Pauk at SemEval-2024 Task 4: A Neuro-Symbolic Method for Consistent Classification of Propaganda Techniques in Memes

Matt Pauk
University of Colorado Boulder
matt.pauk@colorado.edu

Maria Leonor Pacheco
University of Colorado Boulder
maria.pacheco@colorado.edu

Abstract

Memes play a key role in most modern information campaigns, particularly propaganda campaigns. Identifying the persuasive techniques present in memes is an important step in developing systems to recognize and curtail propaganda. This work presents a framework to identify the persuasive techniques present in memes for the SemEval 2024 Task 4, according to a hierarchical taxonomy of propaganda techniques. The framework involves a knowledge distillation method, where the base model is a combination of DeBERTa and ResNET used to classify the text and image, and the teacher model consists of a group of weakly enforced logic rules that promote the hierarchy of persuasion techniques. The addition of the logic rule layer for knowledge distillation shows improvement in respecting the hierarchy of the taxonomy with a slight boost in performance.

1 Introduction

Propaganda has long been used in media as a communication technique to influence people to subscribe to a particular idea or ideology. Identifying the presence of propaganda in media is an important subtask in building systems that can curtail the effect of propaganda campaigns. In modern times, a common way for propaganda to be spread is via memes. A meme is a either a short video or image, often overlayed with text, that is widely circulated over the internet. Identifying the presence and type of persuasion techniques used in memes is an important problem to be solved. The organizers of SemEval 2024 Task 4 propose a shared task for this very problem (Dimitrov et al., 2024). The shared task involves three sub-tasks: sub-task 1, identifying the persuasion technique(s) involved in the textual content of the meme; sub-task 2a, identifying which persuasion techniques are involved in both the visual and textual content of the meme; and sub-task 2b, identifying whether or not any persuasion technique is present in the visual and textual content of the meme. While the training data for all sub-tasks is in English, the evaluation phase includes different test sets for English, Arabic, Bulgarian, and North Macedonian. The proposed framework will focus only on the English version of the first two sub-tasks. Both 1 and 2a are hierarchical multi-label classification problems, where the goal is to classify the correct persuasion techniques used in the meme. The hierarchical nature of the persuasion techniques adds an extra element to this classification. All of the possible persuasion techniques are organized in a Directed Acyclic Graph (DAG), and full credit for a prediction is only given when the correct leaf node is predicted. Partial credit is given when any of its ancestors are given as a prediction.

There is a lot of previous work in the area of propaganda and persuasion identification. Much of this work has been based on previous SemEval shared tasks. The SemEval 2021 Task 6 is almost identical to the task explored here, without the hierarchical label structure (Dimitrov et al., 2021). The best-performing approaches on this task consisted of the use of a fine-tuned, text-based transformer for the textual content of the meme, some CNN or transformer based vision model to extract features from the image, and then a consolidation of the resulting embeddings via simple aggregation such as concatenation or average (Tian et al., 2021; Feng et al., 2021). Another SemEval task in 2023 focused on the identification of propaganda techniques in the text of news articles (Piskorski et al., 2023). In this case the best models used fine-tuned BERT based transformers (Wu et al., 2023; Hromadka et al., 2023).

The approach presented in this paper will also leverage a combination of a text-based transformer and a visual neural model. The key addition will be the incorporation of logic rules that encode the relationship between the possible hierarchical output classes. These relationships can be modeled by

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1424–1434

simple rules where the presence of a particular persuasion technique implies that its parent technique in the hierarchy is also present. For example, the rule `Straw Man` $\implies$ `Distraction` suggests that examples with the *Straw Man* persuasive technique also have it's parent technique *Distraction*. These rules will be used to promote predictions of persuasive techniques that respect the hierarchy. To test hierarchical consistency, we measure the number of hierarchy violations in predictions, where the model predicts a persuasive technique but not one of its ancestors in the hierarchy.

To incorporate logic rules, we take inspiration from the teacher-student logic rule framework proposed by Hu et al. (2016) that distills the information encoded in logic rules into the neural network parameters. The focus of this work is to explore if the incorporation of logic rules can improve on the results of neural based modes, while also producing more consistent results. The intuition is that distilling these hierarchical relationships into the weights of the network will allow the network to better recognize patterns that correspond to types of persuasive techniques, and ultimately make better predictions. We show that the addition of these logic rules does result in much more consistent predictions with a slight improvement in F1 scores.

2 Background and Related Work

Propaganda and Persuasion There is ample existing work in the identification of propaganda techniques. A similar SemEval task was proposed in 2021 to classify memes without including the hierarchy requirement (Dimitrov et al., 2021). Feng et al. (2021) proposed a framework for this task involving a text-based transformer built on RoBERTa, a visual feature extractor, and then a final transformer which takes as input the output of RoBERTa and the visual feature extractor. The authors consider two methods for this final encoder, a text pretrained transformer and a multi-modal transformer. The multi-modal transformer approach works the best, and they slightly improve on this score by combining several models together in an ensemble, which gives them state of the art results for the task. Tian et al. (2021) take a slightly different approach with this task. They similarly decompose the problem into a BERT based transformer and a visual feature extractor, but use a simpler method for combining the output embeddings by simply concatenating the output features before using a final classifier. They also experiment with a few different types of image feature extractors: a model specifically tuned to recognize faces, an Optical Character Recognition (OCR) model tuned to recognize text in an image, and the best performing extractor, which is a region based image feature extraction model that feeds into a multi-modal model.

Other previous SemEval tasks have focused just on the identification of propaganda techniques in text by leveraging multilingual datasets of news articles (Piskorski et al., 2023). Wu et al. (2023) and Hromadka et al. (2023) presented the two top performing systems for this task. Both use very similar approaches, leveraging a fine tuned version of RoBERTa for classification. Wu et al. (2023) had an interesting additional class weighting mechanism to try to improve performance on the under represented classes in the dataset. The approach proposed will leverage a lot of the same ideas regarding the usage of BERT based transformers for textual content, as well as visual feature extractors. Where this approach differs is in the incorporation of logic rules representing the relationship between propaganda techniques.

Hierarchical Classification Hierarchical multi label classification problems are split into two types: local methods, where an independent classifier is used for each node or for each level of the hierarchy; and global methods, which consider the entire hierarchy all at once (Levatić et al., 2014). In this paper, the interest is in exploring a global method that leverages logic rules to represent the relationship between classes in the hierarchy. Similar work has already been done in this area. Giunchiglia and Lukasiewicz (2021) propose a Coherent Hierarchical Multi-Label Classification Network (C-HMCNN) which uses a constraint layer on top of the regular network, as well as a specialized loss function to require the hierarchical constraints be satisfied. The constraint used is a simple one: the output probability of a subclass of a particular class in the hierarchy must be less than or equal to the output probability of its super-class. The approach for this project will be similar, but will follow more closely to the general logic rule integration method proposed by Hu et al. (2016). This framework consists of a teacher-student network, where the teacher network encodes logic rules as soft logic and distills that knowledge into the student network. This framework allows for more flexibility in the kinds of logic rules inte-

grated into a network. Additionally, it is less strict when enforcing the rules on the outputs of the neural network, allowing for a better balance between the signal coming from the direct supervision and the hierarchical knowledge.

3 System Overview

The proposed model architecture is based on a student-teacher knowledge distillation approach consisting of several components, which vary depending on the sub-task, but share a common structure. Regardless of task, base classifiers are used to encode raw textual and/or visual content. These resulting embeddings are then concatenated and used for predictions by the student network. The teacher network consists of a logic rule layer on top of the base student model that encodes the hierarchical information of the propaganda techniques. This logic layer is based on the teacher-student framework proposed by Hu et al. (2016). The framework distills the knowledge from the teacher network into the student network by training the student network to simultaneously emulate the gold labels and the teacher predictions.

3.1 Base Classifiers

Textual Model The textual model used will be the transformer-based model DeBERTa. DeBERTa was chosen based on its state of the art performance on short text datasets (Karl and Scherp, 2023). Additionally, this model was shown to do well on previous approaches for a very similar SemEval task (Feng et al., 2021; Tian et al., 2021). The version used is pre-trained version of DeBERTa proposed by He et al. (2021). This model is then fine-tuned on the textual meme content supplied by the SemEval 2024 task 6 organizers (Dimitrov et al., 2024).

Visual Model A ResNet-50 architecture is used for the vision component of sub-task 2a (He et al., 2016). This CNN based model is a medium sized model that achieved impressive results on the ImageNet classification task (Deng et al., 2009).

Combining the Textual and Visual Models The outputs of both textual and visual models need to be considered when making a prediction and therefore need to be combined in some way. A simple concatenation will be used and then inputted into a final feed forward network with sigmoid activations for each label.

3.2 Hierarchical Constraints

To introduce hierarchical constraints, we use a logical constraint layer on top of the base classifiers. The role of this layer is to distill knowledge coming from a set of logic rules into the weights of the classifiers. The implementation of this layer is based on the framework originally proposed by Hu et al. (2016) for the tasks of sentiment analysis and named entity recognition. The logic layer takes as input the sigmoid output for each label type from the base model and applies softened logic rules as regularization terms to obtain new predictions. The base network then learns weights based not only on the gold labels, but also based on the outputs of this logic rule regularized layer. This joint learning task is described by the following weight update equation which is a slightly modified version of the original formulation by Hu et al. (2016).

$$\theta^{t+1} = arg\min_{\theta \in \Theta} \frac{1}{N} \sum_{n=1}^{N} l_1(y_n, \sigma_\theta(x_n)) + l_2(s_n^t, \sigma_\theta(x_n)) \quad (1)$$

where θ represents the network parameters, N is the number of samples, l_1 is the loss function for the student network and gold labels, l_2 is the loss function for the student network predictions and teacher network predictions, y_n are the gold labels, $\sigma_\theta(x_n)$ is a vector of label probabilities outputted by the base network, and s_n^t is the output of the logic rule layer.

The output of the logic-rule based layer s_n^t is obtained by evaluating the following equation, also originally formulated by Hu et al. (2016).

$$q^*(Y|X) = p_\theta(Y|X) \exp\left\{ -\sum_{l,g_l} \lambda_l (1 - r_{l,g_l}(X,Y)) \right\} \quad (2)$$

Where $p_\theta(Y|X)$ is the output of the base model, λ is a weighting parameter used to determine how strictly to follow a particular rule, and r_{l,g_l} is a softened first order logic rule. The strategy for softening as well as the rules used for this particular application are described in the next section.

Representing Logic Rules This framework supports any FOL rules that can be grounded in the inputs, the output probabilities of the base network, and/or the gold labels. The rules can be softened using the following t-norms as found in the work done by Bach et al. (2017) regarding probabilistic

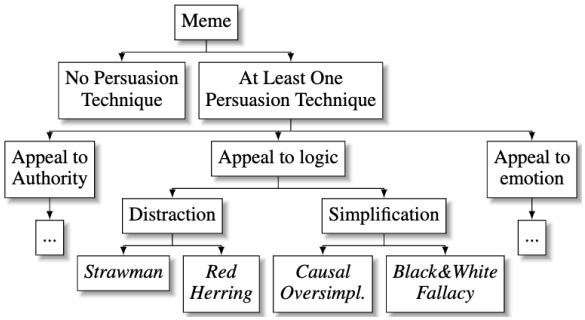

Figure 1: A subset of the hierarchy of propaganda techniques, provided by Dimitrov et al. (2024)

soft logic.

$$A \wedge B = \max(A + B - 1, 0)$$
$$A \vee B = \min(A + B, 1)$$
$$\neg A = 1 - A$$

After the rules have been converted to continuous representations they can be incorporated into the network via Equation (2). The following section describes the constraints that will be used for this specific classification task.

Hierarchical Logic Rules The rules for this application will focus on the hierarchical relationship between propaganda technique labels. The hierarchy of these labels is provided by the SemEval 2024 task 4 organizers, a sub-section of the hierarchy is shown in Figure 1 (Dimitrov et al., 2024). The full hierarchy includes 22 different persuasion techniques.

Given that the dataset is not balanced between all possible labels, we hypothesize that incorporating some of this hierarchical information into the model should allow for better prediction on lower coverage labels based on their relationship in the hierarchy to higher coverage labels. This hierarchical information will be encoded via logic rules that can be represented in the constraint layer via Equation (2), and will be distilled into the base model parameters via Equation (1) to help improve predictions.

The entire hierarchy can be represented via a sequence of rules:

$$\forall l \in L \text{ s.t. } l \neq root,\ l \implies par(l) \tag{3}$$

where L is the set of all possible labels and $par(l)$ represents the parent node of l in the hierarchy. Restructuring this rule slightly and grounding it in the outputs of the base model gives the following result:

$$\forall x \in X, \forall l \in L \text{ s.t. } l \neq root,\ \neg\sigma_l(x) \vee \sigma_{par(l)}(x) \tag{4}$$

where $\sigma_l(x)$ represents the probability that example x contains label l. Softening this logic gives the following expression for each rule:

$$\min(1 - \sigma_l(x) + \sigma_{par(l)}(x), 1) \tag{5}$$

Intuitively, these rules will enforce the hierarchy by penalizing predictions where the probability of a particular label is high, but the probability of its parent label is low.

4 Experimental Setup and Results

The method outlined above is evaluated on two separate but related tasks provided by the organizers of the SemEval 2024 Task 4 (Dimitrov et al., 2024). The details of these tasks and results are described below. Model implementations for both sub-tasks leverage Tensorflow for modeling and Hugging Face for the DeBERTa and ResNet models (Abadi et al., 2015; Wolf et al., 2019). [1]

4.1 SubTask 1

Experimental Setup The goal of subtask 1 is to identify which of the 20 persuasion techniques are present in the textual content of a meme. The dataset contains 7,000 labeled examples in the train set, 500 examples in the validation set, 1,000 in the dev set, and 1,500 in the test set. Hyper-parameter tuning is done on the validation set, with the final evaluation done on the dev and test sets.

Hierarchical F1, precision, and recall are used as the main evaluation metrics as defined by the SemEval Task organizers. However, micro F1, macro F1, and a count of the number of hierarchical violations are used as supplementary evaluation metrics. The hierarchical F1 score was originally formulated by Kiritchenko et al. (2006). This metric is the micro F1 of the label predictions, including both the actual persuasive techniques and their ancestor node categories in the hierarchy. The hierarchical violation metric is a count of the number of final true predictions which have an ancestor incorrectly marked as false. This metric is specifically used to evaluate whether the hierarchical logic rules are having an effect. The models presented will be compared against other approaches on the same

[1] Code can be found here: https://github.com/mappauk/Neuro-Symbolic-Final-Project

Rank	Model	HF1	Precision	Recall
1	914isthebest	0.752	0.684	0.836
2	BCAmirs	0.699	0.668	0.732
3	OtterlyObs...	0.697	0.648	0.755
⋮	⋮	⋮	⋮	⋮
15	**Pauk**	**0.627**	**0.716**	**0.573**
⋮	⋮	⋮	⋮	⋮
31	Baseline	0.369	0.477	0.300
32	WhatsaMeme	0.347	0.347	0.346
33	IIMAS1UTM1...	0.199	0.755	0.115

(a) Sub-task 1 (English) test set evaluation leaderboard. Our system, Pauk, places 15th.

Rank	Model	HF1	Precision	Recall
1	HierarchyEv...	0.746	0.867	0.655
2	NLPNCHU	0.707	0.782	0.645
3	BCAmirs	0.705	0.784	0.641
⋮	⋮	⋮	⋮	⋮
7	**Pauk**	**0.675**	**0.745**	**0.617**
⋮	⋮	⋮	⋮	⋮
12	BDA	0.504	0.477	0.493
13	Baseline	0.447	0.688	0.331
14	WhatsaMeme	0.366	0.313	0.440

(b) Sub-task 2a (English) test set evaluation leaderboard. Our system, Pauk, places 7th.

Table 1: Subtask 1 and 2a (English) test set evaluation results.

Model	Micro F1	Macro F1	Violations
NeuroSym(10)	0.574 ±0.009	0.283 ±0.015	13 ±7.8
NeuroSym(100)	0.583 ±0.006	0.301 ±0.018	7 ±4.2
NeuroSym(Max)	0.558 ±0.009	0.263 ±0.008	5 ±0.8
Baseline	0.581 ±0.002	0.307 ±0.018	42 ±20.9

Table 2: Sub-task 1 validation set results, comparing versions of NeuroSym with varying rule confidences against a baseline model leveraging only the classifiers and no logic layer.

Model	Micro F1	Macro F1	Violations
NeuroSym(10)	0.654 ±0.001	0.329 ±0.001	14.3 ±6.3
NeuroSym(100)	0.654 ±0.006	0.316 ±0.002	14.3 ±11.1
NeuroSym(Max)	0.651 ±0.003	0.329 ±0.01	13.3 ±5.8
Baseline	0.650 ±0.004	0.330 ±0.007	30.6 ±1.9

Table 3: Subtask 2a validation set results, comparing versions of NeuroSym with varying rule confidences against a baseline model leveraging only the classifiers and no logic layer.

task, as well as against themselves to measure the impact of the logical constraints.

The final hyper-parameters selected after tuning and used for evaluation on the dev/test sets are a learning rate of 3e-5, batch size of 4, and dropout of 0.1 after the BERT layer. Additionally, the teacher network rule confidences, represented by λ in Equation (2), are set to 100 for all rules. For dev set and validation set evaluations, the model is trained for 2 epochs on the training set. The model used to for the test set evaluation is trained for 3 epochs on the combined train and dev sets. Binary cross entropy is used as the loss function between the gold labels and student predictions, while KL Divergence is used between the student and teacher predictions.

Results The evaluation results for sub-task 1 on the English test set are shown in Table 1a. Dev set results can be found in Appendix A.1. Our system ranks 15th out of 33 submissions against the test set, as evaluated on the metric of hierarchical F1. The hierarchical F1 metric is a measure of micro F1 over all possible classes in the hierarchy after performing a post hoc operation to add the ancestors of predicted techniques to the list of predicted techniques for a particular example. Due to our systems focus on consistency of predictions with respect to the hierarchy, we also evaluate our system on micro F1 over all possible techniques without this post-hoc operation. Table 2 compares the average results of 3 runs against the validation set for our neuro-symbolic model with varying rule confidences of 10, 100, and Python's `sys.maxint`. In addition, a baseline version of the model without the hierarchical logic rule layer on top is added for comparison. The results for the individual runs can be found in Appendix A.2. Along with the F1 metrics, we present a measure of hierarchical violations over all predictions made.

The results show that regardless of the rule confidence used, the logic rule layer makes a noticeable improvement with regard to violations in the hierarchy of outputted predictions, with stronger rule confidences leading to less violations. This suggests that the rules are having their intended effect of making predictions consistent with the hierarchy. The results also show that very strong rule confidences seem to have a negative effect on F1 scores without much improvement in violations. The rule confidence of 100 seems to have the best compromise between consistent predictions and F1 scores, with even a slight improvement in F1 over the baseline model.

4.2 Sub-Task 2a

Experimental Setup The goal of sub-task 2a is to identify which of the 22 persuasion techniques are present in the textual and image content of a

Model	Micro F1	Macro F1
NeuroSym	0.374 ±0.014	0.162 ±0.006
Baseline(H)	0.433 ±0.016	0.227 ±0.017
Baseline	0.429 ±0.002	0.167 ±0.005

(a) Subtask 1 leaf node evaluation against the validation set.

Model	Micro F1	Macro F1
NeuroSym	0.474 ±0.005	0.218 ±0.005
Baseline(H)	0.488 ±0.005	0.233 ±0.02
Baseline	0.498 ±0.005	0.245 ±0.01

(b) Subtask 2a leaf node evaluation against the validation set.

Table 4: Leaf node evaluation to measure the effectiveness of the hierarchy usage in performance on leaf node propaganda technique predictions.

meme. The dataset distribution into train, validation, dev, and test is the same as task 1. The same hyper-parameter tuning method, final selected hyper-parameters, and evaluation metrics as used in sub-task 1 are also used here.

Results The results for sub-task 2a on the test set are displayed in Table 1b, with dev set results in Appendix A.1. Similar to sub-task 1, our model is in the middle of the pack, ranking 7th out of 14 for submissions on the test set. Table 3 is similar to Table 2 for sub-task 1, showing the results of experimenting on the validation set with varying rule confidences and a comparison to a baseline with no logic rule layer. Once again, we observe that the logic layer is leading to more hierarchically consistent predictions and a slight improvement in F1 scores. Additional results showing the model performance for each persuasive technique can be found in Appendix A.4.

Outside of consistency, one of the goals of using this logic rule student-teacher framework is to get the teacher model to distill information about the hierarchical relationship between the persuasive techniques into the student model and improve predictions on the actual leaf nodes representing specific persuasive techniques. In order to evaluate if this is actually the case, we perform an experiment evaluating just the predictions on the leaf nodes. For this experiment, the baseline model is trained and evaluated on only the leaf nodes of the hierarchy; Baseline(H) is trained on the full hierarchical data, but evaluated only on the leaf nodes; and NeuroSym includes the logic rule layer taking advantage of the hierarchical training data but also evaluated only on the leaf nodes. The results of the experiment averaged over three runs are shown in tables 4a and 4b. As shown in the results, the NeuroSym model has the lowest F1 scores when evaluated on both sub-task 1 and 2a. This indicates that the consistency enforced by the logic rule layer is actually negatively affecting leaf node predictions. The Baseline(H) model outperforms the baseline on sub-task 1 but performs worse on sub-task 2a, leaving inconclusive results as to whether the hierarchical data is helpful in leaf node prediction. The results of the individual runs, can be found in Appendix A.3.

5 Conclusion and Future Work

The framework presented attempts to solve the task of identification of persuasive techniques in memes. The key innovation involved in this framework compared to previous work done in this space is the integration of a logic rule knowledge distillation layer that weakly applies rules encoding the hierarchy of persuasion techniques. This layer is applied on top of a base model using a transformer based DeBERTa model for the textual component and a ResNet for the image component. We find that the logic rule network has some positive effect, consistently resulting in fewer hierarchical violations and a slight improvement in micro F1 scores. However, these logic rules do not lead to better predictions on the leaf node techniques themselves.

There were some difficulties in integrating these logic rules. The way the framework is set up, violations of the rules result in low probabilities for predictions by the teacher model. The part of the loss function that involves the KL divergence between these student and teacher predictions can cause the network to learn one of two aspects to minimize this loss. The first option is to raise the prediction probability of the rule violating ancestor label in the student network which will result in no rule violation and therefore no addition to the loss. Alternatively, the student network predictions can be lowered even further which also minimizes the KL divergence. The goal is for the former result to be learned, but it seems that often the latter is learned especially when the rule confidences are very high. Further work can be done to explore alternative logic rule interactions or loss function formulations to ensure the latter is always learned by the network.

In addition to improvements in the hierarchical logic rule integrations themselves, more work can be done to improve the base model by exploring other image processing methods outside of using a basic ResNet. Additionally, more intelligent ways of combining the textual hidden states and image hidden states can be explored, such as the use of a basic attention mechanism. Finally, the exploration of additional logic rules that promote parts of the textual content of an example that may indicate a particular persuasion technique could be experimented with. This may be especially useful for persuasion techniques with low coverage in the dataset.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Stephen H Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. 2017. Hinge-loss markov random fields and probabilistic soft logic.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

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.

Zhida Feng, Jiji Tang, Jiaxiang Liu, Weichong Yin, Shikun Feng, Yu Sun, and Li Chen. 2021. Alpha at SemEval-2021 task 6: Transformer based propaganda classification. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 99–104, Online. Association for Computational Linguistics.

Eleonora Giunchiglia and Thomas Lukasiewicz. 2021. Multi-label classification neural networks with hard logical constraints. *Journal of Artificial Intelligence Research*, 72:759–818.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Timo Hromadka, Timotej Smolen, Tomas Remis, Branislav Pecher, and Ivan Srba. 2023. KInITVeraAI at SemEval-2023 task 3: Simple yet powerful multilingual fine-tuning for persuasion techniques detection. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 629–637, Toronto, Canada. Association for Computational Linguistics.

Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard Hovy, and Eric Xing. 2016. Harnessing deep neural networks with logic rules. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2410–2420, Berlin, Germany. Association for Computational Linguistics.

Fabian Karl and Ansgar Scherp. 2023. Transformers are short text classifiers: A study of inductive short text classifiers on benchmarks and real-world datasets.

Svetlana Kiritchenko, Stan Matwin, Richard Nock, and A Fazel Famili. 2006. Learning and evaluation in the presence of class hierarchies: Application to text categorization. In *Advances in Artificial Intelligence: 19th Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI 2006, Québec City, Québec, Canada, June 7-9, 2006. Proceedings 19*, pages 395–406. Springer.

Jurica Levatić, Dragi Kocev, and Sašo Džeroski. 2014. The importance of the label hierarchy in hierarchical multi-label classification. *Journal of Intelligent Information Systems*, 45:247—-271.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multilingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.

Junfeng Tian, Min Gui, Chenliang Li, Ming Yan, and Wenming Xiao. 2021. MinD at SemEval-2021 task 6: Propaganda detection using transfer learning and multimodal fusion. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1082–1087, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Ben Wu, Olesya Razuvayevskaya, Freddy Heppell, João A. Leite, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2023. SheffieldVeraAI at SemEval-2023 task 3: Mono and multilingual approaches for news genre, topic and persuasion technique classification. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1995–2008, Toronto, Canada. Association for Computational Linguistics.

A Appendix

A.1 Dev Results

Tables 5a and 5b show the results of our model on the dev set for both subtask 1 and subtask 2a respectively. Our model performs in the middle of the pack for both subtasks, finishing 16th out of 33 for subtask 1 and 7th out of 11th on subtask 2a.

A.2 Validation Result Individual Runs

Tables 6 and 7 show the individual runs on the validation set measuring the effectiveness of the logic rule layer for both subtasks. The baseline model uses the classifiers with no logic rule layer, while the NeuroSym models use the classifiers and the logic rule layer with varying levels of confidence in the logic rules.

Model	Micro F1	Macro F1	Violations
NeuroSym(10)	0.571	0.270	7
NeuroSym(10)	0.586	0.304	8
NeuroSym(10)	0.565	0.275	24
NeuroSym(100)	0.576	0.318	13
NeuroSym(100)	0.582	0.276	4
NeuroSym(100)	0.591	0.308	4
NeuroSym(Max)	0.567	0.271	6
NeuroSym(Max)	0.558	0.267	4
NeuroSym(Max)	0.545	0.252	5
Baseline	0.583	0.285	39
Baseline	0.583	0.329	69
Baseline	0.578	0.308	18

Table 6: Individual runs on the validation set for subtask 1.

Model	Micro F1	Macro F1	Violations
NeuroSym(10)	0.654	0.328	12
NeuroSym(10)	0.653	0.331	23
NeuroSym(10)	0.656	0.329	8
NeuroSym(100)	0.653	0.314	5
NeuroSym(100)	0.652	0.312	8
NeuroSym(100)	0.658	0.322	30
NeuroSym(Max)	0.649	0.340	7
NeuroSym(Max)	0.653	0.331	12
NeuroSym(Max)	0.651	0.316	21
Baseline	0.648	0.336	28
Baseline	0.657	0.320	32
Baseline	0.647	0.334	32

Table 7: Individual runs on the validation set for subtask 2a.

A.3 Leaf Node Evaluation Individual Runs

Tables 8 and 9 show the results of the individual runs for the leaf node experiment for both subtask 1 and 2a. The Baseline model uses just the base classifiers trained and evaluated only on the leaf nodes of the hierarchy. The Baseline(H) model also uses only the base classifiers and is evaluated on only the leaf nodes of the hierarchy, but is trained on the full hierarchical data. Finally, the NeuroSym model is also evaluated on the leaf nodes but leverages the logic rule layer and the full hierarchical data.

Model	Micro F1	Macro F1
NeuroSym	0.376	0.161
NeuroSym	0.356	0.155
NeuroSym	0.389	0.169
Baseline(H)	0.431	0.203
Baseline(H)	0.414	0.242
Baseline(H)	0.453	0.237
Baseline	0.428	0.174
Baseline	0.427	0.163
Baseline	0.431	0.165

Table 8: Individual runs on the validation set for subtask 1 evaluating only the performance of predictions on the leaf node propaganda techniques.

Rank	Model	HF1	Precision	Recall
1	CLaC	0.881	0.808	0.967
2	OtterlyObs...	0.690	0.636	0.754
3	GreyBox	0.685	0.657	0.716
⋮	⋮	⋮	⋮	⋮
16	**Pauk**	**0.611**	**0.654**	**0.573**
⋮	⋮	⋮	⋮	⋮
31	nowhash	0.495	0.379	0.711
32	SINAI	0.430	0.315	0.677
33	Baseline	0.358	0.466	0.291

(a) Subtask 1 dev set evaluation results. Our system, Pauk, is ranked 16th out of 33.

Rank	Model	HF1	Precision	Recall
1	BCAmirs	0.699	0.770	0.640
2	NLPNCHU	0.697	0.767	0.639
3	SuteAlbastre	0.688	0.675	0.700
⋮	⋮	⋮	⋮	⋮
7	**Pauk**	**0.669**	**0.715**	**0.629**
⋮	⋮	⋮	⋮	⋮
9	Lomonoso...	0.648	0.774	0.557
10	hariswaqar	0.646	0.703	0.598
11	Baseline	0.446	0.685	0.331

(b) Subtask 1 and 2a dev set evaluation results. Out system, Pauk, is ranked 7th out of 11.

Table 5: Subtask 1 and 2a dev set leaderboards.

Model	Micro F1	Macro F1
NeuroSym	0.471	0.215
NeuroSym	0.481	0.224
NeuroSym	0.471	0.214
Baseline(H)	0.494	0.259
Baseline(H)	0.488	0.230
Baseline(H)	0.483	0.209
Baseline	0.503	0.259
Baseline	0.491	0.234
Baseline	0.499	0.243

Table 9: Individual runs on the validation set for subtask 2a evaluating only the performance of predictions on the leaf node propaganda techniques.

A.4 Results By Propaganda Technique

Tables 10 and 11 show the results of the dev set predictions on a per class basis. For both subtasks, we see the best performance for those classes higher up in the hierarchy due to the presence of the logic rules in the network as well as a larger number of training examples. Unsurpisingly, we see very poor performance for those techniques with very few training examples, ex: Obfuscation, Reductio ad hitlerum, Straw Man, and Red Herring. Unexpectedly, we see the F1 scores decrease for many of the leaf node propaganda techniques for subtask 2a despite having access to much more training data and getting overall higher F1 scores in aggregate. It appears this lift in F1 is due to the increase in F1 for the higher up nodes in the hierarchy such as Ethos, Pathos, and Ad Hominem as well as a drastic increase in a few leaf node techniques such as Smears and Loaded Language. Additionally, we see several more techniques with a F1 score of 0. This suggests that the images are helpful for classifying high level propaganda techniques and certain leaf techniques but actually confuse the model and lead to worse predictions than the textual model for many of the leaf node techniques.

Class	F1	Precision	Recall	Examples
Logos	0.76	0.76	0.76	545
Repetition	0.39	0.43	0.35	46
Obfuscation, Intentional vagueness, Confusion	0.00	0.00	0.00	8
Reasoning	0.56	0.53	0.60	278
Justification	0.70	0.73	0.65	343
Slogans	0.39	0.54	0.31	111
Bandwagon	0.22	1.00	0.22	16
Appeal to authority	0.85	0.83	0.87	136
Flag-waving	0.48	0.62	0.39	89
Appeal to fear/prejudice	0.19	0.40	0.12	66
Simplification	0.46	0.47	0.44	215
Causal Oversimplification	0.20	0.39	0.13	53
Black-and-white Fallacy/Dictatorship	0.38	0.39	0.37	98
Thought-terminating cliché	0.21	0.24	0.18	78
Distraction	0.30	0.38	0.25	72
Misrepresentation of Someone's Position (Straw Man)	0.00	0.00	0.00	10
Presenting Irrelevant Data (Red Herring)	0.00	0.00	0.00	10
Whataboutism	0.23	0.47	0.15	52
Ethos	0.80	0.79	0.81	610
Glittering generalities (Virtue)	0.41	0.47	0.37	71
Ad Hominem	0.71	0.71	0.70	506
Doubt	0.18	0.26	0.13	45
Name calling/Labeling	0.50	0.64	0.40	262
Smears	0.52	0.51	0.52	282
Reductio ad hitlerum	0.00	0.00	0.00	11
Pathos	0.65	0.68	0.62	427
Exaggeration/Minimisation	0.31	0.62	0.21	62
Loaded Language	0.55	0.63	0.48	303

Table 10: Results for each propaganda technique when evaluated against the submitted dev predictions for subtask 1.

Class	F1	Precision	Recall	Examples
Logos	0.77	0.79	0.75	583
Repetition	0.04	1.00	0.02	46
Obfuscation, Intentional vagueness, Confusion	0.00	0.00	0.00	12
Reasoning	0.54	0.56	0.53	284
Justification	0.69	0.77	0.62	379
Slogans	0.35	0.44	0.30	115
Bandwagon	0.00	0.00	0.00	18
Appeal to authority	0.85	0.81	0.90	143
Flag-waving	0.48	0.50	0.46	123
Appeal to fear/prejudice	0.00	0.00	0.00	78
Simplification	0.51	0.49	0.54	214
Causal Oversimplification	0.00	0.00	0.00	56
Black-and-white Fallacy/Dictatorship	0.37	0.34	0.41	103
Thought-terminating cliché	0.26	0.28	0.24	78
Distraction	0.16	0.53	0.10	83
Misrepresentation of Someone's Position (Straw Man)	0.00	0.00	0.00	11
Presenting Irrelevant Data (Red Herring)	0.00	0.00	0.00	10
Whataboutism	0.14	0.71	0.08	62
Ethos	0.91	0.89	0.94	847
Glittering generalities (Virtue)	0.39	0.38	0.39	92
Ad Hominem	0.81	0.79	0.84	660
Doubt	0.13	0.50	0.08	52
Name calling/Labeling	0.57	0.60	0.54	261
Smears	0.73	0.67	0.80	504
Reductio ad hitlerum	0.00	0.00	0.00	16
Pathos	0.73	0.75	0.70	635
Exaggeration/Minimisation	0.03	1.00	0.01	68
Loaded Language	0.64	0.70	0.58	306
Transfer	0.40	0.59	0.30	274
Appeal to (Strong) Emotions	0.03	0.33	0.02	56

Table 11: Results for each propaganda technique when evaluated against the submitted dev predictions for subtask 2a.