HierarchyEverywhere at SemEval-2024 Task 4: Detection of Persuasion Techniques in Memes Using Hierarchical Text Classifier

Omid Ghahroodi
NLP & DH Lab,
Computer Engineering Department
Sharif University of Technology, Tehran, IR
oghahroodi98@gmail.com

Ehsaneddin Asgari
Language Technologies Group
Qatar Computing Research Institute
Doha, Qatar
easgari@hbku.edu.qa

Abstract

Text classification is an important task in natural language processing. **H**ierarchical **T**ext **C**lassification (**HTC**) is a subset of text classification task-type. HTC tackles multi-label classification challenges by leveraging tree structures that delineate relationships between classes, thereby striving to enhance classification accuracy through the utilization of interclass relationships. Memes, as prevalent vehicles of modern communication within social networks, hold immense potential as instruments for propagandistic dissemination due to their profound impact on users. In SemEval-2024 Task 4, the identification of propaganda and its various forms in memes is explored through two sub-tasks: (i) utilizing only the textual component of memes, and (ii) incorporating both textual and pictorial elements. In this study, we address the proposed problem through the lens of HTC, using state-of-the-art hierarchical text classification methodologies to detect propaganda in memes. Our system achieved first place in **English Sub-task 2a**, underscoring its efficacy in tackling the complexities inherent in propaganda detection within the meme landscape.

1 Introduction

1.1 Propaganda Techniques in Memes

Propaganda can be defined as the deliberate dissemination of information, often with a biased or misleading nature, aimed at promoting or publicizing a particular political cause, ideology, or viewpoint. This communication tactic takes various forms, including persuasive messaging, advertising campaigns, and the dissemination of ideas through media channels. The primary objective of propaganda is to influence people's beliefs, attitudes, or behaviors towards a specific agenda or ideology. Examples of propaganda can range from political advertisements designed to sway voters, to ideological messaging spread through social media platforms.

Memes have emerged as one of the most prevalent communication tools in digital media. Their utilization of both text and image allows for the transmission of substantial information, underscoring the critical need for detecting propaganda within them.

1.2 Task Overview

SemEval-2024 Task 4 (Dimitrov et al., 2024) addressed the challenge of propaganda technique detection within memes in three sub-tasks (**1**, **2a**, **2b**) and four languages (**English**, **Bulgarian**, **North Macedonian**, **Arabic**). The organizers focused on different aspects of meme analysis: **Task 1** concentrated on detecting propaganda techniques from the textual content of memes, while **Tasks 2a** and **2b** respectively tackled the identification of techniques and the presence or absence of propaganda in a multimodal format. The SemEval-2024 Task 4 introduced three distinct sub-tasks across four languages. **English** language data was provided in supervised learning, whereas **Bulgarian**, **North Macedonian**, and **Arabic** language datasets were presented in a zero-shot learning framework. It is important to note that this task presented propaganda techniques in the form of a hierarchy, illustrated in Figure 1.

1.3 Hierarchical Text Classification

Hierarchical **T**ext **C**lassification (**HTC**) is a method wherein classes are organized in a hierarchical structure. This approach aims to enhance the accuracy of text classification models by leveraging the relationships within this hierarchy. We used the previous state-of-the-art (SOTA) hierarchical text classification model (**HPT** (Wang et al., 2022b)) to identify propaganda techniques in memes based on the hierarchical structure of propaganda techniques.

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1727–1732
June 20-21, 2024 ©2024 Association for Computational Linguistics

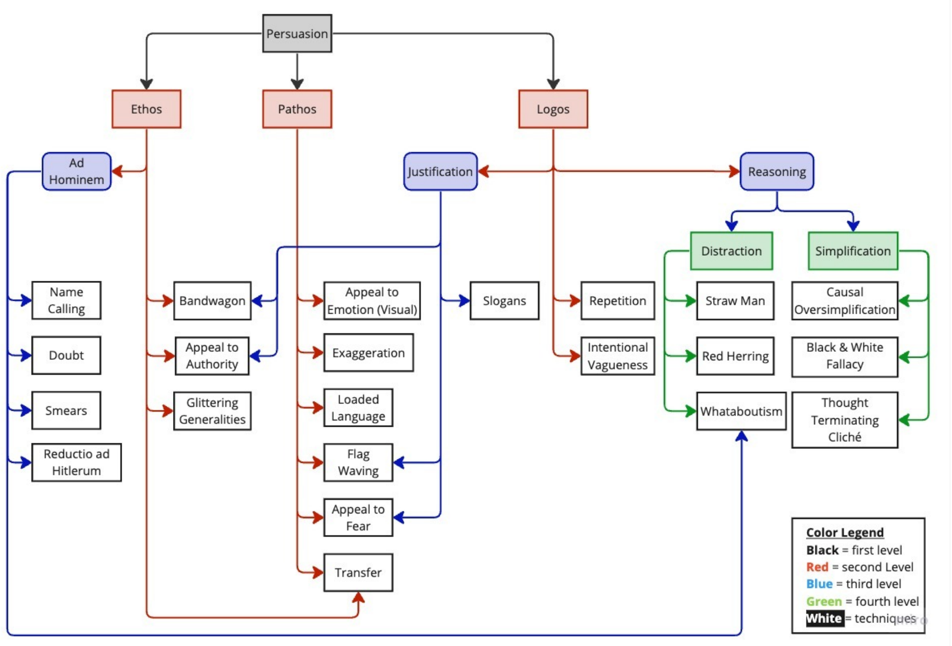

Figure 1: The diagram depicts propaganda techniques, represented as white nodes, organized in a directed acyclic graph (DAG). This image is sourced from the task description paper (Dimitrov et al., 2024)

In the multimodal section, we focused only on the textual content of memes, disregarding the accompanying images.

1.4 Our Discoveries

Our investigation revealed that employing hierarchical text classification models significantly enhanced text classification accuracy compared to various other methodologies and baseline approaches. Intriguingly, our decision to exclude the image component from consideration in **Task 2** resulted in the highest accuracy among all participating teams in **Task 2a**. We attribute this outcome to the inherent limitations of multimodal models in comprehending the intricate semantic relationships between images and text, particularly in the context of propaganda detection. Incorporating the image data would likely have increased model complexity and reduced accuracy, as observed in the performance of other teams. For the multilingual part, we rely on translation for non-English memes.

We utilized the HPT (Wang et al., 2022b) model source code[1], making necessary modifications to adapt it to our specific use case. The final version of our system's code has been made publicly available on GitHub for transparency and reproducibility[2].

2 Background

2.1 Dataset

The dataset utilized in this study comprises both textual and pictorial content extracted from memes along with associated propaganda technique tags (Dimitrov et al., 2024). Specifically, **Task 1** involves texts extracted from memes alongside propaganda technique tags, except **Loaded Language** and **Name Calling/Labeling** techniques, which are not included in the tags. **Task 2a** expands upon **Task 1** by incorporating images of memes, thereby presenting a multi-label classification task in a multi-modal format. **Task 2b** is similar to **Task 1a**, except that it involves binary classification regarding the presence or absence of propaganda. The organizers released the three tasks for the English language in a supervised manner and for **English**, **Bulgarian**, **North Macedonian**, and **Arabic** language in a zero-shot manner. The organizers released the dataset in three parts: training, validation, and testing sets.

2.2 Propaganda Detection

In recent research, the task of detecting propaganda in various forms of media has gained significant attention. (Da San Martino et al., 2020) introduce a task focused on identifying propaganda in news articles, comprising two subtasks: detecting spans containing propaganda and identifying

[1] https://github.com/wzh9969/HPT

[2] https://github.com/language-ml/SemEval-2024-Task-4

specific propaganda techniques from a predefined set of 14 techniques. On the other hand, (Dimitrov et al., 2021) presents a task aimed at detecting propaganda techniques in memes, without considering the hierarchy relation between techniques.

2.3 Hierarchical Text Classification

In this paper, we categorize existing hierarchical text classification models into three main categories: local methods, global methods, and generative methods.

1. **Local Methods**: Local methods tackle the hierarchical classification problem by addressing individual categories within the hierarchy. (Banerjee et al., 2019) employ binary classifications for each category and mitigate the issue of data scarcity at lower levels through transfer learning from parent to child categories. (Kowsari et al., 2017) adopt a strategy of training a multi-label classifier for each node, while (Dumais and Chen, 2000) employ SVM per level. (Shimura et al., 2018) leverage multiple CNNs to address classification at each level of the hierarchy.

2. **Global Methods**: Global methods take a holistic approach by employing a single classifier to predict all classes within the hierarchical structure. HiAGM (Zhou et al., 2020) utilizes two encoders, TreeLSTM and GCN, to derive the tree representation. They introduce two models, HiAGM-LA and HiAGM-TP, which respectively utilize attention mechanisms on classes and text propagation within the graph encoder. (Deng et al., 2021) aims to enhance the HiAGM by using information theory. (Mao et al., 2019) frame the hierarchical classification as a reinforcement learning problem, seeking an optimal policy for traversing suitable labels within the tree. (Zhu et al., 2023) employ structural entropy to construct the code tree, followed by using HiAGM-TP. (Wang et al., 2022a) introduce contrastive learning and positive samples to incorporate hierarchy into the text encoder. (Chen et al., 2021) attempt to unify label embedding and text embedding in a single space using triplet loss. In (Wang et al., 2022b), soft prompt tuning is employed, whereby each row of the hierarchy is fed into a graph attention network. Subsequently, the representations obtained from each row are provided as input to the BERT model. The model is trained to predict the correct label corresponding to the output of these tokens.

3. **Generative Methods**: (Yu et al., 2022) addressed the challenge of hierarchical classification by employing a method that generates a sequence of labels. Their approach involves training a T5 model to generate paths within the hierarchy. (Kwon et al., 2023) also tackled hierarchical classification through label generation. Notably, their approach enabled the model to generate n-grams not explicitly present in the predefined problem categories.

3 System overview

In **sub-task 1**, we addressed the proposed problem using hierarchical text classification and used a state-of-the-art (SOTA) HTC model for propaganda technique detection with some modifications. We utilized HPT (Wang et al., 2022b) as the hierarchical text classifier.

Convert Task to HTC Problem: The hierarchical structure of propaganda techniques was represented as a **D**irected **A**cyclic **G**raph (DAG). The hierarchy of propaganda techniques is depicted in Figure 1. To use the HPT model (Wang et al., 2022b), it was imperative to transform this DAG into a hierarchical tree. This transformation involved converting nodes with multiple parents into new nodes. For instance, the node "Whataboutism" with two parents, "Distraction" and "Ad hominem" was split into two nodes labeled "Distraction_Whataboutism" and "Ad hominem_Whataboutism". Two methods can be employed for organizing the first level of the hierarchy tree: **(1)** placing two nodes labeled "propagandistic" and "non-propagandistic" at the initial level, followed by the entire hierarchy of propaganda techniques under the "propagandistic" node, or **(2)** directly utilizing the hierarchy tree without this initial categorization. Our observations indicate that **method 1** yields superior performance.

Additional Datasets: We utilized two additional datasets, (Da San Martino et al., 2020) and (Dimitrov et al., 2021), as supplementary sources for training our model. In employing the data from (Da San Martino et al., 2020), we focused on its **TC sub-task**, which involves identifying propaganda techniques within news articles. The format of the data provided by (Da San Martino et al.,

Task	Model	HF1	HP	HR	Rank
English - Subtask 1	Best model	0.75247	0.68419	0.83590	1/33
	Our system	0.64252	0.63618	0.64899	12/33
	Our system†	0.65286†	0.63041†	0.67697†	9/34†
	Baseline	0.36865	0.47711	0.30036	31/33
English - Subtask 2a	Our system	0.74592	0.86682	0.65461	1/14
	Baseline	0.44706	0.68778	0.33116	13/14
Bulgarian - Subtask 1	Best model	0.56833	0.51955	0.62722	1/20
	Our system	0.46757	0.48301	0.45310	9/20
	Baseline	0.28377	0.31881	0.25567	18/20
Bulgarian - Subtask 2a	Best model	0.62693	0.70278	0.56586	1/8
	Our system	0.46414	0.67080	0.35483	7/8
	Baseline	0.50000	0.80428	0.36276	5/8
North Macedonian - Subtask 1	Best model	0.51244	0.51824	0.50677	1/20
	Our system	0.41713	0.48609	0.36531	10/20
	Baseline	0.30692	0.31403	0.30012	17/20
North Macedonian - Subtask 2a	Best model	0.63681	0.75019	0.55320	1/8
	Our system	0.35693	0.68903	0.24085	8/8
	Baseline	0.55525	0.90219	0.40103	4/8
Arabic - Subtask 1	Best model	0.47593	0.39140	0.60702	1/17
	Our system	0.40545	0.35638	0.47018	7/17
	Baseline	0.35897	0.35000	0.36842	14/17
Arabic - Subtask 2a	Best model	0.52613	0.55311	0.50166	1/8
	Our system	0.43685	0.50998	0.38206	6/8
	Baseline	0.48649	0.65000	0.38870	3/8

Table 1: The table presents the performance results of the hierarchical text classification model in comparison to both the baseline model and the best-performing model in sub-tasks 1 and 2a across four different languages: English, Bulgarian, North Macedonian, and Arabic. For each sub-task, the metrics HF1 (hierarchical F1 score), HP (hierarchical precision), and HR (hierarchical recall) are reported. † refers to the model trained initially on the (Dimitrov et al., 2021) dataset and subsequently fine-tuned on the task dataset, submitted after the test phase.

Task	Model	F1 macro	F1 micro	Rank
English - Subtask 2b	Best model	0.81030	0.82500	1/20
	Our system	0.56309	0.66167	16/20
	Baseline	0.25000	0.33333	20/20
Bulgarian - Subtask 2b	Best model	0.67100	0.81000	1/15
	Our system	0.48547	0.63000	10/15
	Baseline	0.16667	0.20000	15/15
North Macedonian - Subtask 2b	Best model	0.68627	0.84000	1/15
	Our system	0.50624	0.62000	6/15
	Baseline	0.09091	0.10000	15/15
Arabic - Subtask 2b	Best model	0.61487	0.63125	1/15
	Our system	0.56196	0.66875	5/15
	Baseline	0.22705	0.29375	15/15

Table 2: The table presents the performance results of the hierarchical text classification model in comparison to both the baseline model and the best-performing model in sub-task 2b across four different languages: English, Bulgarian, North Macedonian, and Arabic. For each sub-task, the Macro F1 and Micro F1 are reported.

2020) consists of spans within the news text annotated with corresponding propaganda techniques. To integrate this data into our model, we adopted an approach where if a span within a news article contained a propaganda technique, we assigned that particular technique to the entire article. It's important to note, however, that the dataset from (Da San Martino et al., 2020) does not encompass all the propaganda techniques featured in the SemEval-2024 task 4 dataset. Our analysis revealed that utilizing the data from (Da San Martino et al., 2020) in this manner led to a decrease in model accuracy. We attribute this reduction to two primary factors: **(1)** the broad attribution of propaganda techniques to entire news articles and **(2)** the differing distribution characteristics between news articles and meme text. The task of detecting propaganda techniques from memes, as outlined in (Dimitrov et al., 2021), served as another additional dataset for our study. Our analysis revealed that incorporating the data provided by (Dimitrov et al., 2021) enhanced the accuracy of our model.

[CLS] Token: Many memes comprise multiple sentences distributed across different picture boxes, delineated by "\n\n" in the dataset. To establish coherence between sentence boundaries, we utilized **"[CLS]" Token** between sentences. We observed that the inclusion of this token between sentences improves the performance of the model.

Other Tasks: In subtasks **2a** and **2b**, the image component of memes was disregarded, and only the textual content was provided to the model. Furthermore, for all the sub-tasks that are non-English, we used Google Translation API to translate them into English and used the model of the previous part

Baseline: According to the task description, the baseline for each sub-task is the most common label.

4 Experimental Setup

The organizers provided the data in three parts: training, evaluation, and testing sets. We employed the HPT model, utilizing the **bert-base-uncased** language model for our study. For training purposes, we combine the training and evaluation data, randomly picking **10%** for evaluation, and reserving the remaining **90%** for training. Our training comprised a batch size of **8** and a learning rate of **3e-5**. The remaining hyperparameters are similar to the HPT paper. To use additional data, we initially trained the model on this additional dataset before continuing training on the task data.

5 Results

The results for **sub-tasks 1** and **2a** are presented in Table 1, while the outcomes for **sub-task 2b** are shown in Table 2. Our system has exhibited strong performance in English language Task 1. In **sub-task 2a** for English, despite our model solely leveraging textual content from memes without considering images, it achieved the top ranking. We attribute this observation to two main factors: **(1)** The challenge of discerning the connection between images and text in propaganda detection **(2)** A substantial portion of the requisite information for propaganda detection likely resides within the textual component in addition to the image itself.

However, our system encountered challenges in non-English sub-tasks, displaying poor performance. We attribute this to potential translation errors and the absence of a pre-processing pipeline for these languages.

6 Conclusion

In this study, we addressed the challenge of detecting propaganda techniques in memes through two distinct sub-tasks: textual and multimodal analysis, conducted in both supervised and zero-shot settings across various languages. To tackle this issue, we employed hierarchical text classification. In the multimodal sub-tasks, we focused solely on the textual content of memes, achieving notable performance. However, when dealing with sub-tasks in languages other than English, our system's performance suffered. We concluded by presenting the metrics and conducting a thorough analysis of the results. Moving forward, our next objective is to develop a better hierarchical text classification model with better performance.

References

Siddhartha Banerjee, Cem Akkaya, Francisco Perez-Sorrosal, and Kostas Tsioutsiouliklis. 2019. Hierarchical transfer learning for multi-label text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6295–6300, Florence, Italy. Association for Computational Linguistics.

Haibin Chen, Qianli Ma, Zhenxi Lin, and Jiangyue Yan. 2021. Hierarchy-aware label semantics matching network for hierarchical text classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4370–4379, Online. Association for Computational Linguistics.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.

Zhongfen Deng, Hao Peng, Dongxiao He, Jianxin Li, and Philip Yu. 2021. HTCInfoMax: A global model for hierarchical text classification via information maximization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3259–3265, Online. Association for Computational Linguistics.

Dimitar Dimitrov, Firoj Alam, Maram Hasanain, Abul Hasnat, Fabrizio Silvestri, Preslav Nakov, and Giovanni Da San Martino. 2024. Semeval-2024 task 4: Multilingual detection of persuasion techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation*, SemEval 2024, Mexico City, Mexico.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. SemEval-2021 task 6: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 70–98, Online. Association for Computational Linguistics.

Susan Dumais and Hao Chen. 2000. Hierarchical classification of web content. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, page 256263, New York, NY, USA. Association for Computing Machinery.

Kamran Kowsari, Donald E. Brown, Mojtaba Heidarysafa, Kiana Jafari Meimandi, Matthew S. Gerber, and Laura E. Barnes. 2017. Hdltex: Hierarchical deep learning for text classification. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 364–371.

Jingun Kwon, Hidetaka Kamigaito, Young-In Song, and Manabu Okumura. 2023. Hierarchical label generation for text classification. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 625–632, Dubrovnik, Croatia. Association for Computational Linguistics.

Yuning Mao, Jingjing Tian, Jiawei Han, and Xiang Ren. 2019. Hierarchical text classification with reinforced label assignment. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 445–455, Hong Kong, China. Association for Computational Linguistics.

Kazuya Shimura, Jiyi Li, and Fumiyo Fukumoto. 2018. HFT-CNN: Learning hierarchical category structure for multi-label short text categorization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 811–816, Brussels, Belgium. Association for Computational Linguistics.

Zihan Wang, Peiyi Wang, Lianzhe Huang, Xin Sun, and Houfeng Wang. 2022a. Incorporating hierarchy into text encoder: a contrastive learning approach for hierarchical text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7109–7119, Dublin, Ireland. Association for Computational Linguistics.

Zihan Wang, Peiyi Wang, Tianyu Liu, Binghuai Lin, Yunbo Cao, Zhifang Sui, and Houfeng Wang. 2022b. HPT: Hierarchy-aware prompt tuning for hierarchical text classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3740–3751, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chao Yu, Yi Shen, and Yue Mao. 2022. Constrained sequence-to-tree generation for hierarchical text classification. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 18651869, New York, NY, USA. Association for Computing Machinery.

Jie Zhou, Chunping Ma, Dingkun Long, Guangwei Xu, Ning Ding, Haoyu Zhang, Pengjun Xie, and Gongshen Liu. 2020. Hierarchy-aware global model for hierarchical text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1106–1117, Online. Association for Computational Linguistics.

He Zhu, Chong Zhang, Junjie Huang, Junran Wu, and Ke Xu. 2023. HiTIN: Hierarchy-aware tree isomorphism network for hierarchical text classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7809–7821, Toronto, Canada. Association for Computational Linguistics.