approaches tend to polarize audiences and support specific political positions.

We further split the dataset into subsets based on hierarchical relationships, with narratives and sub-narratives grouped by their main topic (CC or URW). This organization forced the model to learn classification patterns that respect the category hierarchies. All language materials were combined into a unified multilingual dataset.

### 3.1 Evaluation Metric

The text classification task was evaluated using the F1-samples metric, which computes the F1 score averaged across all narrative and sub-narrative labels for each document, with systems ranked on the leaderboard based on this metric.

### 3.2 Experiment Setup

We implemented our methodology on NVIDIA A4000 and NVIDIA RTX3090 GPUs, using CUDA 12.4 and PyTorch 2.5.0. Our preprocessing pipeline included label binarization to convert labels into binary vectors, while preserving hierarchical relationships through manual subdivision of the dataset. The dataset contains samples in all languages, but inference was only performed on the English test set. For CCU model, we applied class weighting

Table 1: Text classification results on the test set.

| Model | F1 Macro (Coarse) | F1 St. Dev. (Coarse) | F1 (Samples) | F1 St. Dev. (Samples) |
|---|---|---|---|---|
| CCU | 0.486 | 0.363 | 0.345 | 0.360 |
| BST-EST | 0.354 | 0.440 | 0.311 | 0.437 |

to address label imbalance using inverse frequency weighting.

We experimented with several base models including BERT, ModernBERT, and all-MiniLM-L12-v2(Devlin et al., 2019)(Warner et al., 2024)(Wang et al., 2020). The document embeddings produced by the encoder were used for parent category classification (CC, URW, or Other) and then routed to specialized domain pathways. For domain adaptation, we implemented a GRL that adversarially aligns embeddings from different topics by reversing the gradient direction during backpropagation.

The pathways were designed to make narrative and sub-narrative classification to form up an end-to-end deep learning approach to handle the complexity of different levels of label granularity.

The training process applied the AdamW optimizer, with linear warmup in 10 percent steps followed by cosine decay, a batch size of 32 with gradient accumulation over 4 steps. Regularization methods applied including dropout (at p=0.1) in all dense layers, mixup interpolation(a=0.4 beta distribution), and gradient clipping(max norm=1.0).

For the BST-EST model, the base models included Logistic Regression, SVC, Decision Tree, Random Forest, GBM, XGBoost, LightGBM and ExtraTrees. Among these, ExtraTrees outperformed the others across evaluation metrics. Semantic meaning was captured through the use of embeddings produced by one of ModernBERT or BERT, using VADER, TextBlob, SentiWordNet for sentiment representation features along with LDA for topic based features.

Hyperparameter optimization was performed using a Tree-structured Parzen Estimator (TPE) method from Optuna, with a focus on the most important hyperparameters for each model architecture. Regularisation parameters were explored to control complexity and prevent over-fitting and for ensemble-based tree models, the number of estimators explored were limited to prevent over-fitting, given the small dataset. A range of other key hyperparameters, including those specific to each algorithm, were also searched to optimise model performance, focusing on a wider range for the better performing models. Model evaluation was also performed using a range of k (k=5, 6, ..., 10) for k-fold cross-validation.

## 4 Discussion

In this SemEval-2025 Task 10 Subtask 2, we successfully implemented our approaches and achieved 6th place. We proposed two methods to solve this multi-label multi-class multilingual text classification task. The CCU model, which integrates a pre-trained language model with a UNet cascaded classifier, hierarchical dataset organization, gradient reversal for domain adaptation, asymmetric loss, and contrastive learning. The BST-EST method using pre-trained transformer encoder model, to extract embeddings, also the sentiment message was added to the embeddings then classified with machine learning classifiers.

Furthermore, limitations still remain unsolved. For the CCU model, labels are represented using one-hot encoding rather than semantic text embeddings, which results in the loss of inherent label meaning.

## 5 Conclusion

This article presents an integrated implementation of existing solutions for multi-class, multi-label text classification to address SemEval-2025 Task 10 Subtask 2. Our team developed two distinct approaches to enhance classification performance. One built upon a cascaded UNet model with contrastive learning, and another leveraged multiple NLP feature extraction methods combined with machine learning classification techniques. Our efforts resulted in achieving 6th place with the main metric F1 sample of 0.345.

The CCU model addressed the challenges identified during experimentation. We mitigated data imbalance through augmentation process, while contrastive learning techniques helped overcome data insufficiency. The cascaded model architecture, with specialized pathways, was designed to tackle the complexities of multi-class learning across dif-

ferent label granularities. These solutions contributed to performance improvements.

Finally, the result remains below the optimal level for a reasonable classification. Furthermore, the integration of existing research offers limited novelty in the research field. Additionally, the pathway learning issue we identified, where the model would learn one pathway at the expense of another, suggests opportunities for developing more balanced training techniques for hierarchical classification models.

## References


Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. 2021. Asymmetric loss for multi-label classification.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Andrea Esuli, Fabrizio Sebastiani, et al. Sentiwordnet: A publicly available lexical resource for opinion mining.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France. PMLR.

C. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225.

Ke Ji, Yixin Lian, Jingsheng Gao, and Baoyuan Wang. 2023. Hierarchical verbalizer for few-shot hierarchical text classification.

Tian Li, Nicolay Rusnachenko, and Huizhi Liang. 2024. Chinchunmei at WASSA 2024 empathy and personality shared task: Boosting LLM's prediction with role-play augmentation and contrastive reasoning calibration. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 385–392, Bangkok, Thailand. Association for Computational Linguistics.

Huizhi Liang, Umarani Ganeshbabu, and Thomas Thorne. 2020. A dynamic bayesian network approach for analysing topic-sentiment evolution. *IEEE Access*, 8:54164–54174.

Jakub Piskorski, Tarek Mahmoud, Nikolaos Nikolaidis, Ricardo Campos, Alípio Jorge, Dimitar Dimitrov, Purificação Silvano, Roman Yangarber, Shivam Sharma, Tanmoy Chakraborty, Nuno 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, Vienna, Austria.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788. Curran Associates, Inc.

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.

Shuhei Watanabe. 2023. Tree-structured Parzen estimator: Understanding its algorithm components and their roles for better empirical performance. In *arXiv:2304.11127*.

Seonghyeon Ye, Jiseon Kim, and Alice Oh. 2021. Efficient contrastive learning via novel data augmentation and curriculum learning.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.