SheffieldVeraAI at SemEval-2024 Task 4: Prompting and fine-tuning a Large Vision-Language Model for Binary Classification of Persuasion Techniques in Memes

Charlie Grimshaw, Kalina Bontcheva and **Xingyi Song**
Department of Computer Science
University of Sheffield
Sheffield, UK
{cgrimshaw1, k.bontcheva, x.song}@sheffield.ac.uk

Abstract

This paper describes our approach for SemEval-2024 Task 4: Multilingual Detection of Persuasion Techniques in Memes. Specifically, we concentrate on Subtask 2b, a binary classification challenge that entails categorizing memes as either "propagandistic" or "non-propagandistic". To address this task, we utilized the large multimodal pretrained model, LLaVa. We explored various prompting strategies and fine-tuning methods, and observed that the model, when not fine-tuned but provided with a few-shot learning examples, achieved the best performance. Additionally, we enhanced the model's multilingual capabilities by integrating a machine translation model. Our system secured the 2nd place in the Arabic language category.

1 Introduction

Research of online misinformation is growing (Chaudhari and Pawar, 2021) as fake news and propagandistic content spreads further and replaces more real news on social media, detrimentally impacting society, including loss of lives, loss of health and economic loss (Muhammed T and Mathew, 2022). A common online propaganda format is a meme, where text and image(s) are combined to share a message, often political (Guo et al., 2020). This paper describes SheffieldVeraAI's approach for SemEval 2024 Task 4 Subtask 2b, involving detecting the presence of persuasion technique(s) within memes, a binary visual/textual classification task (Dimitrov et al., 2024). The previous research on this task including (Feng et al., 2021) (Tian et al., 2021) (Li et al., 2021) used non-autoregressive encoder representation techniques, using models such as BERT (Devlin et al., 2018) or RoBERTa (Liu et al., 2019), fused with a vision representation model such as ResNet (He et al., 2015).

Unlike the previous research, we experimented with a different approach to use and train an auto-regressive vision-language model that receives image and text as input, outputting text only. Prompting the model with a meme and expecting the model to generate a classification output. Specifically, we use the LLaVa-1.5 model (Liu et al., 2023a), which directly projects an image encoding into tokens computed as text tokens by an LLM. This technique allows us to utilise the LLMs' "knowledge" of persuasion techniques they have learnt through massive pre-training and explained outputs through prompting, improving the model's interoperability and error analysis.

1.1 Contributions

- Show that a pre-trained auto-regressive large visual language model can be prompted for binary persuasion classification.
- Show that prompting with translated text is a viable method, achieving 2nd place in the Arabic leaderboard, using an English-only model.

2 Background

Previous SemEval tasks have looked at this problem of online misinformation/persuasion/propaganda:

- **SemEval 2020 Task 11 - "Detection of Propaganda Techniques in News Articles" (Martino et al., 2020).** This task involved span and technique investigation on text-only news articles.
- **SemEval 2021 Task 6 - "Detection of Persuasion Techniques in Texts and Images" (Dimitrov et al., 2021).** The first task relating to persuasion techniques involved a subtask with images, which required classifying propaganda techniques within memes.

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 2051–2056
June 20-21, 2024

- **SemEval 2023 Task 3 - "Detecting the Genre, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup" (Piskorski et al., 2023).** This task is similar to SemEval 2020 Task 11, which contains no visual content or news articles while adding genre and framing detection.

This theme of persuasion technique detection is prominent in recent years of SemEval, with the most similar task being SemEval 2021 Task 6 Subtask 3, which involved visual and textual persuasion techniques in memes.

2.1 Task Description

In this work, our focus is on Subtask 2b, which aims to determine whether at least one persuasion technique is present in the meme or no technique is present. This task provides both the original image and the text transcriptions. The detailed data structure is outlined as follows:

- unique *id* of the sample. e.g. 12345
- The image of the meme, an example can be found in Figure 1
- A transcription of text within the image content of the meme. For example: “GIVE A THUMBS UP IF YOU\\nSTILL SUPPORT TRUMP\\n”
- A label which is either **propagandisitic** or **non-propagandistic**. A meme is propagandistic if it contains one or more of the 22 persuasion techniques defined by the task organisers.

The language of the meme and transcription is either English, Bulgarian, North Macedonian or Arabic. The language of the meme and transcription always match.

3 System Overview

Our system follows these steps:

1. Fine-tune LLaVa with LoRA (Hu et al., 2021) using pre-processed English training data. (Optional; our final system is untrained).
2. Translate Bulgarian, North Macedonian and Arabic transcriptions to English using NLLB (Team et al., 2022).

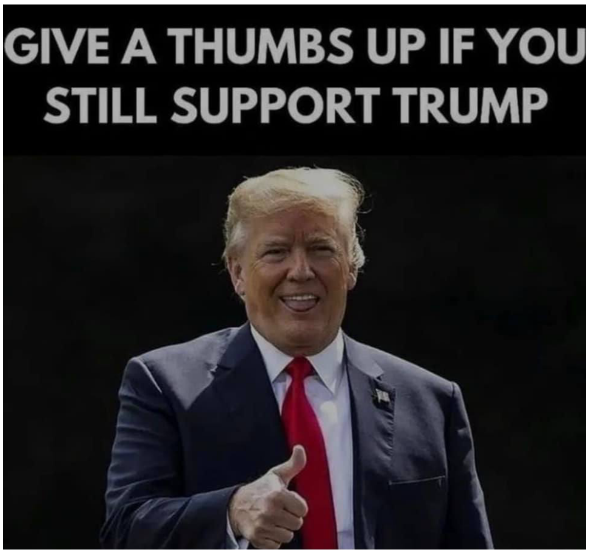

Figure 1: Example of a propagandistic image from the task

3. Prompt LLaVa for binary classification of persuasion techniques, giving a few-shot example.

3.1 LLaVa

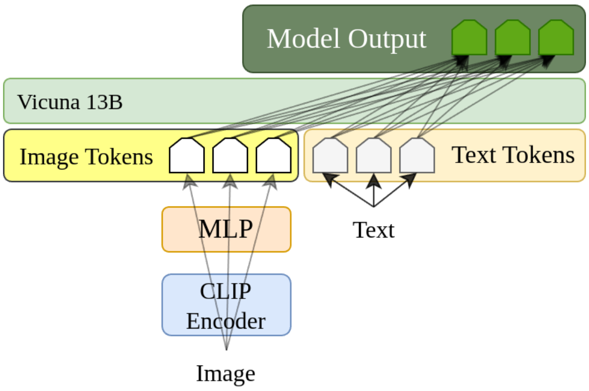

Figure 2: Diagram of LLaVa 1.5 architecture, modelled from original paper (Liu et al., 2023b).

The model we use for this task is called LLaVa (Large Language and Vision Assistant) (Liu et al., 2023b) (Liu et al., 2023a). We are using the 13B parameter version of LLaVa-1.5. We use the original author's public code, available on GitHub[1], for training and inference. LLaVa is an English-only end-to-end fine-tuned Large Vision-Language Model (LVLM) trained on Chat-GPT4 (OpenAI, 2023) generated instruction-following data. It handles image inputs using a trained projection, a multi-layer perceptron (MLP) in LLaVa-

[1]https://github.com/haotian-liu/LLaVA

1.5, that projects image features from a CLIP encoder (Ramesh et al., 2022) into the word embedding space of an LLM, Vicuna-13B (Chiang et al., 2023). We experiment with prompting LLaVa for binary classification with a meme image and other information that could improve the models' performance and fine-tune the model using the training set for the task. We fine-tune the model using LoRA, a widely used training technique for reducing the number of trainable parameters. We use LoRA to reduce training time and required GPU memory.

3.2 Machine Translation

For the three unseen languages (Bulgarian, North Macedonian and Bulgarian), that are part of the test set, we use the machine translation model NLLB, as this model can translate English into all three unseen languages and is trained at sentence-level which matches the short form text within memes. As LLaVa is English only, we will translate all non-English transcriptions from the test set into English and use these as inputs for LLaVa, allowing LLaVa still to see the visual content of the original meme while receiving text input it understands.

4 Experimental Setup

4.1 Data processing

The dataset provided was split into 3 sets: 1200 training, 150 validation and 300 unlabeled development examples used for early testing and a leaderboard available before the test set was released. These splits were entirely in English. The final test set contained 600 memes in English, 100 in Bulgarian, 100 in North Macedonian and 160 in Arabic. We used the training set for fine-tuning our model, the validation set for finding the best prompts for our model, and the development set to get our results when the labels were released.

To preprocess the data, we removed all new lines and non-Latin characters from English and translated all non-English text from the test set into English before inputting them into the model.

4.2 Hyperparameters

The two hyperparameters we experimented with were the LoRA parameters rank (r) and α. r controls the trainable parameters for fine-tuning, and α is a scaling parameter that affects how much the LoRA adaption weights affect the base model weights. We experimented with every combination of the following values

- r - [8, 16]
- α - [4, 8, 16, 32]

We did experiment with numbers outside this range, but they only worsened the model's performance. We trained the model for 1 epoch, using a single 80GB A100 GPU. We used Python 3.10.13 and the Hugging Face models *liuhaotian/llava-v1.5-13b*[2] and *facebook/nllb-200-3.3B*[3].

4.3 Prompting

We experimented with different prompting techniques. We report the results in Table 1. We tested each technique using the development set as follows:

- **Basic Prompt:**

  USER: <image>\n
  Does this meme contain any propagandistic or persuasive techniques? Answer with "yes" or "no"\n
  ASSISTANT:

- **Meme Text Included:**

  USER: <image>\n
  This meme contains the text: <text>. Does this meme contain any propagandistic or persuasive techniques? Answer with "yes" or "no"\n
  ASSISTANT:

- **Persuasive/Propaganda:**
  Here, we experimented with using different words for the techniques.

  USER: <image>\n
  This meme contains the text: <text>. Does this meme contain any <propaganda/persuasive> techniques? Answer with "yes" or "no"\n
  ASSISTANT:

- **Examples of Persuasion techniques:**
  Here we experiment by providing an example of some persuasion techniques. We tested every combination of 1-5 persuasion techniques from subtask 2b and found the following prompt to perform the best.

[2]https://huggingface.co/liuhaotian/llava-v1.5-13b
[3]https://huggingface.co/facebook/nllb-200-3.3B

USER: <image>\n
You are tasked with detecting the presence of propaganda techniques in memes. Examples of propaganda techniques are: [Black-and-white Fallacy/Dictatorship, Doubt, Slogans, Appeal to authority, Bandwagon] This meme contains the text: <text>. Does this meme contain any propaganda techniques? Answer with just "Yes" or "No" \n
ASSISTANT:

- **Few-shot example prompt:** We experimented with providing an example of a propagandistic meme within the prompt, hoping to improve the model's classification performance. We could only give the model the transcription from a propagandistic meme, as the LLaVa model was only trained to receive one input image.

  USER: <image>\n
  You are tasked with detecting the presence of propaganda techniques in memes. Some but not all examples of propaganda techniques are: [Black-and-white Fallacy/Dictatorship, Doubt, Slogans, Appeal to authority, Bandwagon]. For example, a meme that contains the text: [American democracy and the Soviet system may peacefully exist side by side and compete with each other. But one cannot evolve into the other. (J. Stalin)] contains propaganda techniques. This meme contains the text: [<Meme Transcription>]. Does this meme contain any propaganda techniques? Answer with just "Yes" or "No".\n
  ASSISTANT:

  We use this final prompt when testing and fine-tuning our model. For fine-tuning, we pair it with a desired output of *yes* if the meme is propagandistic and *no* otherwise. Before evaluating the models, we convert *yes* and *no* back to their corresponding labels.

5 Results

Table 2 presents the results of fine-tuning our model on training data and testing it on the development set. As reported, fine-tuning using our experimental setup only worsened the model's performance, so we chose the untrained baseline LLaVa-1.5 model for submission using the test set.

Technique	Macro-F1
Basic Prompt	0.42
Meme Text Included	0.45
Persuasive	0.50
Propagandistic	0.60
Example techniques	0.65
Few-shot example	**0.66**

Table 1: Results from using different prompting techniques. The best results are marked as **bold**

.

r	α	Macro-F1	Micro-F1
16	32	0.62	0.73
	16	0.60	0.72
	8	0.53	0.70
	4	0.54	0.70
8	32	0.65	0.74
	16	0.61	0.72
	8	0.50	0.70
	4	0.57	0.71
Untrained		**0.69**	**0.75**

Table 2: Results on the dev set from our standard training strategy. Best results are marked as **bold.**

Table 3 presents the final results for the leaderboard for our attempt at the task. Our worst performing result was in the English language. However, we performed better in the unseen languages, with our best result being **2nd place in Arabic.**

5.1 Error Analysis

Figure 3 shows a confusion matrix for the untrained model on the development set. We can see that by proportion, it made more errors predicting propagandistic memes as non-propagandistic than otherwise.

By prompting the model, we can "ask" the model why it made a classification for any meme, which increases the explainability and error analysis capabilities of a model like this. We provide output explanations for two examples where it predicts pro-

Final Submission		
Language	Macro-F1	Place
English	0.64	15
Bulgarian	0.54	8
North Macedonian	0.46	10
Arabic	0.61	2

Table 3: Subtask 2b results from the final submission, with placement in the leaderboard for each language.

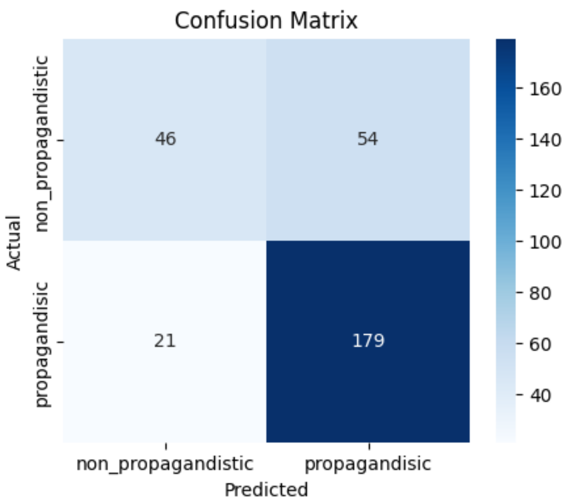

Figure 3: Confusion Matrix of untrained LLaVa on the dev set.

pagandistic (P) as propagandistic (NP) and where it predicts NP as P.

- **Classifying P as NP** - Dev Set ID 28493:

  > "The meme does not contain any of the propaganda techniques listed. It is a simple image of two men shaking hands, with a caption that is not promoting any particular ideology or using any manipulative techniques. The meme is not trying to persuade or influence the viewer's opinion, but rather it is simply presenting a visual representation of a positive interaction between two individuals."

  The original meme requires context about Russian President Vladimir Putin's relationship with intergovernmental forum G20, which the model fails to recognise. It claims the meme doesn't contain any of the propaganda techniques listed. This is likely referring to the example persuasion technique provided in the original prompt, which the model may be confusing for the **only** persuasion techniques to look for.

- **Classifying NP as P** - Dev Set ID 45439:

  > "Yes, this meme contains a propaganda technique called "Slogans." The meme is promoting a political message, specifically advocating for expanding the court. The use of the hashtag WENEEDNINE and the phrase "Expand the Court" are examples of slogans that aim to influence public opinion and support for the political cause. Slogans are often used in propaganda to create a memorable and catchy phrase that can be easily shared and repeated, thus increasing the reach and impact of the message."

  Although the meme does contain slogans in its image, they aren't part of the message and joke of the meme; the slogans are there because the meme is about a politician who happens to be photographed standing behind a slogan. However, the model fails to make this distinction.

6 Conclusion

We have presented our system for SemEval 2024 Task 4 Subtask 2b. We prompted and fine-tuned an auto-regressive large visual language model and showed that LLaVa can be used for non-English persuasion technique detection in memes through improved prompting and machine translation, with our best ranking being 2nd on the Arabic leaderboard. We experimented with different techniques for prompting to discover which produced the best-performing output. We also analysed the type of errors LLaVa can produce when classifying memes, showing how this model can easily be prompted for explainability. Further work is required to improve training LVLMs for this task, as we could not see improvements through fine-tuning.

Acknowledgements

This work has been co-funded by the European Union under the Horizon Europe vera.ai (grant 101070093) and Vigilant (grant 101073921) projects and the UK's innovation agency (Innovate UK) grants 10039055 and 10039039.

References

Deptii Devendra Chaudhari and Ambika Vishal Pawar. 2021. Propaganda analysis in social media: a bibliometric review. *Information Discovery and Delivery*, 49(1):57–70. Publisher: Emerald Publishing Limited.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. SemEval-2021 Task 6: Detection of Persuasion Techniques in Texts and Images. ArXiv:2105.09284 [cs].

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

Bin Guo, Yasan Ding, Lina Yao, Yunji Liang, and Zhiwen Yu. 2020. The future of false information detection on social media: New perspectives and trends. *ACM Comput. Surv.*, 53(4).

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Peiguang Li, Xuan Li, and Xian Sun. 2021. 1213Li at SemEval-2021 task 6: Detection of propaganda with multi-modal attention and pre-trained models. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1032–1036, Online. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved Baselines with Visual Instruction Tuning. ArXiv:2310.03744 [cs].

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual Instruction Tuning. ArXiv:2304.08485 [cs].

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.

G. Da San Martino, A. Barrón-Cedeño, H. Wachsmuth, R. Petrov, and P. Nakov. 2020. SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles. ArXiv:2009.02696 [cs].

Sadiq Muhammed T and Saji K. Mathew. 2022. The disaster of misinformation: a review of research in social media. *International Journal of Data Science and Analytics*, 13(4):271–285.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. SemEval-2023 Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

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