SemEval-2024 Task 4: Multilingual Detection of Persuasion Techniques in Memes

Dimitar Dimitrov[1], **Firoj Alam**[2], **Maram Hasanain**[2], **Abul Hasnat**[34],
Fabrizio Silvestri[5], **Preslav Nakov**[6] and **Giovanni Da San Martino**[7]
[1]Sofia University "St. Kliment Ohridski", [2]Qatar Computing Research Institute, HBKU, Qatar
[3]APAVI.AI, France, [4]BlackBird.AI, USA, [5]Sapienza University of Rome, Italy,
[6]Mohamed bin Zayed University of Artificial Intelligence, UAE
[7]Department of Mathematics, University of Padova, Italy
ilijanovd@fmi.uni-sofia.bg, {fialam,mhasanain}@hbku.edu.qa
fsilvestri@diag.uniroma1.it,hasnat@blackbird.ai
preslav.nakov@mbzuai.ac.ae, dasan@math.unipd.it

Abstract

The automatic identification of misleading and persuasive content has emerged as a significant issue among various stakeholders, including social media platforms, policymakers, and the broader society. To tackle this issue within the context of memes, we organized a shared task at SemEval-2024, focusing on the multilingual detection of persuasion techniques. This paper outlines the dataset, the organization of the task, the evaluation framework, and the outcomes.The task targets memes in four languages, with the inclusion of three surprise test datasets in Bulgarian, North Macedonian, and Arabic. It encompasses three subtasks: *(i)* identifying whether a meme utilizes a persuasion technique; *(ii)* identifying persuasion techniques within the meme's "textual content"; and *(iii)* identifying persuasion techniques across both the textual and visual components of the meme (a multimodal task). Furthermore, due to the complex nature of persuasion techniques, we present a hierarchy that groups the 22 persuasion techniques into several levels of categories. This became one of the attractive shared tasks in SemEval 2024, with 153 teams registered, 48 teams submitting results, and finally, 32 system description papers submitted.

1 Introduction

The rise of online social media platforms has enabled people to share their views and feelings openly. This increase in freedom of speech has significantly expanded the volume of digital content, offering valuable resources for initiatives like citizen journalism, raising public awareness, and supporting political campaigns. However, this freedom has also facilitated negative uses, leading to an increase in online hostility, as evidenced by the spread of content such as disinformation, hate speech, propaganda, and cyberbullying (Brooke, 2019; Joksimovic et al., 2019; Schmidt and Wiegand, 2017; Davidson et al., 2017; Da San Martino et al., 2019a; Van Hee et al., 2015).

Social media posts often combine various modalities, such as text, images, and videos. In recent years, *Internet memes* have become a prevalent form of content on these platforms. A meme is defined as "a collection of digital items that share common characteristics in content, form, or stance, which are created through association and widely circulated, imitated, or transformed over the Internet by numerous users." (Shifman, 2013) Memes generally consist of one or more images accompanied by textual content (Shifman, 2013; Suryawanshi et al., 2020). While memes are primarily aimed at humor, they can also convey persuasive narratives or content that may mislead audiences. To automatically identify such content, there have been research efforts directed towards addressing offensive content (Gandhi et al., 2020), identifying hate speech across different modalities (Gomez et al., 2020; Wu and Bhandary, 2020), and detecting propaganda techniques in memes (Dimitrov et al., 2021a).

Focusing on propaganda detection, research efforts have been specifically directed towards defining techniques and addressing the issue in news articles (Da San Martino et al., 2019), tweets (Alam et al., 2022b), memes (Dimitrov et al., 2021a), and textual content in multiple languages (Piskorski et al., 2023b). The associated shared tasks include SemEval-2020 Task 11 on news articles (Da San Martino et al., 2020), SemEval-2021 Task 6 on memes (Dimitrov et al., 2021b), WANLP-2022 and ArabicNLP-2023 focusing on Arabic

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 2009–2026
June 20-21, 2024 ©2024 Association for Computational Linguistics

(Alam et al., 2022b; Hasanain et al., 2023), and SemEval-23 Task 3 on news articles in multiple languages (Piskorski et al., 2023b).

The SemEval-2024 shared task extends previous tasks but introduces multilinguality, covering four languages, and features the largest dataset in English, along with a new hierarchical evaluation method. It has attracted significant participation. The task consists of three subtasks and was run in two phases: *(i)* the development phase and *(ii)* the evaluation phase. In the remainder of this paper, we define the tasks, describe the datasets, and provide an overview of participating systems and their official scores.

2 Related Work

2.1 Persuasion Techniques Detection

Past research on propaganda detection focused on analyzing documents as a whole to assess whether they contained propaganda. Barrón-Cedeno et al. (2019) created a corpus categorized into *propaganda* and *non-propaganda*, exploring the writing style and readability levels. Their results indicated that using distant supervision combined with comprehensive representations could lead the model to predict the source of the article instead of accurately differentiating between propaganda and non-propaganda content. An alternative approach to research has concentrated on identifying the use of specific propaganda techniques within texts. For example, Habernal et al. (2017, 2018) constructed a corpus containing 1.3k arguments, each annotated with different fallacies directly associated with propaganda techniques.

Building on previous work, Da San Martino et al. (2019b) created a corpus of news articles annotated for eighteen fine-grained propaganda techniques, approaching the problem as a task of span detection and classification. The majority of these studies have primarily focused on English. To address this gap in multilingual settings, Piskorski et al. (2023c) developed a dataset of news articles encompassing nine languages (Piskorski et al., 2023c). This dataset has enabled research into developing multilingual models.

Focusing on multimodality, specifically on memes, Dimitrov et al. (2021a) developed a corpus consisting of 950 memes and investigated various transformer models for automatic detection.

2.2 Multimodal Content

Multimodal content has been effectively utilized for propagating information and generating positive impacts. At the same time, it has also been used to cause harm (Sharma et al., 2022) or spread mis- and dis-information (Alam et al., 2022a). Research in this area include predicting misleading information (Volkova et al., 2019), detecting deception (Glenski et al., 2019), emotions and propaganda (Abd Kadir et al., 2016), hateful memes (Kiela et al., 2020), and propaganda in images (Seo, 2014).

To address the problem, current state-of-the-art research includes fine-tuning transformer models such as ViLBERT (Lu et al., 2019), Multimodal Bitransformers (Kiela et al., 2019), and VisualBERT (Li et al., 2019). Several studies have also explored the use of prompting strategies for hateful meme classification (Cao et al., 2022), aiming for detection from both text and visual modalities by leveraging (Prakash et al., 2023). For a recent survey, please refer to the work by Hee et al. (2024), which reports on the role of multimodality and LLMs in hateful content moderation.

2.3 Related Shared Tasks

To foster community engagement, several shared tasks on propaganda detection have been organized in the past. *SemEval-2020 task 11 on Detection of Persuasion Techniques in News Articles* (Da San Martino et al., 2020) focused on news articles, and asked to detect the text spans where propaganda techniques are used, and to predict their type (14 techniques). Closely related to that is the *NLP4IF-2019 task on Fine-Grained Propaganda Detection* (Da San Martino et al., 2019), which asked to detect the spans of use in news articles of each of 18 propaganda techniques. The *SemEval-2023 task 3 Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup* was focused on news articles covering nine languages Piskorski et al. (2023b). The WANLP'2022 and ArabicNLP'2023 shared task asked to detect the use of 20 propaganda techniques in Arabic tweets and news articles (Alam et al., 2022b; Hasanain et al., 2023).

The *SemEval-2021 Task 6 on Detection of Persuasion Techniques in Texts and Images* focused on identifying 22 persuasion techniques in memes (Dimitrov et al., 2021b). Following this prior work, we have significantly extended the size of the English dataset to 10K memes and added three surprise languages. The task is divided into three subtasks and also presents the persuasion techniques in a newly formed hierarchy allowing for better system predictions in cases of low confidence when predicting persuasion.

3 Tasks and Dataset

3.1 Tasks

The objective of the shared task is to develop models capable of identifying persuasion techniques (see Table 2 for a list and Dimitrov et al. (2021b) for a detailed description). This involves one subtask focused solely on analyzing the textual content of memes and another two subtasks dedicated to a multimodal analysis, where both textual and visual content are examined together. The subtasks are defined as follows:

Subtask 1 (ST 1): Given only the "textual content" of a meme, identify which persuasion techniques, organized in a hierarchy, it uses. If the ancestor node of a technique is selected, only a partial reward is given. This is a multilingual hierarchical multilabel classification problem.

Subtask 2a (ST 2a): Given a meme, identify which persuasion techniques, organized in a hierarchy, are used both in the textual and in the visual content of the meme (multimodal task). If the ancestor node of a technique is selected, only a partial reward is given. This is a multilingual hierarchical multilabel classification problem.

Subtask 2b (ST 2b): Given a meme, identify whether it contains a persuasion technique. This is a binary classification problem.

	EN					BG	MK	AR
Subtask	Train	Val	Dev	Test	**Total**	Test	Test	Test
ST 1	7,000	500	1,000	1,500	10,000	436	259	100
ST 2a	7,000	500	1,000	1,500	10,000	436	259	120
ST 2b	1,200	150	300	600	2,250	100	100	160

Table 1: Number of memes for every language on each subtask and associated data splits. Note that **only** the test split contains all four languages. EN=English, BG=Bulgarian, MK=North Macedonian, AR=Arabic

3.2 Dataset

Collection: We collected English, Bulgarian, North Macedonian, and Arabic memes from our personal Facebook accounts by scraping public Facebook groups, which focus on politics, vaccines, COVID-19, gender equality, and the Russo-Ukrainian War. However, Facebook groups did not provide enough memes for North Macedonian and Arabic therefore we collected some of the memes for these languages from Instagram. We considered a meme to be a "*photograph style image with a short text on top of it*", and we removed examples that did not fit this definition, e.g., cartoon-style memes, memes whose textual content was strongly dominant or non-existent, memes with a single-color background image, etc.

Annotation: The list of persuasion techniques and the annotation process are as described in (Dimitrov et al., 2021b). For each meme, we first annotated its textual content, and then the entire meme. We performed each of these two annotations in two phases: in the first phase, the annotators independently annotated the memes; afterward, all annotators met together with a consolidator to discuss and select the final gold label(s). This process was applied to each language, however, for English we had an additional step in the process where an expert linguist reviewed random samples of consolidated memes and communicated his observations back to the team of annotators. This was done to ensure we maintained high-quality annotations throughout the whole annotation campaign, considering the high cognitive complexity of the task.

Statistics: Table 1 shows the number of memes for each subtask in all four languages. The data for every subtask was split into train, validation, dev, and test as shown in the table. We introduced a validation set to allow parameter optimization on a predefined set of data, making it comparable across different systems. Bulgarian, North Macedonian, and Arabic were only used for the test set as they were surprise languages.

Table 2 and Table 3 show the label distribution for all subtasks. *Transfer* and *Appeal to (Strong) Emotions* do not apply to text, i.e., to Subtask 1. For Subtasks 1 and 2a, each technique can be present at most once per example. From the persuasion technique distribution we can see that the dataset is extremely imbalanced with some labels being present in more than 50% of the memes (Smears) and others in less than 1% (Obfuscation, Intentional Vagueness, Confusion). Moreover, Figure 2 (in Appendix A) shows that most of the memes contain more than one persuasion technique.

Persuasion Techniques	Subtask 1				Subtask 2a			
	EN	BG	MK	AR	EN	BG	MK	AR
Smears	2,838	41	25	17	5,159	320	220	63
Loaded Language	2,636	160	110	24	2,644	162	111	35
Name Calling/Labeling	2,284	140	83	26	2,294	148	95	33
Appeal to Authority	1,251	18	4	1	1,315	26	10	1
Black-and-White/Dictatorship	1,079	5	–	–	1,115	7	–	–
Slogans	994	62	23	–	1,024	68	26	–
Flag-Waving	834	28	6	1	1,179	43	14	2
Thought-Terminating Cliché	760	20	6	1	762	22	6	1
Glittering Generalities (Virtue)	703	5	–	2	991	29	5	2
Exaggeration/Minimisation	537	31	18	18	590	51	48	31
Appeal to Fear/Prejudice	527	35	13	8	643	73	52	43
Doubt	487	17	9	5	567	25	14	15
Repetition	442	19	3	1	445	19	3	1
Whataboutism	407	23	9	1	474	37	15	1
Causal Oversimplification	391	7	4	2	419	17	4	2
Bandwagon	144	2	–	1	157	6	–	1
Reductio ad Hitlerum	94	–	–	–	170	–	2	–
Straw Man	91	7	3	1	106	16	15	2
Presenting Irrelevant Data	87	3	1	1	91	–	1	3
Confusion	43	–	–	2	84	3	1	2
Transfer	–	–	–	–	2,286	141	113	–
Appeal to (Strong) Emotions	–	–	–	–	537	24	19	–
Total	**16,629**	**737**	**401**	**130**	**23,052**	**1,254**	**778**	**245**

Table 2: Persuasion techniques distribution for subtasks 1 and 2a in every language. For each technique, we show the number of instances.

Label	**EN**	**BG**	**MK**	**AR**
propagandistic	1,500	80	90	113
non propagandistic	750	20	10	47
Total	2,250	100	100	160

Table 3: Subtask 2b label distribution

We also observe a higher number of memes with 2 or more labels in ST2a which shows that a lot of memes require not only the text but the visual content to form enough context.

4 Evaluation Framework

4.1 Hierarchy

We introduce a hierarchy to allow the assignment of high-level categories in case of high uncertainty when predicting persuasion techniques. The persuasion techniques are grouped in a hierarchy, to be more precise a directed acyclic graph, as shown in Figure 3. The leaves of the hierarchy are the 22 persuasion techniques. The internal nodes are defined according to (Sourati et al., 2023; Piskorski et al., 2023a). Starting from the ROOT, we have the first level with *Ethos*, *Pathos*, and *Logos*. On the next level under Ethos – *Ad Hominem* and under Logos – *Justification* and *Reasoning*. Finally, Reasoning branches into *Distraction* and *Simplification*.

4.2 Evaluation Measures

Considering the hierarchical setup of the task, the evaluation metrics have to take into account the possibility of label assignment different than the original 22 persuasion techniques. Additionally, the metrics need to support a multilabel setting. We use adjusted F_1, $precision$, and $recall$ for hierarchical evaluation (Kiritchenko et al., 2006). For example, given the hierarchy in Figure 1, Let G be the ground truth value and H the predicted value, then to calculate the hierarchical measures we extend G to a set of its ancestor classes $S_{gold} = \{G, E, B, C\}$ and then do the same for H – $S_{pred} = \{H, E, B, C\}$. Then hierarchical $precision$, $recall$, and F_1 (hP, hR, and hF_1) would be:

$$hP = \frac{|S_{gold} \cap S_{pred}|}{|S_{pred}|} = \frac{|\{E, B, C\}|}{|\{H, E, B, C\}|} = \frac{3}{4} \tag{1}$$

$$hR = \frac{|S_{gold} \cap S_{pred}|}{|S_{gold}|} = \frac{|\{E, B, C\}|}{|\{G, E, B, C\}|} = \frac{3}{4} \tag{2}$$

$$hF_1 = \frac{2 \cdot hP \cdot hR}{hP + hR} = \frac{2 \cdot \frac{3}{4} \cdot \frac{3}{4}}{\frac{3}{4} + \frac{3}{4}} = \frac{3}{4} \tag{3}$$

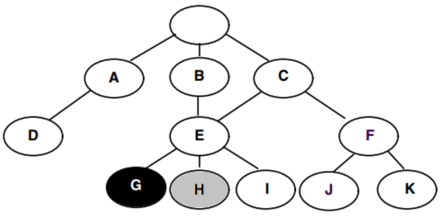

Figure 1: Example graph for hierarchical evaluation

Subtask	#team	#subm	#team	#subm
	EN Dev	EN Dev	EN Test	EN Test
ST 1	38	1159	35	130
ST 2a	11	61	14	28
ST 2b	20	457	20	51
Total	42	1677	42	209
	BG Test	BG Test	MK Test	MK Test
ST 1	20	29	20	29
ST 2a	8	13	8	10
ST 2b	15	20	15	21
Total	27	62	27	60
	AR Test	AR Test		
ST 1	17	36		
ST 2a	8	19		
ST 2b	15	24		
Total	24	79		

Table 4: Submission statistics. Note that only English has dev submissions, as the other languages were only released for test. *#team*: Number of teams that submitted results; *#subm*: Number of submissions.

Subtasks 1 and 2a are hierarchical multilabel classification problems. We used $hierarchical - F_1$ as the official evaluation measure. We also computed $hierarchical\ precision$ and $hierarchical\ recall$.

Subtask 2b is a binary classification problem. We used $macro\ F_1$ as the official evaluation measure. We also computed $micro\ F_1$.

4.3 Task Organization

The shared task was run in two phases:

Development Phase: During the development phase, we made *training* and *development* sets available for the participants. However, gold standard labels were not released for the development set. The participants submittd systems' results on the development set. They could make an unlimited number of submissions, and the best score for each team, regardless of the submission time, was shown in real time on a public leaderboard.

Test Phase: In this phase, we have released the test set and the *development* set together with the gold labels. The participants were given a week to submit their final predictions on the *test* set. It is important to note that the test data included memes in three additional languages such as Bulgarian, North Macedonian, and Arabic, which were not disclosed to the participants in advance as surprise languages. Similar to the development phase, participants could submit multiple entries; however, they have not received any feedback on their performance. Only the latest submission from each team was considered official and used to determine the final team rankings. Overall, 153 teams registered for the task, out of which 48 made official submissions. Moreover, 24 teams submitted results for all four languages. Specifically, 17 teams submitted results for all languages for ST1, 8 for ST2a, and 14 for ST2b, respectively. The total number of submissions across both phases was 2,087, with 1,677 on the development set and 410 on the test set. More details on submission statistics can be found in Table 4.

The results for the development and the test phases are available on the leaderboard page.[1] After the competition was over, we left the submission system open for the test dataset for post-shared task evaluations and to monitor the state of the art for the different subtasks across the languages.

4.4 Baseline Systems

Due to the highly imbalanced dataset, as seen in Table 2, the baseline for each subtask is the most common label or majority class baseline, i.e., for each meme, we make a prediction with the most frequent label. *Smears* is the most frequent label for Subtasks 1 and 2a, and *propagandistic* is the most frequent label for Subtask 2b. Note that the baseline is chosen according to the most common label across all languages.

5 Results

5.1 English Subtasks

The results for the three English subtasks are presented in Tables 5 and 6. All systems outperformed the baseline and the winning system is noticeably better than the second in subtasks 1 and 2a. In Subtask 2b there are three teams with top performance, two winning systems ex-aequo, and a third with a 0.001 difference in F_1.

We now briefly describe some of the top systems for each subtask. In Subtask 1 **914isthebest** (Li et al., 2024a) developed a transformer-based model

[1] https://propaganda.math.unipd.it/semeval2024task4/leaderboard.php

with in-domain pre-training. For system training, the training dataset was augmented following a Chain-of-Thought-based data augmentation approach using GPT-3.5. The main classification architecture includes four RoBERTa models and one DeBERTa model initialized using different random seeds. A soft voting approach, which averages the predicted probabilities of each label from all five models, is used to predict labels.

In Subtask 2a **HierarchyEverywhere** (Ghahroodi and Asgari, 2024) adapted the hierarchical text classification (HTC) model to the task by placing “propagandistic" and “non-propagandistic" nodes at the initial level and utilizing the “[CLS]" Token between sentences in memes enhanced model performance (Wang et al., 2022). Moreover, they employed additional datasets. Interestingly, 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, Google Translation API was used to translate them into English.

In Subtask 2b, **LMEME** (Li et al., 2024b) proposed a detection system that employs a Teacher Student Fusion framework. Initially, a Large Language Model serves as the teacher, engaging in abductive reasoning on multimodal inputs to generate background knowledge on persuasion techniques, assisting in the training of a smaller downstream model. The student model adopts CLIP as an encoder for text and image features, incorporating an attention mechanism for modality alignment.

5.2 Bulgarian Subtasks

The results for the Bulgarian subtasks are reported in Tables 7 and 8. For Subtask 1, seventeen out of nineteen systems outperformed the baseline; for Subtask 2a, four out of seven systems outperformed the baseline; and for Subtask 2b, all systems outperformed the baseline.

We briefly describe some of the top systems for each subtask. The top system, **OtterlyObsessedWithSemantics** (Wunderle et al., 2024) for Subtask 1, used a custom classification head that is designed to be applied atop a large language model. For the non-English test sets, the system was used after translating all documents to English using GPT-4.

For Subtask 2a, the top system is **BCAmirs** (Abaskohi et al., 2024). It involved using GPT-4 to generate a descriptive caption of the meme. The caption is then combined

R	Team	hF1	hP	hR
	English - Subtask 1			
1	914isthebest	0.752	0.684	0.836
2	BCAmirs	0.699	0.668	0.732
3	OtterlyObsessedWithSemantics	0.697	0.648	0.755
4	TUMnlp	0.674	0.638	0.714
5	GreyBox	0.670	0.652	0.688
6	NLPNCHU	0.663	0.610	0.726
7	Puer	0.660	0.648	0.673
8	EURECOM	0.655	0.628	0.685
9	SuteAlbastre	0.652	0.633	0.673
10	UMUTeam	0.648	0.708	0.597
11	RDproj	0.643	0.575	0.728
12	HierarchyEverywhere	0.643	0.636	0.649
13	nowhash	0.641	0.612	0.673
14	ShefCDTeam	0.640	0.662	0.618
15	Pauk	0.627	0.716	0.557
16	IUSTNLPLAB	0.625	0.632	0.618
17	whatdoyoumeme	0.617	0.598	0.638
18	LomonosovMSU	0.613	0.712	0.539
19	SoftMiner	0.607	0.649	0.569
20	MagnumJUCSE	0.603	0.547	0.673
21	IITK	0.591	0.596	0.586
22	CLaC	0.578	0.501	0.685
23	BAMBAS	0.577	0.501	0.679
24	MemeSifters	0.575	0.576	0.573
25	fralak	0.557	0.478	0.668
26	IIITG	0.526	0.614	0.459
27	Two	0.522	0.526	0.518
28	Scalar	0.505	0.433	0.606
29	SINAI	0.425	0.312	0.667
30	McRock	0.423	0.301	0.708
31	Baseline	0.369	0.477	0.300
32	WhatsaMeme	0.347	0.347	0.346
33	IIMAS1UTM1LaSalle	0.199	0.755	0.115
	English - Subtask 2a			
1	HierarchyEverywhere	0.746	0.867	0.655
2	NLPNCHU	0.707	0.782	0.645
3	BCAmirs	0.705	0.784	0.641
4	UMUTeam	0.690	0.768	0.627
5	SuteAlbastre	0.685	0.718	0.655
6	TUMnlp	0.677	0.781	0.598
7	Pauk	0.675	0.745	0.617
8	CodeMeme	0.666	0.607	0.739
9	LomonosovMSU	0.656	0.792	0.560
10	IITK	0.636	0.763	0.545
11	BERTastic	0.613	0.816	0.491
12	BDA	0.504	0.515	0.493
13	Baseline	0.447	0.688	0.331
14	WhatsaMeme	0.366	0.313	0.440

Table 5: Official results for English - Subtasks 1 and 2a. Runs ranked by the official measure (Hierarchical F1).

Rank	Team	F1 macro	F1 micro
1	LMEME	0.810	0.825
2	SuteAlbastre	0.810	0.835
3	DUTIR938	0.809	0.837
4	BCAmirs	0.803	0.825
5	Snarci	0.799	0.827
6	BDA	0.793	0.823
7	NLPNCHU	0.788	0.822
8	UMUTeam	0.787	0.807
9	TUMnlp	0.784	0.802
10	CodeMeme	0.782	0.807
11	LomonosovMSU	0.772	0.798
12	BERTastic	0.716	0.762
13	Hidetsune	0.714	0.790
14	Scalar	0.702	0.753
15	SheffieldVeraAI	0.642	0.687
16	HierarchyEverywhere	0.563	0.662
17	WhatsaMeme	0.515	0.530
18	nowhash	0.498	0.515
19	IITK	0.483	0.490
20	Baseline	0.250	0.333

Table 6: Official results for English - Subtask 2b. Runs ranked by the official measure (Hierarchical F1).

with the meme text, before being passed to a RoBERTa model. A vision encoder utilizing a pre-trained vision transformer model (CLIP-ViT), is used to encode and analyze the meme image. Finally, a multi-layer perceptron classifier takes the combined visual and textual representations and classifies the meme. The models used were monolingual, and thus, for non-English tasks, the system was applied to test sets translated using Google Translate.

In Subtask 2b, the top system is **LMEME** (Li et al., 2024b), which was also the top system for Subtask 2b in English, and presented in Section 5.1.

5.3 North Macedonian Subtasks

The results for the North Macedonian subtasks are presented in Tables 9 and 10. Our observations for the North Macedonian subtasks closely align with those for the Bulgarian subtasks. For subtasks 1 and 2a, a few teams were unable to surpass the baseline results. However, for Subtask 2b, all teams exceeded the baseline performance.

As with the Bulgarian Subtask 1 and Subtask 2a, the top systems for North Macedonian were also **OtterlyObsessedWithSemantics** (Wunderle et al., 2024) and **BCAmirs**, respectively.

Team **BERTastic** (Mahmoud and Nakov, 2024) achieved the best performance for Subtask 2b. The system uses three representations of the input meme, including the image, associated text, and a generic description of the meme generated

R	Team	hF1	hP	hR
	Bulgarian - Subtask 1			
1	OtterlyObsessedWithSemantics	0.568	0.520	0.627
2	RDproj	0.541	0.435	0.714
3	NLPNCHU	0.517	0.536	0.500
4	MagnumJUCSE	0.500	0.470	0.533
5	nowhash	0.486	0.460	0.516
6	MemeSifters	0.481	0.491	0.472
7	GreyBox	0.476	0.438	0.522
8	whatdoyoumeme	0.473	0.502	0.446
9	HierarchyEverywhere	0.468	0.483	0.453
10	fralak	0.464	0.374	0.613
11	914isthebest	0.463	0.477	0.450
12	CLaC	0.449	0.400	0.512
13	BCAmirs	0.448	0.387	0.533
14	IITK	0.434	0.404	0.470
15	ShefCDTeam	0.366	0.454	0.307
16	EURECOM	0.345	0.367	0.325
17	SINAI	0.341	0.214	0.849
18	Baseline	0.284	0.319	0.256
19	SuteAlbastre	0.236	0.134	1.000
20	IIMAS1UTM1LaSalle	0.183	0.654	0.107
	Bulgarian - Subtask 2a			
1	BCAmirs	0.627	0.703	0.566
2	SuteAlbastre	0.611	0.660	0.569
3	NLPNCHU	0.549	0.707	0.448
4	BERTastic	0.544	0.812	0.409
5	Baseline	0.500	0.804	0.363
6	BDA	0.483	0.523	0.450
7	HierarchyEverywhere	0.464	0.671	0.355
8	IITK	0.446	0.541	0.379

Table 7: Bulgarian - Subtasks 1 and 2a

	Bulgarian - Subtask 2b		
Rank	**Team**	**F1 macro**	**F1 micro**
1	LMEME	0.671	0.810
2	Snarci	0.668	0.840
3	BERTastic	0.662	0.750
4	BCAmirs	0.647	0.770
5	NLPNCHU	0.647	0.820
6	MemeSifters	0.611	0.830
7	SuteAlbastre	0.594	0.650
8	SheffieldVeraAI	0.536	0.570
9	BDA	0.506	0.620
10	HierarchyEverywhere	0.485	0.630
11	IITK	0.473	0.530
12	DUTIR938	0.434	0.570
13	nowhash	0.434	0.450
14	Hidetsune	0.327	0.330
15	Baseline	0.167	0.200

Table 8: Bulgarian - Subtask 2b

R	Team	hF1	hP	hR
	North Macedonian - Subtask 1			
1	OtterlyObsessedWithSemantics	0.512	0.518	0.507
2	RDproj	0.499	0.434	0.587
3	MagnumJUCSE	0.483	0.486	0.480
4	fralak	0.464	0.359	0.658
5	NLPNCHU	0.462	0.546	0.400
6	EURECOM	0.442	0.520	0.384
7	MemeSifters	0.441	0.539	0.373
8	GreyBox	0.434	0.440	0.429
9	nowhash	0.426	0.414	0.438
10	HierarchyEverywhere	0.417	0.486	0.365
11	CLaC	0.395	0.371	0.422
12	BCAmirs	0.393	0.332	0.482
13	IITK	0.383	0.344	0.432
14	914isthebest	0.369	0.401	0.341
15	whatdoyoumeme	0.362	0.399	0.331
16	ShefCDTeam	0.319	0.436	0.251
17	Baseline	0.307	0.314	0.300
18	SINAI	0.301	0.183	0.846
19	SuteAlbastre	0.204	0.113	0.996
20	IIMAS1UTM1LaSalle	0.137	0.529	0.079
	North Macedonian - Subtask 2a			
1	BCAmirs	0.637	0.750	0.553
2	SuteAlbastre	0.576	0.492	0.692
3	BERTastic	0.573	0.866	0.428
4	Baseline	0.555	0.902	0.401
5	BDA	0.501	0.546	0.463
6	NLPNCHU	0.487	0.706	0.372
7	IITK	0.440	0.545	0.369
8	HierarchyEverywhere	0.357	0.689	0.241

Table 9: North Macedonian - Subtasks 1 and 2a

by a vision-language model. A multilingual model, MPNet, was used to extract embeddings from text elements, while a multimodal multilingual model, CLIP-ViT-B-32, was used to represent both text and image. All extracted features were fused into a single feature vector, followed by logistic regression for classification.

5.4 Arabic Subtasks

In Tables 11 and 12, we report the results for Arabic subtasks. Here, we also observe similar patterns to Bulgarian and North Macedonian. The top systems for Subtask 1 and Subtask 2a are also **OtterlyObsessedWithSemantics** (Wunderle et al., 2024) and **BCAmirs**, respectively. **BCAmirs** also achieves the top performance for Subtask 2b.

Considering all non-English languages, we see that many systems struggled to surpass the baseline for Subtask 2a specifically. This can be due to the difficult nature of this subtask, as it is a hierarchical multilabel classification task that also requires considering multimodal content. Such difficulty also affected the number of participants, with a relatively smaller number of systems submitted to this

	North Macedonian - Subtask 2b		
Rank	**Team**	**F1 macro**	**F1 micro**
1	BERTastic	0.686	0.840
2	MemeSifters	0.660	0.900
3	LMEME	0.591	0.780
4	BCAmirs	0.561	0.770
5	NLPNCHU	0.520	0.790
6	HierarchyEverywhere	0.506	0.620
7	IITK	0.485	0.630
8	Snarci	0.479	0.720
9	DUTIR938	0.469	0.660
10	SheffieldVeraAI	0.458	0.510
11	BDA	0.435	0.600
12	nowhash	0.429	0.520
13	Hidetsune	0.389	0.460
14	SuteAlbastre	0.177	0.180
15	Baseline	0.091	0.100

Table 10: North Macedonian - Subtask 2b

R	Team	hF1	hP	hR
	Arabic - Subtask 1			
1	OtterlyObsessedWithSemantics	0.476	0.391	0.607
2	NLPNCHU	0.475	0.428	0.533
3	fralak	0.428	0.309	0.698
4	whatdoyoumeme	0.424	0.328	0.600
5	RDproj	0.411	0.333	0.537
6	IITK	0.408	0.339	0.512
7	HierarchyEverywhere	0.405	0.356	0.470
8	nowhash	0.404	0.360	0.460
9	BCAmirs	0.396	0.320	0.519
10	MagnumJUCSE	0.395	0.346	0.460
11	CLaC	0.381	0.308	0.498
12	MemeSifters	0.360	0.355	0.365
13	914isthebest	0.360	0.314	0.421
14	Baseline	0.359	0.350	0.368
15	SINAI	0.258	0.154	0.793
16	SuteAlbastre	0.234	0.198	0.288
17	EURECOM	0.177	0.343	0.119
	Arabic - Subtask 2a			
1	BCAmirs	0.526	0.553	0.502
2	SuteAlbastre	0.516	0.469	0.573
3	Baseline	0.486	0.650	0.389
4	NLPNCHU	0.483	0.595	0.407
5	IITK	0.455	0.457	0.453
6	HierarchyEverywhere	0.437	0.510	0.382
7	BDA	0.416	0.382	0.457
8	BERTastic	0.388	0.613	0.284

Table 11: Arabic - Subtasks 1 and 2a

Arabic - Subtask 2b			
Rank	**Team**	**F1 macro**	**F1 micro**
1	BCAmirs	0.615	0.631
2	SheffieldVeraAI	0.610	0.613
3	BERTastic	0.603	0.606
4	NLPNCHU	0.585	0.594
5	HierarchyEverywhere	0.562	0.669
6	MemeSifters	0.557	0.694
7	Snarci	0.555	0.556
8	Hidetsune	0.528	0.544
9	BDA	0.510	0.606
10	SuteAlbastre	0.501	0.544
11	nowhash	0.498	0.531
12	DUTIR938	0.469	0.519
13	IITK	0.467	0.469
14	LMEME	0.362	0.388
15	Baseline	0.227	0.294

Table 12: Arabic - Subtask 2b

subtask compared to the other two subtasks.

6 Conclusions and Future Work

We presented SemEval-2024 Task 4 on *Multilingual Detection of Persuasion Techniques in Memes*. The task consists of detecting persuasion techniques in memes in a multimodal setting. The task offered a significantly larger dataset for English (10K memes) than previous ones, and three surprise languages: Arabic, Bulgarian, and North Macedonian.

The task attracted a lot of attention: 153 teams registered for the task and 30 teams submitted a task description paper. Fine-tuning transformer-based architectures was the most dominant approach followed by most teams. The majority of teams participating in Subtask 2 considered both the text and image components of the data, utilizing corresponding transformer models. Finally, several teams designed hierarchical classification techniques, to tackle the hierarchy of labels in Subtask 1 and Subtask 2a. As for the surprise languages, at least a third of the submitting teams used automatic translation to translate the datasets into English.

7 Limitations

The dataset we have collected originates from various public Facebook groups, with a primary focus on politics. Consequently, the representativeness of this dataset may be limited for other domains and topics. The highly imbalanced distribution of the labels in the dataset may affect the model's performance. Therefore, it is important to develop models with this aspect in mind.

Ethics and Broader Impact

Our dataset solely comprises memes, and we have not collected any user information; therefore, the privacy risk is nonexistent.

Any biases present in the dataset are unintentional, and our intention is not to cause harm to any group or individual. It's important to acknowledge that annotating propaganda techniques involves a degree of subjectivity, making biases in our gold-labeled data or label distribution unavoidable. To mitigate these concerns, we have collected examples from a diverse range of users and groups. Furthermore, we adhere to a well-defined schema with clear definitions, which has enabled us to achieve high inter-annotator agreement. Additionally, our annotation team was diverse, consisting of six members, including both females and males.

We advise researchers of the risk that our dataset could be exploited to biasly moderate memes, potentially due to biases related to demographics or specifics in the text. To prevent this, the implementation of human moderation is crucial.

References

Amirhossein Abaskohi, Amirhossein Dabiriaghdam, Lele Wang, and Giuseppe Carenini. 2024. Bcamirs at semeval-2024 task 4: Beyond words: A multimodal and multilingual exploration of persuasion in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1402–1412, Mexico City, Mexico. Association for Computational Linguistics.

Shamsiah Abd Kadir, Anitawati Lokman, and T. Tsuchiya. 2016. Emotion and techniques of propaganda in YouTube videos. *Indian Journal of Science and Technology*, Vol (9).

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022a. A survey on multimodal disinformation detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6625–6643, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Firoj Alam, Hamdy Mubarak, Wajdi Zaghouani, Giovanni Da San Martino, and Preslav Nakov. 2022b. Overview of the WANLP 2022 shared task on propaganda detection in Arabic. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop*, Abu Dhabi, UAE.

Ion Anghelina, Gabriel Buță, and Alexandru Enache. 2024. Sutealbastre at semeval-2024 task 4: Predicting propaganda techniques in multilingual memes using joint text and vision transformers. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 430–436, Mexico City, Mexico. Association for Computational Linguistics.

Alberto Barrón-Cedeno, Israa Jaradat, Giovanni Da San Martino, and Preslav Nakov. 2019. Proppy: Organizing the news based on their propagandistic content. *Information Processing & Management*, 56(5):1849–1864.

Sian Brooke. 2019. "condescending, rude, assholes": Framing gender and hostility on stack overflow. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 172–180, Florence, Italy. Association for Computational Linguistics.

Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2022. Prompting for multimodal hateful meme classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 321–332, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nishan Chatterjee, Marko Pranjic, Boshko Koloski, Lidia Pivovarova, and Senja Pollak. 2024. whatdoyoumeme at semeval-2024 task 4: Hierarchical-label aware cross-lingual persuasion detection using translated texts. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1548–1554, Mexico City, Mexico. Association for Computational Linguistics.

Shreenaga Chikoti, Shrey Mehta, and Ashutosh Modi. 2024. Iitk at semeval-2024 task 4: Hierarchical embeddings for detection of persuasion techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1790–1798, Mexico City, Mexico. Association for Computational Linguistics.

Abu Nowhash Chowdhury and Michal Ptaszynski. 2024. nowhash at semeval-2024 task 4: Exploiting fusion of transformers for detecting persuasion techniques in multilingual memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 133–138, Mexico City, Mexico. 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 International Workshop on Semantic Evaluation*, SemEval '20, Barcelona, Spain.

Giovanni Da San Martino, Alberto Barrón-Cedeño, and Preslav Nakov. 2019. Findings of the NLP4IF-2019 shared task on fine-grained propaganda detection. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, NLP4IF '19, pages 162–170, Hong Kong, China.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019a. Fine-grained analysis of propaganda in news articles. In *EMNLP*.

Giovanni Da San Martino, Seunghak Yu, Alberto Barron-Cedeno, Rostislav Petrov, and Preslav Nakov. 2019b. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, EMNLP '19, pages 5636–5646, Hong Kong, China.

Jiaxu Dao, Zhuoying Li, Youbang Su, and Wensheng Gong. 2024. Puer at semeval-2024 task 4: Fine-tuning pre-trained language models for meme persuasion technique detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 64–69, Mexico City, Mexico. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11 of *AAAI '17*.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021a. Detecting propaganda techniques in memes. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, ACL-IJCNLP '21, pages 6603–6617.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021b. Task 6 at SemEval-2021: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th International Workshop on Semantic Evaluation*, SemEval '21, Bangkok, Thailand.

Shreyansh Gandhi, Samrat Kokkula, Abon Chaudhuri, Alessandro Magnani, Theban Stanley, Behzad Ahmadi, Venkatesh Kandaswamy, Omer Ovenc, and Shie Mannor. 2020. Scalable detection of offensive and non-compliant content / logo in product images. *WACV*, pages 2236–2245.

Omid Ghahroodi and Ehsaneddin Asgari. 2024. Hierarchyeverywhere at semeval-2024 task 4: Detection of persuasion techniques in memes using hierarchical text classifier. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1738–1743, Mexico City, Mexico. Association for Computational Linguistics.

Meredith Gibbons, Maggie Mi, Xingyi Song, and Aline Villavicencio. 2024. Shefcdteam at semeval-2024 task 4: A text-to-text model for multi-label classification. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1872–1879, Mexico City, Mexico. Association for Computational Linguistics.

Maria Glenski, E. Ayton, J. Mendoza, and Svitlana Volkova. 2019. Multilingual multimodal digital deception detection and disinformation spread across social platforms. *ArXiv*, abs/1909.05838.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *WACV*, pages 1470–1478.

Charlie Grimshaw, Kalina Bontcheva, and Xingyi Song. 2024. Sheffieldveraai at semeval-2024 task 4: Prompting and fine-tuning a large vision-language model for binary classification of persuasion techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2035–2040, Mexico City, Mexico. Association for Computational Linguistics.

Shih-Wei Guo and Yao-Chung Fan. 2024. Nlpnchu at semeval-2024 task 4: A comparison of mdhc strategy and in-domain pre-training for multilingual detection of persuasion techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1880–1887, Mexico City, Mexico. Association for Computational Linguistics.

Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. 2017. Argotario: Computational argumentation meets serious games. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, EMNLP '17, pages 7–12, Copenhagen, Denmark.

Ivan Habernal, Patrick Pauli, and Iryna Gurevych. 2018. Adapting serious game for fallacious argumentation to German: Pitfalls, insights, and best practices. In *LREC*. European Language Resources Association (ELRA).

Maram Hasanain, Firoj Alam, Hamdy Mubarak, Samir Abdaljalil, Wajdi Zaghouani, Preslav Nakov, Giovanni Da San Martino, and Abed Freihat. 2023. ArAIEval shared task: Persuasion techniques and disinformation detection in Arabic text. In *Proceedings of ArabicNLP 2023*, pages 483–493, Singapore (Hybrid). Association for Computational Linguistics.

Ming Shan Hee, Shivam Sharma, Rui Cao, Palash Nandi, Preslav Nakov, Tanmoy Chakraborty, and Roy Ka-Wei Lee. 2024. Recent advances in hate speech moderation: Multimodality and the role of large models. *arXiv preprint arXiv:2401.16727*.

Srecko Joksimovic, Ryan S. Baker, Jaclyn Ocumpaugh, Juan Miguel L. Andres, Ivan Tot, Elle Yuan Wang, and Shane Dawson. 2019. Automated identification of verbally abusive behaviors in online discussions. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 36–45, Florence, Italy. Association for Computational Linguistics.

Adnan Khurshid and Dipankar Das. 2024. Magnum jucse at semeval-2024 task 4: Multilingual detection of persuasion techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1004–1007, Mexico City, Mexico. Association for Computational Linguistics.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. In *Proceedings of the NeurIPS 2019 Workshop on Visually Grounded Interaction and Language*, ViGIL@NeurIPS '19.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, NeurIPS '20.

Svetlana Kiritchenko, Richard Nock, and Fazel Famili. 2006. Learning and evaluation in the presence of class hierarchies: Application to text categorization. volume 4013, pages 395–406.

Katarina Laken. 2024. Fralak at semeval-2024 task 4: combining rnn-generated hierarchy paths with simple neural nets for hierarchical multilabel text classification in a multilingual zero-shot setting. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 583–588, Mexico City, Mexico. Association for Computational Linguistics.

Dailin Li, Chuhan Wang, Xin Zou, Junlong Wang, Peng Chen, Jian Wang, Liang Yang, and Hongfei Lin. 2024a. 914isthebest at semeval-2024 task 4: Cot-based data augmentation strategy for persuasion techniques detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation*, SemEval 2024, Mexico City, Mexico.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Shiyi Li, Yike Wang, Liang Yang, Shaowu Zhang, and Hongfei Lin. 2024b. Lmeme at semeval-2024 task 4: Teacher student fusion - integrating clip with llms for enhanced persuasion detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 615–620, Mexico City, Mexico. Association for Computational Linguistics.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the Conference on Neural Information Processing Systems*, NeurIPS '19, Vancouver, Canada.

Tarek Mahmoud and Preslav Nakov. 2024. Bertastic at semeval-2024 task 4: State-of-the-art multilingual propaganda detection in memes via zero-shot learning with vision-language models. In *Proceedings of* *the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 490–497, Mexico City, Mexico. Association for Computational Linguistics.

Kota Shamanth Ramanath Nayak and Leila Kosseim. 2024. Clac at semeval-2024 task 4: Decoding persuasion in memes – an ensemble of language models with paraphrase augmentation. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 175–180, Mexico City, Mexico. Association for Computational Linguistics.

Mohammad Osoolian, Erfan Moosavi Monazzah, and Sauleh Eetemadi. 2024. Iustnlplab at semeval-2024 task 4: Multilingual detection of persuasion techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1081–1085, Mexico City, Mexico. Association for Computational Linguistics.

ronghao pan, José Antonio García-Díaz, and Rafael Valencia-García. 2024. Umuteam at semeval-2024 task 4: Multimodal identification of persuasive techniques in memes through large language models. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 642–652, Mexico City, Mexico. Association for Computational Linguistics.

Matt Pauk and Maria Leonor Pacheco. 2024. Pauk at semeval-2024 task 4: A neuro-symbolic method for consistent classification of propaganda techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1413–1423, Mexico City, Mexico. Association for Computational Linguistics.

Youri Peskine, Raphael Troncy, and Paolo Papotti. 2024. Eurecom at semeval-2024 task 4: Hierarchical loss and model ensembling in detecting persuasion techniques. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1166–1171, Mexico City, Mexico. Association for Computational Linguistics.

Jakub Piskorski, Nicolas Stefanovitch, Valerie-Anne Bausier, Nicolo Faggiani, Jens Linge, Sopho Kharazi, Nikolaos Nikolaidis, Giulia Teodori, Bertrand De Longueville, Brian Doherty, Jason Gonin, Camelia Ignat, Bonka Kotseva, Eleonora Mantica, Lorena Marcaletti, Enrico Rossi, Alessio Spadaro, Marco Verile, Giovanni Da San Martino, Firoj Alam, and Preslav Nakov. 2023a. News categorization, framing and persuasion techniques: Annotation guidelines. Technical Report JRC-132862, European Commission Joint Research Centre, Ispra (Italy).

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

Jakub Piskorski, Nicolas Stefanovitch, Nikolaos Nikolaidis, Giovanni Da San Martino, and Preslav Nakov. 2023c. Multilingual multifaceted understanding of online news in terms of genre, framing, and persuasion techniques. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3001–3022, Toronto, Canada. Association for Computational Linguistics.

Nirmalendu Prakash, Han Wang, Nguyen Khoi Hoang, Ming Shan Hee, and Roy Ka-Wei Lee. 2023. PromptMTopic: Unsupervised multimodal topic modeling of memes using large language models. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 621–631, New York, NY, USA. Association for Computing Machinery.

Nathan Roll and Calbert Graham. 2024. Greybox at semeval-2024 task 4: Progressive fine-tuning (for multilingual detection of propaganda techniques). In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 875–880, Mexico City, Mexico. Association for Computational Linguistics.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Hyunjin Seo. 2014. Visual propaganda in the age of social media: An empirical analysis of Twitter images during the 2012 Israeli–Hamas conflict. *Visual Communication Quarterly*, 21(3).

Shivam Sharma, Firoj Alam, Md. Shad Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, and Tanmoy Chakraborty. 2022. Detecting and understanding harmful memes: A survey. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, IJCAI '22, pages 5597–5606, Vienna, Austria. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Victoria Sherratt, Sedat Dogan, Ifeoluwa Wuraola, Lydia Bryan-Smith, Oyinkansola Onwuchekwa, and Nina Dethlefs. 2024. Bda at semeval-2024 task 4: Detection of persuasion in memes across languages with ensemble learning and external knowledge. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 123–132, Mexico City, Mexico. Association for Computational Linguistics.

Limor Shifman. 2013. *Memes in digital culture*. MIT press.

Marco Siino. 2024. Mcrock at semeval-2024 task 4: Mistral 7b for multilingual detection of persuasion techniques in memes. In *Proceedings of the*

18th International Workshop on Semantic Evaluation (SemEval-2024), pages 53–59, Mexico City, Mexico. Association for Computational Linguistics.

Gleb Skiba, Mikhail Pukemo, Dmitry Melikhov, and Konstantin Vorontsov. 2024. Lomonosovmsu at semeval-2024 task 4: Comparing llms and embedder models to identifying propaganda techniques in the content of memes in english for subtasks №1, №2a, and №2b. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1555–1559, Mexico City, Mexico. Association for Computational Linguistics.

Zhivar Sourati, Vishnu Priya Prasanna Venkatesh, Darshan Deshpande, Himanshu Rawlani, Filip Ilievski, Hông-Ân Sandlin, and Alain Mermoud. 2023. Robust and explainable identification of logical fallacies in natural language arguments. *Know.-Based Syst.*, 266(C).

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *TRAC*, pages 32–41.

Hidetsune Takahashi. 2024. Hidetsune at semeval-2024 task 4: An application of machine learning to multilingual propagandistic memes identification using machine translation. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 363–366, Mexico City, Mexico. Association for Computational Linguistics.

Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Veronique Hoste. 2015. Detection and fine-grained classification of cyberbullying events. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 672–680, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Arthur Vasconcelos, Luiz Felipe de Melo, Eduardo Goncalves, Eduardo Bezerra, Aline Paes, and Alexandre Plastino. 2024. Bambas at semeval-2024 task 4: How far can we get without looking at hierarchies? In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 442–449, Mexico City, Mexico. Association for Computational Linguistics.

Svitlana Volkova, Ellyn Ayton, Dustin L. Arendt, Zhuanyi Huang, and Brian Hutchinson. 2019. Explaining multimodal deceptive news prediction models. In *Proceedings of the International Conference on Web and Social Media*, ICWSM '19, Munich, Germany.

Zihan Wang, Peiyi Wang, Tianyu Liu, Binghuai Lin, Yunbo Cao, Zhifang Sui, and Houfeng Wang. 2022. 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.

Ching Seh Wu and Unnathi Bhandary. 2020. Detection of hate speech in videos using machine learning. In *CSCI*, pages 585–590.

Julia Wunderle, Julian Schubert, Antonella Cacciatore, Albin Zehe, Jan Pfister, and Andreas Hotho. 2024. Otterlyobsessedwithsemantics at semeval-2024 task 4: Developing a hierarchical multi-label classification head for large language models. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 589–599, Mexico City, Mexico. Association for Computational Linguistics.

Erchen Yu, Junlong Wang, Xuening Qiao, Jiewei Qi, Zhaoqing Li, Hongfei Lin, Linlin Zong, and Bo Xu. 2024. Dutir938 at semeval-2024 task 4: Semi-supervised learning and model ensemble for persuasion techniques detection in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 629–635, Mexico City, Mexico. Association for Computational Linguistics.

Luca Zedda, Alessandra Perniciano, Andrea Loddo, Cecilia Di Ruberto, Manuela Sanguinetti, and Maurizio Atzori. 2024. Snarci at semeval-2024 task 4: Themis model for binary classification of memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 840–845, Mexico City, Mexico. Association for Computational Linguistics.

Yuhang Zhu. 2024. Rdproj at semeval-2024 task 4: An ensemble learning approach for multilingual detection of persuasion techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 181–187, Mexico City, Mexico. Association for Computational Linguistics.

A Additional Dataset Details

Figure 2 shows statistics about the distribution of the number of persuasion techniques per meme for Subtasks 1 and 2a. The techniques hierarchy in 3 shows the details of coarse and fine-grained categories.

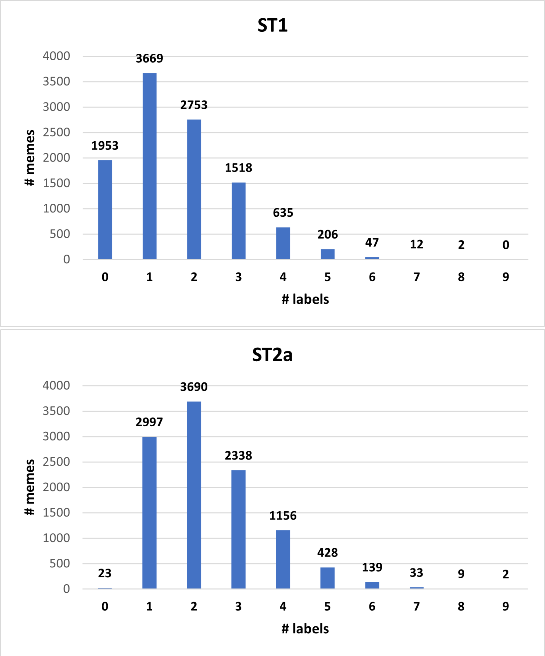

Figure 2: Subtasks 1 and 2a number of labels distributions.

B Overview of Participating Systems

In this section, we provide a summary of the approach followed by each of the participating systems.

BDA (Sherratt et al., 2024) The team participated in both Subtask 2a and Subtask 2b. For Subtask 2a, the proposed architecture is an ensemble of models operating on two modalities, text and images. For text, an ensemble of mBERT and XLM-RoBERTa is used, while CLIP and a monolingual BERT model is used to process visual entities extracted from images using Google Vision. Finally, a late fusion engine is used to merge predictions; generate additional translated task data; and modify the prediction confidence threshold based on the task hierarchy. As for Subtask 2b, the system is an ensemble of three models: 1) XLM-RoBERTa, that is trained on augmented task data, 2) VGG19 trained on task images and 3) a BERT model trained on visual entities extracted from the images using Google Vision. Late fusion is applied to join predictions from the models.

OtterlyObsessedWithSemantics (Wunderle et al., 2024) For Subtask 1, a custom classification head that is designed to be applied atop of a large language model was used. This approach includes reconstructing the hierarchy across multiple fully connected layers, allowing for incorporation of previous foundational decisions in subsequent, more fine-grained layers. For the non-English tasks, the same system was used after translating all documents to English.

BAMBAS (Vasconcelos et al., 2024) The proposed system for Subtask 1 does not consider the hierarchy of labels. First, text embeddings are extracted leveraging a multilingual tweets-based language model, Bernice. Next, those embeddings are used to train a separate binary classifier for each label, in a binary-relevance style, adopting independent oversampling strategies in each model.

nowhash (Chowdhury and Ptaszynski, 2024) In their submission to Subtask 1, the team starts from meme texts as input to the system and fine-tunes a Language-agnostic BERT sentence embedding (LaBSE) model on top of Flair's Transformer Document Embeddings. Further, those document vectors are then fed to a single-layer feed-forward linear classifier to obtain the prediction label.

For Subtask 2b, the proposed system operates on both meme images and texts. The architecture includes a vision transformer and XLMRoBERTa to extract effective contextual information from both modalities. Finally, the features are fused, to be passed to a single feed-forward linear layer. The architecture is fine-tuned given the task training data.

RDproj (Zhu, 2024) In their participation in Subtask 1, the team built an ensemble learning system employing a soft voting strategy. Propaganda techniques were grouped into ten subsets based on their representation in the training subset. Subsequently, one classifier including XLM-RoBerta$_{large}$ with a classification head is trained on each of these training sets. Finally, a classifier with the same architecture is used to learn a weighted average of the label's probability generated by the other classifiers.

BERTastic (Mahmoud and Nakov, 2024) For Subtask 2, the proposed architecture covers three

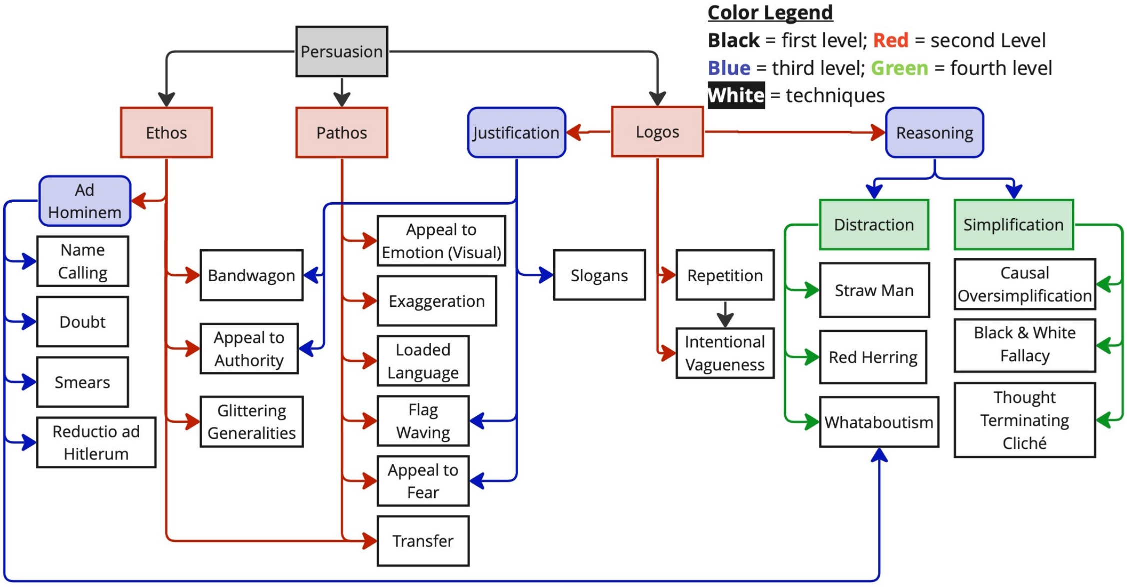

Figure 3: Persuasion techniques hierarchy.

representations of the input meme, including the image, associated text, and a generic description of the meme generated by a vision-language model. A multilingual model, MPNet, is used to extract embeddings from text elements, while a multimodal multilingual model, CLIP-ViT-B-32, is used to represent both text and image. All extracted features are fused into a single feature vector, followed by logistic regression for classification. During training, models' weights were frozen.

GreyBox (Roll and Graham, 2024) For Subtask 1, GPT 3.5 Turbo was fine-tuned in multiple stages using the training and validation datasets from all subtasks. Then, zero-shot prompting was used to generate predictions. The team also experimented with the original GPT 3.5 Turbo, Llama 2 70B Chat model, and Mistral AI's Mixtral 8x7B instruct, mixture of experts model.

SuteAlbastre (Anghelina et al., 2024) In submitting to Subtask 1, a BERT model was fine-tuned on the provided data. As for Subtask 2a: The backbone of the solutions is a BERT + ViT architecture where the BERT-based model creates embeddings from the text data while the ViT creates features from the image data. The two embeddings are concatenated and the resulting one is passed to a fully connected layer to obtain the scores for each persuasion technique. The same architecture was used for Subtask 2b, except the output of the final fully connected layer was adjusted for binary classification on whether the provided meme is propagandistic or non-propagandistic.

Pauk (Pauk and Pacheco, 2024) For Subtask 1, a student-teacher knowledge distillation approach was implemented. DeBERTa was adopted as the student model, in addition to a softened logic rule layer on top with a collection of logic rules that encode the hierarchical relationship between possible output labels. The student model then learns by both emulating the gold labels as well as the teacher's predictions that respect the hierarchy. The same knowledge distillation approach was used for Subtask 2a. However, the student model consists of DeBERTa for processing the textual content and ResNet for processing the image content with output embeddings concatenated and fed into a feed-forward network for predictions.

DUTIR938 (Yu et al., 2024) For Subtask 2b, the team developed a dual-channel model based on semi-supervised learning and model ensemble. Within the image channel, CLIP was used to extract image features from memes. Concurrently, in the text channel, diverse pre-trained language models were utilized. A concatenation and fusion process of the extracted features was applied and the resulting features were subsequently fed into a classification layer. Lastly, a two-stage soft-voting ensemble strategy was used to amalgamate the predictions of multiple models.

CLaC (Nayak and Kosseim, 2024) Similar to several other systems submitted to Subtask 1, the proposed approach was based on fine-tuning individual language models (BERT, XLM-RoBERTa, and mBERT) and leveraging a mean-based ensemble model. Additionally, the training dataset was augmented by a relevant dataset extracted from a previous SemEval task

EURECOM (Peskine et al., 2024) The proposed system for Subtask 1 uses an ensemble of multiple models trained with different parameters. Experiments were conducted with different models (BERT, RoBERTa, DeBERTa, DistillBERT, AlBERT), different training datasets (SemEval 2024, + 2021, + PTC), different loss functions (BCE, CE, Focal, Hierarchical) and data augmentation (back translation, GPT-4-turbo augmented). The best results were obtained by leveraging the hierarchical nature of the data, by outputting ancestor classes and with a hierarchical loss. The official submission was based on the majority voting of our top-3 models for each persuasion technique.

Fralak (Laken, 2024) Different from transformer-based approaches presented so far, the system developed for Subtask 1 involved training an RNN. It was based on restructuring the labels into strings that showed the full path through the label hierarchy, and training a basic RNN that generated these strings based on the multilingual sentence embedding of the meme text. This RNN module was then incorporated into an ensemble model with 2 more models consisting of basic fully connected networks.

Snarci (Zedda et al., 2024) The system submitted to Subtask 2b involved a modular architecture that combines image and language embedding models. As image encoders, several versions of CLIP were used. Similarly, to process the textual part of memes the system resorts to several pre-trained language models (specifically TinyLlama, phi-1.5, and phi-2). The embeddings extracted from the CLIP model undergo an image embedding projection to fit a compatible size for large language models. An optional Token Merger module, inspired by the Patch Merger module proposed in vision transformers, merges tokens from image and text embeddings to focus on relevant meme aspects. This module aims to aggregate similar tokens together, regardless of their original position. To make the system more computationally efficient, freezing techniques were used to maintain the pre-trained weights of both image and language embeddings, and then Low-Rank Adaptation techniques were leveraged to fine-tune the models' weights.

SheffieldVeraAI (Grimshaw et al., 2024) For Subtask 2b, the team approached the problem by prompting and fine-tuning the large vision-language model, LLaVa. Fine-tuning was done using the multi-modal training data through LoRA training technique, however, this did not improve the model's performance. We achieved the best results prompting the baseline LLaVa model. We adapted the model to the unseen languages, by using a machine translation model, NLLB. We translated the meme transcriptions into English and used this translated text prompt with the original meme.

ShefCDTeam (Gibbons et al., 2024) The team participated in subtask 1, exploring sequence-to-sequence modeling for this task using a Flan-T5 model with sequential parameter efficient fine-tuning methods - Low-Rank Adaptation and prompt tuning.

whatdoyoumeme (Chatterjee et al., 2024) Subtask 1 was approached by fine-tuning a transformer model. The hierarchical labels for the task were integrated into the system by extending the training labels to include all ancestors. Experiments were conducted using several models like DistilBERT and mBERT but the best results were achieved with mBART. The model employed the standard classification architecture (mBART+classification head) and was trained using a BCE loss. When running the system over the non-English test sets, the documents were translated to English using the NLLB-200 model.

LomonosovMSU (Skiba et al., 2024) Two approaches were used to solve Subtask 1. 1) A generative approach involving training a generative model to generate explicit responses to questions. 2) A BERT-like approach involving training a simple fully connected network on top of a frozen pre-trained embedding model to solve the hierarchical classification task. Subtask 2 was tackled similarly to Subtask 1, but using multimodal text-to-image embedding models.

HierarchyEverywhere (Ghahroodi and Asgari, 2024) In Subtask 1, a state-of-the-art hierarchical text classification model called HPT was used. This required representing the propaganda techniques hierarchically as a directed acyclic graph. Two supplementary datasets were also added to the training. In Subtask 2a and Subtask 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, Google Translation API was used to translate them into English.

914isthebest (Li et al., 2024a) The team developed a transfer-based model for Subtask 1. For system training, the training dataset was augmented following a Chain-of-Thought-based data augmentation approach using GPT-3.5. The main classification architecture includes four RoBERTa models and one DeBERTa model initialized using different random seeds. A soft voting approach, which averages the predicted probabilities of each label from all five models, is used to predict labels. To predict non-English languages, the testing sets were translated using GPT-3.5.

LMEME (Li et al., 2024b) In Subtask 2b, the team proposed a detection system that employs a Teacher Student Fusion framework. Initially, a large language model serves as the teacher, engaging in abductive reasoning on multimodal inputs to generate background knowledge on persuasion techniques, assisting in the training of a smaller downstream model. The student model adopts CLIP as an encoder for text and image features, incorporating an attention mechanism for modality alignment.

McRock (Siino, 2024) The team approached Subtask 1 by prompting an instruction-tuned large language model called Mistral-7B-Instruct-v0.2. The prompt used included both the definitions of all 20 techniques targeted by the subtask, a short instruction on the task to perform, and the sample to predict on. The post-processed model's outputs were then submitted to the task's leaderboard.

BCAmirs (Abaskohi et al., 2024) The team participated in all subtasks but mainly focused on Subtask 2a. GPT-4 was used to generate a descriptive caption of the meme. The caption is then combined with the meme text before being passed to a RoBERTa model. A vision encoder utilizing a pre-trained vision transformer model (CLIP-ViT), is used to encode and analyze meme images. Finally, a multi-layer perceptron classifier takes the combined visual and textual representations and classifies the meme. The RoBERTa and MLP classifiers are fine-tuned, while CLIP remains frozen.

They conducted a series of experiments exploring different methods of combining the textual and visual data: *text-only* (Vicuna-1.5, BERT, RoBERTa), *image-only* (LLaVa without textual input), *text + image* (VisualBERT, ConcatRoBERTa, LLaVa-1.5), *text + caption + image* (LLaVa-1.5, Vicuna-1.5, VisualBERT, ConcatRoBERTa). Experiments were conducted using LLaVa and GPT-4 generated captions with GPT-4 captions showing consistently better results.

Puer (Dao et al., 2024) The team participated in Subtask 1 on the English test data with a detection system based on RoBERTa, using Roberta-large, which was fine-tuned on a corpus of social media posts. They conducted extensive parameter tuning over the dev set to identify an optimal threshold, epoch, etc. Finally, They compare the performances of other different deep learning model architectures, such as BERT, ALBERT, and XLM-RoBERTa, on multilingual detection of persuasion techniques in memes.

Hidetsune (Takahashi, 2024) The team approached Subtask 2b with a text-only classical NLP solution using SpacyV3 textcat_multilabel classification architecture. The model was trained on the official dataset for Subtask 2b, combined with additional data from Kaggle consisting of non-propagandistic tweets. The team participated in all languages included in Subtask2b by translating non-English text into English and applying the same model for text classification.

UMUTeam (pan et al., 2024) The team participated in all subtasks of the competition focusing only on English data. In Subtask 1 the team fine-tuned the RoBERTa-large model using an epoch-based evaluation strategy. In Subtasks 2a and 2b, they again used RoBERTa-large as their classification model but trained it by combining the textual content of a meme with image descriptions extracted using LlaVa.

MagnumJUCSE (Khurshid and Das, 2024) The team participated in Subtask 1 in all languages. They participated in the subtask with a node-level hierarchical classification system consisting of four phases: data denoising, feature generation, node-level classifier training, and finally inference. They first clean the data, then generate features using pre-trained sentence transformers, afterwards they predict whether an example belongs to a given node or not using SVM (Support Vector Machine). Finally, inference is done in a top-down fashion by selecting the most suitable depth for the prediction results, based on the decision probabilities of the classifier at each node.

IUSTNLPLAB (Osoolian et al., 2024) The team addresses Subtask 1 on the English dataset. Their study focused on fine-tuning language models using the training dataset, including BERT, GPT-2, and RoBERTa, with GPT-2 showing the best performance for the task. Additionally, they used data on persuasion techniques from Semeval 2023 Task 3 increasing the training data with 3,445 new samples, however, this approach did not yield discernable improvements. Finally, the participants adjusted the prediction threshold which lead to a noticeable improvement in model performance.

IITK (Chikoti et al., 2024) The team participated in all three Subtasks in every language. Subtask 1: they presented an approach to meme classification based on HypEmo (pre-trained hyperbolic embeddings) and emotion prediction through a multi-task learning framework, incorporating auxiliary tasks, including masked language modeling (MLM) and class definition prediction to enhance the understanding of emotional concepts. The predictions from HypEmo and the Fine-grained class-definition-based model are merged for the final prediction. Subtask 2a: the team experiments with an ensemble of HypEmo and the class definition-based multi-task learning model for the textual content of the meme and using the CLIP model embeddings from the visual content of the meme. Subtask 2b: the team uses a fusion approach, concatenation pre-trained BERT-base model for textual features and CNN model for visual features. They use weighted binary cross entropy as a loss function due to the dataset imbalance.

NLPNCHU (Guo and Fan, 2024) The team participated in all three Subtasks in every language. They explored various finetuning techniques and classification strategies, such as data augmentation, problem transformation, and hierarchical multi-label classification strategies. In Subtasks 1 and 2a, they explored different classification strategies: Global Classifier (GC), Stacking + GC, and Stacking + Local Classifier per Level (LCL), combined with Distribution-Balanced Loss (DBL) loss to address the long-tail distribution of the data. For Subtask 1 the team compared the performance of XLM-RoBERTa and XLM-RoBERTa-Twitter to asses the impact of domain-specific pre-training. For Subtask 2a the team used XLM-RoBERTa and XLM-RoBERTa-Twitter for extracting textual features and CLIP for extracting visual features combining them through Feature-wise Linear Modulation (FIM), these two encoders encode to obtain a representation embedding vector containing both image and text For Subtask 2b the team employed the same strategy as Subtask 2a applied to a binary classification setting.