BDA at SemEval-2024 Task 4: Detection of Persuasion in Memes Across Languages with Ensemble Learning and External Knowledge

Victoria Sherratt, Sedat Dogan, Ifeoluwa Wuraola,
Lydia Bryan-Smith, Oyinkansola Onwuchekwa and **Nina Dethlefs**
University of Hull
Big Data Analytics Research Group
v.sherratt-2020@hull.ac.uk

Abstract

This paper outlines our multimodal ensemble learning system for identifying persuasion techniques in memes. We contribute an approach which uses the novel inclusion of consistent named visual entities extracted using Google Vision API's as an external knowledge source, joined to our multimodal ensemble via late fusion. As well as detailing our experiments in ensemble combinations, fusion methods and data augmentation, we explore the impact of including external data and summarise post-evaluation improvements to our architecture based on analysis of the task results.

1 Introduction

In this paper, we describe our approach to identifying persuasion techniques for SemEval 2024 Task 4. The task involves the identification of up to 22 persuasion techniques within memes, which are inherently multimodal. We participated in Subtask2a and Subtask2b of the task.

Subtask2a is a multilabel classification task, requiring the identification of 22 persuasion techniques using both the textual and visual content. The subtask was evaluated by a hierarchical F1, as each label is part of a subset of techniques and contains a parent node. Subtask2b is a binary classification task, determining the presence or absence of any persuasion technique within a meme. For both subtasks, training data is provided in the English language and a development set also in English. As well as English, 3 surprise languages in Arabic, North Macedonian and Bulgarian were provided to officially evaluate our approach (Dimitrov et al., 2024).

Our system architecture is an amalgamation of traditional NLP and vision models, exploring late and early fusion techniques as well as carefully crafted confidence thresholds. We extend beyond the training data by incorporating resources such as Google Vision[1], which would provide consistent named visual entities extracted from the image regardless of language; in a multilingual context this reduces reliance on sentence spans or tokens, which can be problematic due to linguistic variations in unseen language data. We also make our code publicly available.[2]

2 Background

Identifying persuasion techniques in memes is necessary endeavour for combating misinformation and fostering critical media consumption among the public, and the focus of a number of ongoing research areas for the prevention of harmful content, propaganda or disinformation spread through memes (Dimitrov et al., 2021a; Dupuis and Williams, 2019; Sharma et al., 2022).

Propaganda is generally referred to as information which is purposefully shaped or presented to support a particular agenda, often utilising persuasion techniques in the shared task. Previous shared tasks have also considered the identification of persuasion techniques in text only (Da San Martino et al., 2020), multimodal contexts using memes (Dimitrov et al., 2021b), and persuasion techniques in multilingual text (Piskorski et al., 2023b). SemEval 2024 Task 4 is a shared task of a similar nature, however the task considers both image and text as well as multilingual test data.

As meaning is often generating through the interaction of both modalities in memes, meme related tasks are typically approached using pretrained convolutional neural networks (Beskow et al., 2020; Hossain et al., 2022; Sherratt et al., 2023; Suryawanshi et al., 2020) or vision transformers (Afridi et al., 2021; Cao et al., 2023) in combination with language models. Our ensemble approach therefore explores successful CNNs for

[1]https://cloud.google.com/vision/docs/detecting-web
[2]https://github.com/vemchance/BDA-SemEval4

the binary classification task; for the more complex multilabel classification, we explore CLIP (Radford et al., 2021) to leverage its significant pretraining on large-scale natural language descriptions and images, as well as its notable performance in zero-shot classification and related downstream multimodal tasks such as social media sentiment analysis (Bryan-Smith et al., 2023).

Our motivation for including external knowledge sources is inspired by previous successful applications of external information (Zhu, 2020) and ongoing research to improve meme-related tasks with the addition of structured knowledge external to the meme itself (Sherratt, 2022; Tommasini et al., 2023).

3 Exploratory Analysis

Before implementing our approach, we explore the task data provided. Exploring Subtask2a, we calculated TF-IDF vectors for texts within each label and calculated the cosine similarity between vectors. We noted that, for the majority of labels, there is significant crossover in textual content. We also examine the number of labels in a single meme, as Subtask2a was a multilabel classification problem where each meme could have more than one persuasion technique, in Figure 1.

Given this crossover, we initially explored leveraging the annotation guidelines provided for the task, which provides concrete examples of how to label each persuasion technique. We noted the annotation guidelines primarily provided examples annotation based on the location of nouns or adjectives per technique, but provided few examples of non-European languages aside from Russian. However, the guidelines did note the presence of 'personal characteristics, organisations, political orientation or opinions' in some techniques (Piskorski et al., 2023a).

As the test data would be provided on unseen languages, we explore a more concise representation of these attributes using the Google Vision API to extracted 'web entities' and visual concepts from an image. For multilingual data, this allows us to rely less on sentence spans or tokens - elements that vary across language - and instead leverage visual entities that could consistently represent information for each label regardless of textual content. In Table 1, we outline an sample of extracted entities from Google Vision's web entities search.

Technique	Entity	Occurrence Count
Appeal to (Strong) Emotions	Russia	48
Appeal to (Strong) Emotions	United States	35
Appeal to (Strong) Emotions	Amnesty International	34
Doubt	Brand	52
Doubt	Politics	48
Doubt	Public Relations	40
Doubt	Speech	39
Red Herring	Entrepreneur	8
Red Herring	Business	7
Red Herring	Ukraine	7
Red Herring	Russia	7

Table 1: Example Entities Extracted via Google Vision

4 System Overview

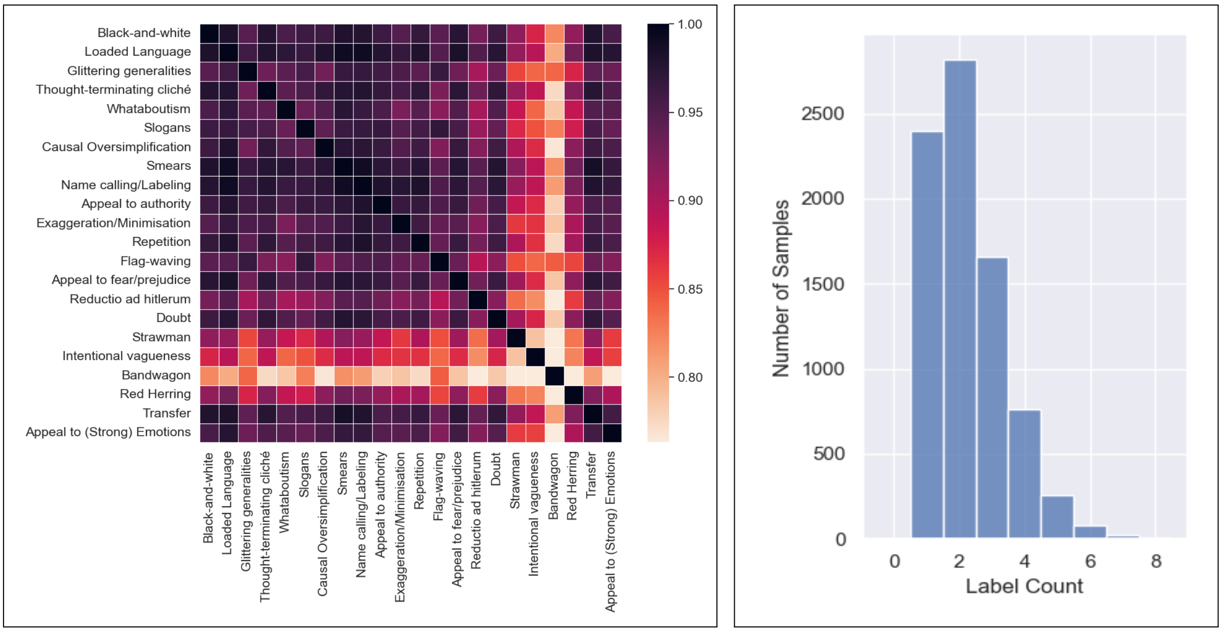
Our main system approach includes ensembling NLP models with vision models for both subtasks. We experimented with BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) family models as well as VGG19 (Simonyan and Zisserman, 2014), ResNet50 (He et al., 2015) and CLIP (Radford et al., 2021).

For Subtask2a, we initially design an architecture that combines multilingual text processing with visual analysis. Our vision stream also includes web entities from Google Vision, processed by a single BERT model. Our Subtask2b system similarly integrated visual and textual modalities with experiments in late and early fusion. We also include additional novel implementations beyond an ensemble of pretrained models:

External Knowledge: We use Google Vision to extract information from meme image. The Google Vision API annotates an image using web detection, returning the a list of predicted labels for objects, people or concepts in an image, as well as matching URLs and the Google Knowledge Graph ID (Singhal, 2012). We utilise only the named visual entities, with an example in Section 3.

Data Augmentation: We experiment with augmenting the task data. English training data is direct translated using GPT-3.5 (Brown et al., 2020) into a number of other languages, and then again translated for model training when the test datasets are released.

F1 Confidence Threshold: For Subtask2a, we leverage the provided hierarchy of techniques (Dimitrov et al., 2024) to change the confidence threshold for our model's predicted labels. The F1 Confidence Threshold reduces both the threshold required to classify a label from 0.50 to 0.40 and a confidence between 0.35 and 0.40 will return the parent node of the label. We detail the impact of the F1 Confidence Threshold in Section 5.2.



(a) TF-IDF Cosine Similarity in Label Groups

(b) Count of Labels Per Meme in Subtask2a

Figure 1: Multilabel Classification Label Exploration

Late Fusion Engine: We implement a late fusion system to combine our separate NLP and vision streams together into a single predictive value. We calculated the per-label accuracy for each model, and used this to weight the contribution of each. In other words:

$$predict_{label} = \frac{(A_{label} \times accA_{label}) + (B_{label} \times accB_{label})}{accA_{label} + accB_{label}}$$

where $accA_{label} \in \{0..1\}$ and $accB_{label} \in \{0..1\}$ refers to the accuracy for the respective models for a given label.

5 Experimental Setup

We combine the training and validation sets for Subtask2a and Subtask2b to train each architecture, a total of 7,500 for Subtask2a and 1,350 for Subtask2b originally in English. We test our approach on the Development Set in English (1,000 samples for Subtask2a and 300 for Subtask2b). Detailed in section 5.1, the total samples are increased by direct translating data for both subtasks.

For all experiments, we set the validation split in the model to 30% of the total training data. When multiple languages are included in the data, we stratify the training and test splits based on language.

The number of epochs is determined by no improvement to validation loss after 5 epochs. We find that the majority of the language models in combination complete around 8 - 10 epochs, whereas CLIP often stops improving around 6 epochs. Table 2 details the specific parameters of our main models. We use pretrained models for both image and text modalities, and therefore the drop-out rate is applied before the respective classification layer detailed in Figure 2.

	mBERT	XLM-RBase	BERT	CLIP
Optimizer	AdamW	AdamW	AdamW	Adam
Dropout	0.4	0.4	0.3	0.5
Weight Decay	1e-5	1e-5	-	-
Learning Rate	1e-5	1e-5	1e-5	5e-5
Batch Size	8	8	8	16

Table 2: Model Parameters

5.1 Additional Data

We explore the use of the Persuasion Techniques Corpus (PTC) (Da San Martino et al., 2020) as additional training data. We use the Google Vision API to extract descriptive entities for all task data images, which is returned in English from the API under the ‘web entities’ search response. We also augment our dataset using GPT-3.5 (Brown et al., 2020) to direct translate a sample of 500 texts from Subtask2a for each unseen language in the task (1,500 additional samples, or 20% of the available training data). We perform the same process for Subtask2b. Notably we do not augment or change

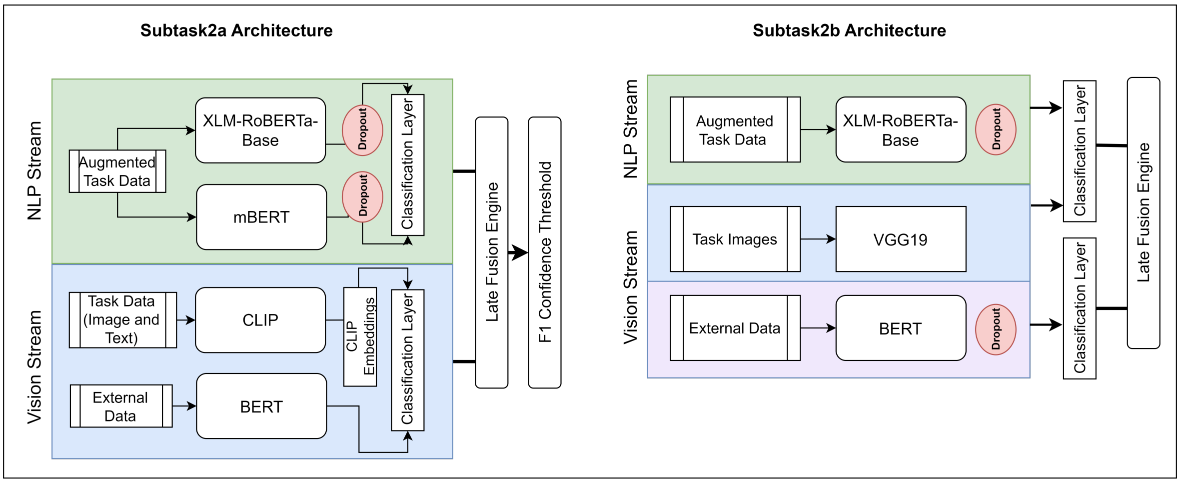


Figure 2: Subtask2a and Subtask2b Architecture

the image for this additional data.

In our results detailed in Section 6, we refer to the Persuasion Techniques Corpus as *PTC*, the original task data as *TD*, the task data with added samples as *ATD* (augmented task data) and data extracted via Google Vision as *ED* (External Data). When external data is used as input, this is followed by (ex) (e.g., BERT(ex)) in Section 6.

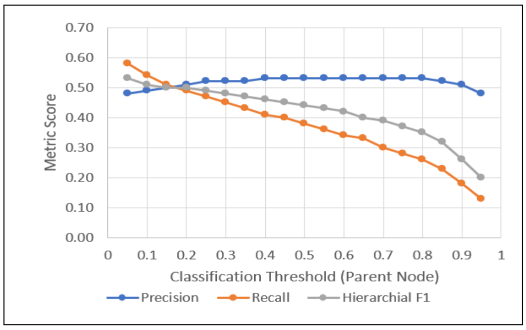


Figure 3: F1 Score Against Parent Node Threshold

5.2 Subtask2a Details

For Subtask2a, we experimented with a number of individual models and ensemble models as detailed in Section 6 as well as different fusion strategies and the inclusion of the F1 Confidence Threshold. In early fusion, models are jointly trained and their learned feature vectors concatenated before passed through final classification layer. In late fusion, we use the late fusion engine detailed in Section 5 on the predicted probabilities of each model. Primarily, we experiment with late fusion when incorporating external knowledge.

The original architecture is detailed in Figure 2. However, as we experimented with a number of model combinations, input data and fusion techniques, we opted to choose the model which performed the best on the English development data for the official submission.

As detailed in Table 3 in Section 6, our original architecture was less effective than other experiments. We therefore utilised a modified version of the architecture, adding an additional mBERT model with a high drop-out rate to combat overfitting in the first instance, which marginally increase performance (the *Triad* model in results). In our final submitted architecture we remove CLIP, so only the BERT model with external data as input remains in the vision stream, and use late fusion to merge this with the Triad NLP architecture.

Detailed in Figure 3, we examine the impact of changing the required confidence threshold for a label, testing a single mBERT model from our ensemble. The F1 Confidence Threshold reduces the threshold required predict a technique, and then introduces another lower threshold to predict the technique label's parent node from the task hierarchy (Dimitrov et al., 2024). We opted to use a configuration which balances the Hierarchical F1, Precision and Recall detailed in Figure 3. In the F1 Hierarchy Threshold, the parent node prediction is always 0.05 less than the label confidence threshold. The configuration used is 0.40 for the label threshold, and 0.35 to return the parent node of the label.

5.3 Subtask2b Details

For Subtask2b, if a model is reused from Subtask2a (e.g., Bert(ex) models to process external data) we reuse the parameters described above. For the vi-

Model	Fusion	Finetune Data	H. F1	Precision	Recall
XLM-RBase	-	PTC	0.213	0.362	0.151
XLM-RBase	-	PTC, ATD	0.387	0.516	0.310
XLM-RBase	-	ATD	0.404	0.521	0.330
mBERT	-	PTC	0.213	0.362	0.151
mBERT	-	PTC, ATD	0.163	0.512	0.097
mBERT	-	ATD	0.463	0.523	0.416
BERT(ex)	-	ED	0.395	0.528	0.316
BERT(ex)^{F1}	-	ED	0.424	0.477	0.382
CLIP	-	TD	0.315	0.375	0.272
CLIP^{F1}	-	TD	0.405	0.413	0.398
mBERT + XLM-RBase	Early	ATD	0.451	0.514	0.402
mBERT + XLM-RBase ^{F1}	Early	ATD	0.480	0.471	0.490
mBERT + XLM-RBase + BERT(ex)^{F1}	Early	ATD, ED	0.475	0.466	0.484
CLIP + BERT(ex)	Early	ATD, ED	0.342	0.374	0.316
CLIP + BERT(ex)	Late	ATD, ED	0.345	0.523	0.257
CLIP + BERT(ex)^{F1}	Early	ATD, ED	0.457	0.420	**0.501**
CLIP + BERT(ex)^{F1}	Late	ATD, ED	0.435	0.488	0.392
Triad	Early	ATD	0.470	0.515	0.433
Triad + BERT(ex)	Early	ATD, ED	0.473	0.467	0.480
Triad + BERT(ex)	Late	ATD, ED	0.476	0.470	0.484
Triad + BERT(ex)^{F1}	Late	ATD, ED	**0.483**	0.526	0.446
Triad + BERT(ex) + CLIP	Late	TD, ATD, ED	0.463	**0.541**	0.405
Triad + BERT(ex) + CLIP^{F1}	Late	TD, ATD, ED	0.455	0.461	0.450

Table 3: Subtask2a Experiment Results on Development Set (English)

sion models, we use a different learning rate for ResNet50 and VGG19 with the AdamW optimizer of 1e-8, a batch size of 8 and the same early stopping parameters as Subtask2a.

Both image models utilise ImageNet weights (Deng et al., 2009). We apply the same dropout rate specified above to the text model before this is passed through a classification layer in the case of early fusion techniques. As Subtask2b is a binary classification task, we do not require the F1 Confidence Threshold technique for this architecture. In our final architecture, VGG19 and XLM-RoBERTa-Base are trained jointly on the augmented task data, and the late fusion engine combines predictions from from the Google Vision web entities.

6 Results

We detail the results of our experiments for Subtask2a in Table 3 and Subtask2b in Table 4. In the Table 3, the F1 Confidence Threshold modification is indicated by [Model] *F1*.

For Subtask2a, we found the Triad combination performed best with BERT (trained on the extracted Google Vision entities) predictions combined with late fusion. In all cases, the F1 Hierarchy threshold increased the score of the same model.

Whilst we explored the use of PTC to finetune our models, we found that, due to the different naming conventions of some techniques, performance did not improve with incorporation of the PTC data. We also noted the PTC data was drawn from a different domain (e.g., news articles) were the context of techniques would be longer than short sentences in memes, and potentially this corpus was less effective as a finetuning dataset for the task.

We originally aimed to leverage CLIP's text and image embeddings to inform a novel early fusion neural network model for multilabel multiclass persuasion techniques classification. However, this architecture including CLIP was slightly less effective than others. The reasons behind this suboptimal performance could be multifaceted, including the complexity and subtlety of propagandistic content within memes, the inherent challenges of cross-modal understanding in this particular domain. One reason is suggested that, whilst the vi-

Model	Fusion	Data	F1 Macro	F1 Micro
BERT(ex)	-	ED	0.577	0.580
CLIP	-	TD	0.618	0.680
CLIP + BERT(ex)	Late	TD, ED	0.634	0.707
Triad	Early	ATD	0.383	0.613
VGG19 + BERT	Early	ATD	**0.753**	**0.806**
VGG19 + mBERT	Early	ATD	0.621	0.740
ResNet50 + mBERT	Early	ATD	0.638	0.700
VGG19 + XLM-RBase	Early	ATD	0.641	0.706
ResNet50 + XLM-RBase	Early	ATD	0.618	0.706
VGG19 + XLM-RBase + BERT(ex)	Early	ATD, ED	0.337	0.360
VGG19 + XLM-RBase + BERT(ex)	Late	ATD, ED	*0.677*	0.717
VGG19 + XLM-RBase + CLIP + BERT(ex)	Late	TD, ATD, ED	0.602	0.707

Table 4: Subtask2b Experiment Results on Development Set (English)

sual modality is important for identifying whether a technique is present, *distinguishing* between the specific types of techniques may primarily be a linguistic task.

For Subtask2b, our architecture achieved overall better scores than Subtask2a. We tested architectures retrained for a binary classification task from Subtask2a on Subtask2b as a comparison, noting these models did not perform as well. In Subtask2b, therefore, the vision modality was significant in the binary classification task. We note from the results monolingual language models outperform multilingual models, and suggest this may be due to the limited sample size for the augmented data in Subtask2b. In line with our system strategy, we include BERT(ex) only in conjunction with multilingual models, as the aim of this additional data is to improve zero-shot classification irrespective of language. We observed significant performance increase using the BERT(ex) model in late fusion for Subtask2b.

6.1 Test Set Performance and Analysis

For the test set, we submitted the best performing model from each subtask experiment. For Subtask2a, this was the Triad + BERT(ex) with late fusion. For Subtask2b, we submitted the VGG19 + BERT model for English test sets and the VGG19 + XLM-RoBERTa-Base + BERT(ex) for all other languages.

Evaluating our results on the test set in Table 5, we found that our model for Subtask2a generalised better on different languages, outperforming the results on the English Development dataset in some cases. Our system performed the best on North Macedonian and the worst in Arabic for this task. The original and augmented task data for Subtask2a was larger than Subtask2b, and we effectively traded English language performance for better generalisability on other languages.

	Rank	F1	F1 < Leader
Subtask2a			
English	12	0.504	-0.242
Bulgarian	6	0.483	-0.144
North Macedonian	**5**	**0.514**	**-0.135**
Arabic	8	0.409	-0.153
Subtask2b			
English	**6**	**0.793**	**-0.017**
Bulgarian	9	0.506	-0.165
North Macedonian	11	0.435	-0.251
Arabic	9	0.510	-0.105

Table 5: Results on Official Test Set Leaderboard

For Subtask2b, our architecture under-performed from tests on the English Development dataset aside from the VGG19+BERT model used in the English test set. This approach was less able to generalise on non-English data than our approach from Subtask2a, with a significant score reduction in North Macedonian, our highest scoring language for Subtask2a.

Whilst the performance drop could equally be attributed to a smaller augmented data sample in Subtask2b, we also examine North Macedonian memes to understand the reduction of performance on this set. Visually, North Macedonian memes were different from memes in other languages, particularly in English; they included a significant number of 'cartoon' type memes and comic strips compared to

Language	Entity	Occurrence Count
English	Politics	68
English	United States	62
English	US President	38
Bulgarian	Product	24
Bulgarian	Bulgaria	17
Bulgarian	Public Relations	14
North Macedonian	Cartoon	78
North Macedonian	Public Relations	38
North Macedonian	Poster	28
Arabic	Product	29
Arabic	Humor	12
Arabic	Laughter	11

Table 6: Sample Web Entities for Test Dataset in Subtask2b

others, which is also reflected in a sample of visual entities outlined in Table 6. As our Subtask2b architecture relied more on the visual modality than Subtask2a, the reduction of performance is therefore expected given this analysis.

6.2 Post-Evaluation Analysis

Post official evaluation, we took the best elements from each subtask and to explore an approved architecture for each task. Whilst these are *not* part of the official SemEval Task 4 leaderboard, we include these as additional experiments.

For Subtask2a, we direct translated a further 500 texts from Subtask1 (a subtask we did not compete in which uses 20 of the techniques, but considers only the text in memes) into Arabic, Bulgarian and North Macedonian (1,500 additional samples). For Subtask2b, we direct translated 200 memes per test language from the Memotion (Sharma et al., 2020) dataset which were considered 'not offensive' and labelled these non-propagandistic, to significantly increase and re-balance the data provided for Subtask2b. In this new augmented data, each test language comprised 10% of the non-propagandistic label whereas English comprised 70%, also drawing memes from Memotion in English to balance the label sample size.

In terms of architecture, for Subtask2a we incorporated the VGG19 model instead of CLIP and removed the second mBERT model with the 80% drop-out rate with the aim to provide more information from the visual modality. For Subtask2b, we attempted to improve the linguistic part of the model by incorporating XLM-Roberta-Large instead.

Despite incorporating the visual modality and additional data, our second attempt at Subtask2a

Subtask2a	Test Language	F1	F1 Change
mBERT+XLM-RBase + VGG19	Bulgarian	0.424	-0.06
mBERT+XLM-RBase + VGG19	North Macedonian	0.358	-0.06
mBERT+XLM-RBase + VGG19	Arabic	0.366	-0.12
Subtask2b			
XLM-RL + VGG19	Bulgarian	0.571	0.065
XLM-RL + VGG19	North Macedonian	0.570	0.135
XLM-RL + VGG19	Arabic	**0.621**	0.111
XLM-RL + VGG19 + BERT(ex)	Bulgarian	0.571	0.065
XLM-RL + VGG19 + BERT(ex)	North Macedonian	0.578	**0.143**
XLM-RL + VGG19 + BERT(ex)	Arabic	0.603	0.093

Table 7: Post-Evaluation Model Results

under-performed. Considering the drop, we did not feel the inclusion of external knowledge via an additional BERT model as in prior experiments would improve performance. As with our initial results, we suggest that the multilabel classification task requires greater attention given to the text in memes.

In Subtask2b, all languages improved without BERT(ex). Performance on Arabic decreased slightly with the inclusion of external knowledge, with no change in Bulgarian and an increase in North Macedonian. The inclusion of external knowledge via late fusion, comparative to the results in Table 4, provided marginal improvement; likely the dataset re-balance and inclusion of a larger language model were also significant in this improvement. The augmented data for this experiment were also more diverse in this case as they were drawn from a different dataset, whereas augmenting the multilabel classes in Subtask2a from another dataset was not possible without native language speakers trained in the specific annotation task.

7 Conclusion and Future Work

We presented our ensemble learning approach to SemEval-2024 Task 4, including a number of experiments with early and late fusion, the inclusion of external knowledge and modifying the label threshold. We found that the inclusion of external sources of knowledge, even basic descriptive entities as in our experiments, improved performance on both subtasks especially using late fusion.

The identification of named entities in visual modality of memes is a potential future area of research, as this would enable drawing on complex stores of knowledge and contribute to deeper cross-modal understanding to disentangle persuasion techniques. We further suggest that there promise in generating more high quality, multilingual data for persuasion techniques across languages based on our experiments with augmented data, particularly for low-resource languages.

We also note there is a cultural element to memes not considered in current research. We identified that North Macedonian memes were visually different from other memes; the different cultural perspectives and practices in developing consuming memes is under-researched, with only limited studies investigating global meme practices (Nissenbaum and Shifman, 2018). As well as additional varied training data, a better understanding of cultural meme production could contribute to defining the most appropriate approach for zero-shot multilingual meme tasks.

Acknowledgements

We acknowledge the Viper High Performance Computing facility of the University of Hull and its support team.

References

Tariq Habib Afridi, Aftab Alam, Muhammad Numan Khan, Jawad Khan, and Young-Koo Lee. 2021. A multimodal memes classification: A survey and open research issues. In *Innovations in Smart Cities Applications Volume 4: The Proceedings of the 5th International Conference on Smart City Applications*, pages 1451–1466. Springer.

David M Beskow, Sumeet Kumar, and Kathleen M Carley. 2020. The evolution of political memes: Detecting and characterizing internet memes with multi-modal deep learning. *Information Processing & Management*, 57(2):102170.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Lydia Bryan-Smith, Jake Godsall, Franky George, Kelly Egode, Nina Dethlefs, and Dan Parsons. 2023. Real-time social media sentiment analysis for rapid impact assessment of floods. *Computers Geosciences*, 178:105405.

Rui Cao, Ming Shan Hee, Adriel Kuek, Wen-Haw Chong, Roy Ka-Wei Lee, and Jing Jiang. 2023. Pro-cap: Leveraging a frozen vision-language model for hateful meme detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5244–5252.

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

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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. 2021a. Detecting propaganda techniques in memes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6603–6617, Online. Association for Computational Linguistics.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021b. 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.

Marc J Dupuis and Andrew Williams. 2019. The spread of disinformation on the web: An examination of memes on social networking. In *2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 1412–1418. IEEE.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. MemoSen: A multimodal dataset for sentiment analysis of memes. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1542–1554, Marseille, France. European Language Resources Association.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Asaf Nissenbaum and Limor Shifman. 2018. Meme templates as expressive repertoires in a globalizing world: A cross-linguistic study. *Journal of Computer-Mediated Communication*, 23(5):294–310.

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.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020.

Chhavi Sharma, Deepesh Bhageria, William Paka, Scott, Srinivas P Y K L, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. SemEval-2020 Task 8: Memotion Analysis-The Visuo-Lingual Metaphor! In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain. Association for Computational Linguistics.

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. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Victoria Sherratt. 2022. Towards contextually sensitive analysis of memes: Meme genealogy and knowledge base. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5871–5872. International Joint Conferences on Artificial Intelligence Organization. Doctoral Consortium.

Victoria Sherratt, Kevin Pimbblet, and Nina Dethlefs. 2023. Multi-channel convolutional neural network for precise meme classification. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, pages 190–198.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Amit Singhal. 2012. Introducing the knowledge graph: Things, not strings.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41, Marseille, France. European Language Resources Association (ELRA).

Riccardo Tommasini, Filip Ilievski, and Thilini Wijesiriwardene. 2023. Imkg: The internet meme knowledge graph. In *The Semantic Web: 20th International Conference, ESWC 2023, Hersonissos, Crete, Greece, May 28–June 1, 2023, Proceedings*, page 354–371, Berlin, Heidelberg. Springer-Verlag.

Ron Zhu. 2020. Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution. *CoRR*, abs/2012.08290.