# BERTastic at SemEval-2024 Task 4: State-of-the-Art Multilingual Propaganda Detection in Memes via Zero-Shot Learning with Vision-Language Models

**Tarek Mahmoud**[†◇], **Preslav Nakov**[†]
[†]Mohamed Bin Zayed University of Artificial Intelligence,[◇] Presight
{tarek.mahmoud,preslav.nakov}@mbzuai.ac.ae

## Abstract

Analyzing propagandistic memes in a multilingual, multimodal dataset is a challenging problem due to the inherent complexity of memes' multimodal content, which combines images, text, and often, nuanced context. In this paper, we use a VLM in a zero-shot approach to detect propagandistic memes and achieve a state-of-the-art average macro F1 of $\mathbf{66.7}\%$ over all languages. Notably, we outperform other systems on North Macedonian memes, and obtain competitive results on Bulgarian and Arabic memes. We also present our early fusion approach for identifying persuasion techniques in memes in a hierarchical multilabel classification setting. This approach outperforms all other approaches in average hierarchical precision with an average score of $\mathbf{77.66}\%$. The systems presented contribute to the evolving field of research on the detection of persuasion techniques in multimodal datasets by offering insights that could be of use in the development of more effective tools for combating online propaganda.

## 1 Introduction

Propaganda is an ancient technique that has existed for thousands of years[1]. The way propaganda is understood today was formalized between 1937 and 1942 by the Institute of Propaganda Analysis through a series of publications (Cantril (1938), Edwards (1938), Lavine et al. (1940), and Brace (1939)). Britannica defines propaganda as the "dissemination of information—facts, arguments, rumours, half-truths, or lies—to influence public opinion."[1]. Propaganda can be beneficial when it unites people behind a noble or beneficial cause. It can also be harmful if it leads to tensions, destabilization, and the death of millions. In our digitally mediated world, transmitting (dis)information to millions of people occurs in seconds. Hence, the adverse effects of propaganda are accelerated and amplified. Propaganda has been used to influence public opinion on Brexit (Rawlinson, 2020), US elections (Chernobrov and Briant, 2020), and the Ukraine crisis (Chernobrov and Briant, 2020). Thus, it is easy to see the damaging effects propaganda has already caused and continues to inflict.

Propaganda takes many forms. It could be broadcast on television (Pan et al., 2020), spread through coordinated communities on social media (Hristakieva et al., 2022), transmitted across national borders through loudspeakers (Seo, 2018), disseminated via news articles (Nakov et al., 2022), published on blogs (Burgers, 2017), or could even exist on postage stamps (Lauritzen, 1988). More recently, memes have become powerful tools for the dissemination of political messages. The visual and textual simplicity of memes, combined with their viral nature, allows them to be rapidly consumed and shared across social media platforms, reaching vast audiences with minimal effort. This level of accessibility makes memes an attractive medium for propagandists seeking to subtly influence public opinion, disseminate misinformation, and polarize communities.

Consequently, there has been an increased need for and interest in propaganda identification in the research community. The most difficult challenge is that propaganda is often based on kernels of truth and is presented in a misleading way, making it seem genuine. Hence, training a model to detect propaganda is challenging, given the subtlety in how propaganda masquerades as an ordinary text or an innocent, funny meme. Moreover, interpreting any such model's results could also be problematic. With memes, the challenge, is further exacerbated due to the inherent complexity of memes' multimodal content, which combines images, text, and often, nuanced context, making the detection of propaganda all the more challeng-

[1] https://www.britannica.com/topic/propaganda



*Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 503–510

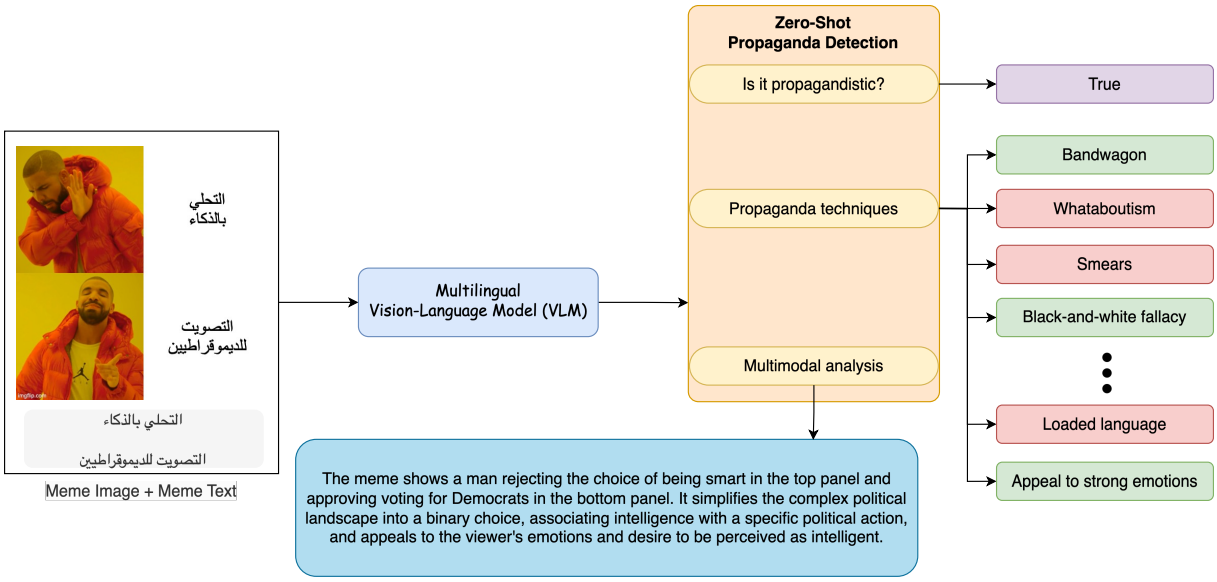



Figure 1: **Zero-shot propaganda detection overview:** This instance shows an Arabic meme that reads "being smart" in the top half and "voting democrat" in the other half. The figure illustrates the comprehensive analysis of a meme and the fine-grained output obtained from a VLM using zero-shot learning. The model accepts as input both the meme image and the meme text, and is prompted to provide three outputs. It provides a multimodal analysis and description of the meme. It also identifies which technique out of 22 possible techniques are present in the meme. Finally, it makes a determination whether the meme is propagandistic or not.

ing. Furthermore, the language used in memes is characteristically concise, often consisting of mere sentence fragments or a few keywords. Consequently, developing systems that consider only the textual content in isolation from the accompanying image presents a significant challenge.

The model explainability challenge has been tackled by Da San Martino et al. (2019) for texts and by Dimitrov et al. (2021) for memes via the introduction of fine-grained propaganda detection tasks and datasets. The tasks required identifying the propaganda technique(s) out of over eighteen techniques in a multi-label classification formulation. This fine level of granularity increases interpretability of propaganda detection models. However, the subtlety challenge is still prevalent. In addition to the subtlety challenge, propaganda detection research is scarce on multimodal datasets in general, and on memes in particular, correlating with a scarcity of datasets. Moreover, the majority of research and datasets are monolingual and consider mainly the English language. Therefore, the difficulty of tackling propaganda in memes also extends to include the scarcity of multilingual meme datasets. There are recent efforts to address this challenge, which are currently spearheaded by the shared task on *Multilingual Detection of Persuasion Techniques in Memes* (Dimitrov et al., 2024) described in detail in 2.1.

In this paper, we capitalize on recent advances in Vision Language Models (VLMs) (Zhang et al., 2023) and Pretrained Language Models (PLMs) (Zhao et al., 2023), and highlight the following contributions:

- We achieve state-of-the-art performance on multilingual, multimodal propaganda detection in memes with an average macro F1 score of $\mathbf{66.7}\%$ using a zero-shot VLM approach (see Figure 1). This includes state-of-the-art performance on North Macedonian memes, and competitive results on Bulgarian and Arabic memes.

- We present a early fusion approach for identifying persuasion techniques in memes in a hierarchical multilabel classification setting. This approach outperforms all other approaches in average hierarchical precision with an average score of $\mathbf{77.66}\%$.

## 2 Background

### 2.1 Task Formulation

The "Multilingual Detection of Persuasion Techniques in Memes" task contains three subtasks addressing the challenge of identifying persuasion

| Subtask | Binary | | Multilabel | |
|---|---|---|---|---|
| Split | Train | Test | Train | Test |
| English | 1650 | 600 | 9050 | 1500 |
| Arabic | - | 160 | - | 120 |
| Bulgarian | - | 100 | - | 436 |
| North Macedonian | - | 100 | - | 259 |

Table 1: Dataset summary. All labeled data from the training, development, and validation sets are merged and included under the training split. We also augment the multilabel training split with non-propagandistic samples from the binary training split.

techniques used in memes out of which we describe two subtasks. One subtask simplifies the challenge to a binary classification task, determining the presence or absence of any persuasion technique in a meme. In more concrete terms, given a text-image pair $p = (m, t)$ where $m$ is the meme image and $t$ is the meme text, the goal is to predict whether $p$ is propagandistic or not.

The other subtask requires the identification of one or more of twenty-two persuasion techniques within a meme. That is, given a text-image pair $p_i = (m, t)$, the goal is to learn a mapping $f : p \rightarrow K$ where $K = [k_1, ..., k_n]$ and $k_j \in \{True, False\}$ denotes whether $p_i$ contains the $j^{th}$ persuasion technique and $n$ denotes the total number of persuasion labels, which is 22 in this subtask.

The task employs macro-F1 scores for binary classification, and hierarchical-F1 scores for multilabel classification.

Table 1 summarizes the dataset. Note that the task introduces memes in languages other than English without any labels in order to evaluate the models' zero-shot learning capabilities. Figure 2 analyzes how balanced the label distribution is in the dataset of the multilabel task. It is clear the dataset is highly imbalanced. The binary task's dataset is also imbalanced with two-thirds of the training data being propagandistic.

### 2.2 Related Work

Recent advancements have highlighted the multimodal nature of modern propaganda, particularly within social media. The integration of text and visual content in memes presents a unique challenge for detection algorithms and models. Recognizing this, Dimitrov et al. (2021) introduce a multi-label multimodal task focused on identifying the specific propaganda techniques used in memes. The authors have compiled and released a corpus of approximately one thousand memes. This collection is annotated with twenty two distinct propaganda techniques. These techniques appear either in the textual content, the image content, or a combination of both. The creation of such a dataset is a significant contribution to the field, providing a foundational resource for developing and evaluating propaganda analysis models on memes. One limitation of this dataset is that it only contains English memes. This challenge is overcome by Dimitrov et al. (2024) who introduce a multilingual meme dataset that contains approximately ten thousand memes in four languages including English, Arabic, Bulgarian, and North Macedonian. In this study, we use this dataset.

Dimitrov et al. (2021) evaluate many baselines on the English meme dataset. These approaches include text models, image models, and multimodal models. Unlike our work, none of these baselines utilize a zero-shot learning approach using VLMs.

## 3 System Overview

### 3.1 Zero-Shot Detection of Propagandistic Memes using Vision-Language Models

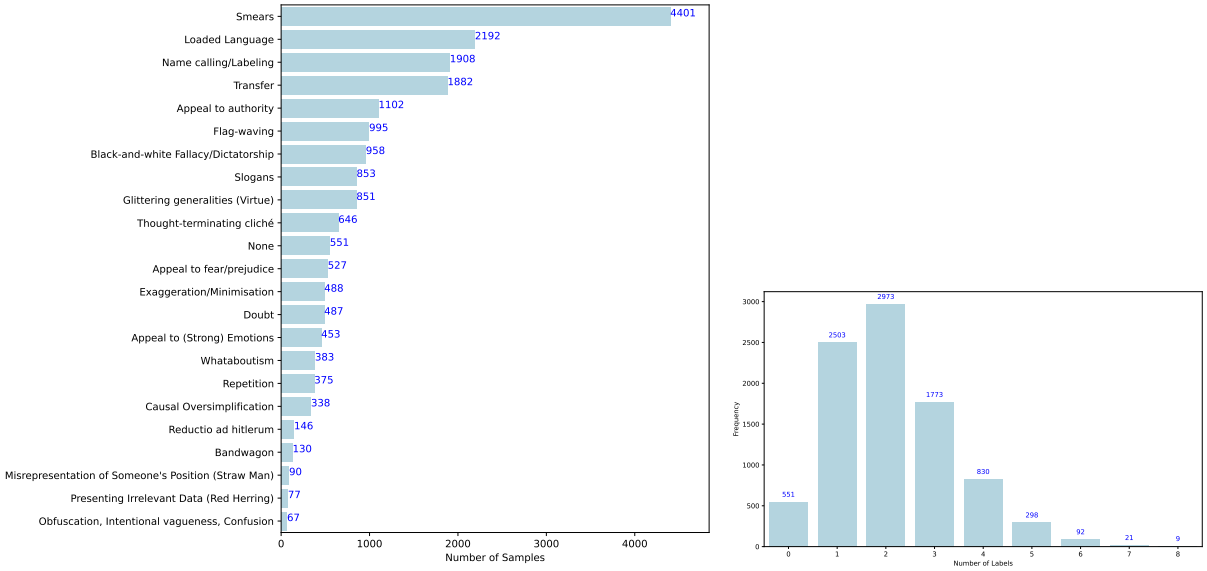
We employ zero-shot detection of propagandistic memes using GPT-4V (OpenAI, 2023). The core objective of our system illustrated in Figures 1 and 4 is to automatically identify and analyze the propagandistic content within memes. Upon processing the meme, the system utilizes GPT-4V to perform a comprehensive analysis that includes the following tasks executed in a single prompt:

1. **Analysis of Meme Text:** The model interprets the text within the meme, considering its semantic and contextual relevance to the image.

2. **Persuasuin Technique Identification:** The model assesses the meme against a predefined list of propaganda techniques, analyzing both the image and text to identify which, if any, techniques are present.

3. **Overall Propagandistic Determination:** The model concludes whether the meme is propagandistic, utilizing both the visual and textual content, and the above analysis, to inform its judgment.

The output of this analysis is structured as a JSON object, which includes three key attributes:




(a) (b)

Figure 2: Label distribution analysis for the multilabel task. (a) **Label frequencies.** The dataset is highly imbalanced with the number of labeled data for each label ranging from as little as 67 samples to as many as 4401 samples. (b) **Label count frequencies**. The majority of samples contain 1 to 4 persuasion techniques. A few samples contain 5 or more techniques and the maximum label count per sample is 8 techniques.

- **Description:** A description of the meme, obtained through multimodal analysis of both the visual and the textual contents of the meme.
- **Techniques:** A list of propaganda techniques identified in the meme.
- **Propagandistic:** A value indicating whether the meme is considered to be propagandistic or not.

This structured output enables a clear, concise, and automated method for identifying and categorizing memes by their propagandistic content, allowing such output to be used in other formulations and experiments that we outline in this paper.

### 3.2 Early Fusion for Multilabel Persuasion Identification

Our method for multilabel persuasion technique identification in memes incorporates an *early fusion* strategy, utilizing embeddings from both text and image modalities to enrich the feature space. This approach, illustrated in Figure 3, involves two key steps:

1. Embeddings Extraction:
   - We use a multilingual MPNet model (Song et al., 2020; Reimers and Gurevych, 2019) to extract embeddings from the meme's text.
   - The multilingual MPNet is also used to generate embeddings from a meme description that was obtained via a VLM as described earlier in Figure 1.
   - A CLIP-ViT-B-32 multilingual model (Radford et al., 2021; Reimers and Gurevych, 2019) processes the meme image alongside the meme text and the VLM-generated description. This model is used to capture the relationship between visual elements and textual information in memes, producing comprehensive multimodal embeddings.
2. Fusion and Classification:
   - The embeddings from the multilingual MPNet (for both meme text and description) and the CLIP-ViT-B-32 model are fused into a single feature vector. This fusion happens before training the classifier, ensuring the classifier operates on a rich representation of each meme. We use logistic regression for classification.
   - Note that the weights of the embedding models (MPNet and CLIP-ViT-B-32) are frozen during training.

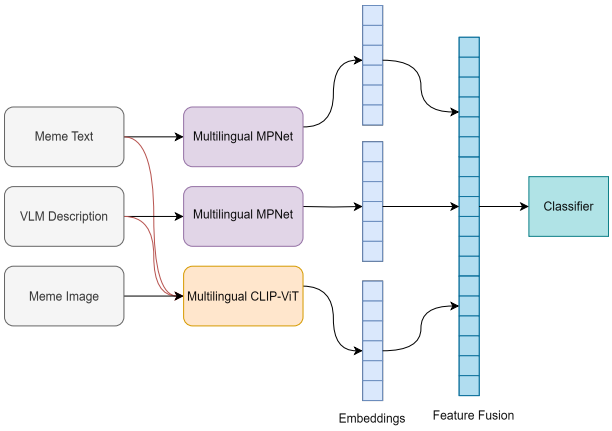



Figure 3: **Early fusion approach overview.** VLM description is obtained as described in Figure 1. CLIP-VIT receives as input all three of the meme image, meme text, and the VLM description. We obtain three separate embeddings which are then concatenated and used to train a classifier.

| Team | Avg F1 | Macro F1 | | | |
|---|---|---|---|---|---|
| | | Arabic | English | Bulgarian | North Macedonian |
| BERTastic (ours) | **66.67** | 60.28 | 71.58 | 66.21 | **68.63** |
| BCAmirs | 65.66 | **61.49** | 80.34 | 64.72 | 56.1 |
| NLPNCHU | 63.51 | 58.52 | 78.8 | 64.71 | 52.0 |
| Snarci | 62.52 | 55.54 | 79.86 | 66.78 | 47.92 |
| LMEME | 60.85 | 36.2 | **81.03** | **67.1** | 59.08 |
| SheffieldVeraAI | 56.16 | 61.03 | 64.2 | 53.62 | 45.79 |
| BDA | 56.1 | 50.97 | 79.29 | 50.62 | 43.54 |
| DUTIR938 | 54.52 | 46.89 | 80.91 | 43.41 | 46.88 |
| HierarchyEverywhere | 52.92 | 56.2 | 56.31 | 48.55 | 50.62 |
| SuteAlbastre | 52.05 | 50.07 | 80.96 | 59.45 | 17.7 |
| Hidetsune | 48.95 | 52.82 | 71.35 | 32.67 | 38.94 |
| IITK | 47.71 | 46.71 | 48.34 | 47.26 | 48.55 |
| nowhash | 46.47 | 49.83 | 49.84 | 43.36 | 42.86 |
| MemeSifters | 45.71 | 55.65 | – | 61.14 | 66.03 |
| UMUTeam | 19.66 | – | 78.66 | – | – |
| TUMnlp | 19.6 | – | 78.41 | – | – |
| CodeMeme | 19.55 | – | 78.2 | – | – |
| LomonosovMSU | 19.31 | – | 77.23 | – | – |
| Baseline | 18.37 | 22.7 | 25.0 | 16.67 | 9.09 |
| Scalar | 17.54 | – | 70.15 | – | – |
| WhatsaMeme | 12.87 | – | 51.49 | – | – |

Table 2: **Results – binary task:** The table shows the results on the test set from the official leaderboard. It shows macro F1 results for all four languages. In addition, we also compute the average macro F1 and sort the teams by this value.

Note that we use multilingual models for all embedding models to ensure our approach generalizes well in the zero-shot scenario. We opted to use a multilingual MPNet which was trained on parallel data on 50+ languages that include Arabic, Bulgarian, and North Macedonian. This decision was made despite the availability of models like XLM-R (Conneau et al., 2020), which, although powerful, were not trained on parallel datasets and thus might not perform as well across languages especially with only English training data. The same rationale applies on why we selected a CLIP-VIT-B-32 model which has a multilingual text encoder that was trained using multilingual knowledge distillation on parallel data.

## 4 Results

We show the results on binary and multilabel classification tasks in tables 2 and 3, respectively. For binary classification, our system ranks first on North Macedonian propaganda detection task, and achieves competitive results ranking third on Arabic and Bulgarian. Collectively, we achieve the top rank in average F1 across all four languages. As for multilabel classification, our system suffers from low recall but compensates for that by a very high precision performance. Our system achieves the highest precision on Bulgarian, the second highest on Arabic, English, and North Macedonian, and the top precision score on average across all four languages. This means our system is very conservative in that it only makes predictions it is highly confident of.

We also experimented with other models for both the binary and multilabel tasks as shown in tables 4 and 5, respectively. In the binary setting, we train several models on embeddings obtained using a CLIP-ViT-B-32 model, but the zero-shot VLM approach performs better in comparison. For the multilabel task, we try several approaches. We attach a classification head and fine-tune an MPNet model on the meme descriptions that we obtained as described earlier in Figure 1. However, this did not yield a high performance. We also fine-tune DeBERTa-V3-Large (He et al., 2023) and XLM-R-Large on different subsets of available text (i.e., meme text and the VLM descriptions). Out of all combinations, we observe that DeBERTa performed best when it was fine-tuned on VLM descriptions only. This is likely due to the fact the meme texts are often short, incoherent and do not form complete sentences; hence, we deduce they may contaminate the much more coherent descriptions generated by the VLM. Moreover, we also observe that we cannot achieve competitive results using only the text modality.

In all experiments involving transformers, We fine-tune with early stopping with a patience of three epochs, use a batch size of eight, a learning rate of $5e-5$, and we accumulate the gradients for eight steps making our effective batch size $64$. We also use a binary-cross entropy loss and set a classification threshold of $0.5$.

| Team | Avg P | Arabic | | | English | | | Bulgarian | | | North Macedonian | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R |
| BERTastic (ours) | **77.66** | 38.82 | 61.29 | 28.41 | 61.34 | 81.58 | 49.14 | 54.36 | **81.16** | 40.86 | 57.33 | 86.59 | 42.85 |
| Baseline | 76.11 | 48.65 | **65.0** | 38.87 | 44.71 | 68.78 | 33.12 | 50.0 | 80.43 | 36.28 | 55.53 | **90.22** | 40.1 |
| HierarchyEverywhere | 70.91 | 43.69 | 50.99 | 38.21 | **74.59** | **86.68** | 65.46 | 46.41 | 67.08 | 35.48 | 35.69 | 68.9 | 24.08 |
| BCAmirs | 69.75 | **52.61** | 55.31 | 50.17 | 70.5 | 78.37 | 64.06 | **62.69** | 70.28 | 56.59 | **63.68** | 75.02 | 55.32 |
| NLPNCHU | 69.73 | 48.32 | 59.47 | 40.70 | 70.68 | 78.16 | 64.5 | 54.86 | 70.69 | 44.83 | 48.71 | 70.58 | 37.18 |
| SuteAlbastre | 58.48 | 51.61 | 46.94 | **57.31** | 68.48 | 71.78 | 65.47 | 61.07 | 65.96 | **56.86** | 57.55 | 49.25 | **69.22** |
| IITK | 57.65 | 45.54 | 45.73 | 45.35 | 63.6 | 76.29 | 54.54 | 44.59 | 54.08 | 37.93 | 44.0 | 54.48 | 36.9 |
| BDA | 49.15 | 41.64 | 38.25 | 45.68 | 50.39 | 51.48 | 49.34 | 48.34 | 52.26 | 44.97 | 50.14 | 54.62 | 46.34 |
| LomonosovMSU | 19.79 | – | – | – | 65.61 | 79.15 | 56.02 | – | – | – | – | – | – |
| TUMnlp | 19.52 | – | – | – | 67.72 | 78.07 | 59.79 | – | – | – | – | – | – |
| UMUTeam | 19.19 | – | – | – | 69.0 | 76.76 | 62.67 | – | – | – | – | – | – |
| Pauk | 18.63 | – | – | – | 67.53 | 74.5 | 61.75 | – | – | – | – | – | – |
| CodeMeme | 15.16 | – | – | – | 66.62 | 60.66 | **73.88** | – | – | – | – | – | – |
| WhatsaMeme | 7.84 | – | – | – | 36.59 | 31.34 | 43.96 | – | – | – | – | – | – |

Table 3: **Results – multilabel task:** The table shows the results on the test set from the official leaderboard. It shows hierarchical F1, precision (P), and recall (R) results for all four languages. In addition, we also compute the average precision and sort the teams by this value.

| Method | Training Data | Macro F1 |
|---|---|---|
| XGBoost | CLIP-ViT embeddings | 64.34 |
| LightGBM | CLIP-ViT embeddings | 70.1 |
| SVM | CLIP-ViT embeddings | 71.91 |
| Zero-shot VLM | Meme image & text | **75.08** |

Table 4: **Additional experiments – binary task.** This table reports results on the development set.

| Method | Training Data | F1 | P | R |
|---|---|---|---|---|
| MPNet + FFN | VLM descriptions | 36.34 | 57.62 | 26.55 |
| DeBERTa-V3-Large | Meme text & VLM descriptions | 41.34 | 31.33 | 60.72 |
| DeBERTa-V3-Large | Meme text | 42.15 | 29.25 | **75.4** |
| DeBERTa-V3-Large | VLM descriptions | 42.93 | 30.18 | 74.36 |
| XLM-R-Large | VLM descriptions | 43.38 | 31.98 | 67.42 |
| XGBoost | BLIP embeddings | 47.55 | 79.79 | 33.87 |
| Zero-shot VLM | Meme image & text | 52.56 | 48.8 | 56.94 |
| Early Fusion | See Figure 3 | **67.84** | **88.93** | 54.83 |

Table 5: **Additional experiments – multilabel task.** This table reports results on the development set. The values reported here are all hierarchical metrics.

## 5 Conclusion and Future Work

In this study, we introduced a state-of-the-art multilingual propaganda detection in memes using zero-shot learning with VLMs. Our approach uniquely addressed the complexities of multimodal content in memes, merging visual and textual cues in a manner that comprehensively understands and identifies propagandistic content across languages. Achieving an average macro F1 score of 66.7% across all assessed languages, our system demonstrated high performance over existing methods, particularly excelling in North Macedonian memes and showing competitive performance in Bulgarian and Arabic in the binary setting. In addition, our early fusion technique for identifying persuasion techniques in memes within a hierarchical multi-label classification setting outperformed all other approaches with an average hierarchical precision score of 77.66%.

Looking forward, our research opens several directions for further exploration and improvement. First, the exploration of advanced fusion techniques that could more intricately combine the strengths of textual and visual analyses may yield even higher accuracies in propaganda detection. Additionally, the adaptability and performance of our model in detecting subtler forms of propaganda and across a broader spectrum of languages present an exciting challenge, especially considering the highly nuanced contextual nature of meme content and its cultural intricacies.

## Acknowledgments

We would like to thank the anonymous reviewers for their time and valuable insights.

## References

Harcourt Brace. 1939. *The Fine Art of Propaganda: A Study of Father Coughlin's Speeches*. Institute for Propaganda Analysis and Lee, A.M.C. and Lee, E.B.

Matt Burgers. 2017. Yup, the russian propagandists were blogging lies on medium too | wired uk.

Hadley Cantril. 1938. Propaganda analysis. *The English Journal*, 27(3):217–221.

Dmitry Chernobrov and Emma L Briant. 2020. Competing propagandas: How the united states and russia represent mutual propaganda activities. *Politics*, 42(3):393–409.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco

Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

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

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

Violet Edwards. 1938. *Group Leader's Guide to Propaganda Analysis: Revised Edition of Experimental Study Materials for Use in Junior and Senior High Schools, in College and University Classes, and in Adult Study Groups*. Institute for propaganda analysis, Incorporated.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing.

Kristina Hristakieva, Stefano Cresci, Giovanni Da San Martino, Mauro Conti, and Preslav Nakov. 2022. The spread of propaganda by coordinated communities on social media. In *14th ACM Web Science Conference 2022*, WebSci '22, page 191–201, New York, NY, USA. Association for Computing Machinery.

Frederick Lauritzen. 1988. Propaganda art in the postage stamps of the third reich. *The Journal of Decorative and Propaganda Arts*, 10:62–79.

H. Lavine, J.A. Wechsler, and Institute for Propaganda Analysis. 1940. *War Propaganda and the United States*. International propaganda and communications. Yale University Press.

Preslav Nakov, Giovanni Da San Martino, and Firoj Alam. 2022. Fact-checking, fake news, propaganda, media bias, and the covid-19 infodemic. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 1632–1634, New York, NY, USA. Association for Computing Machinery.

OpenAI. 2023. [link].

Jennifer Pan, Zijie Shao, and Yiqing Xu. 2020. The effects of television news propaganda: Experimental evidence from china. *Available at SSRN 3579148*.

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.

Francis Rawlinson. 2020. "how press propaganda paved the way to brexit", by francis rawlinson - consilium.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Yoonjung Seo, Joshua Berlinger. 2018. South korea stops blasting propaganda as summit looms | cnn.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2023. Vision-language models for vision tasks: A survey.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models.

# Appendix

## A Prompt Template

The system illustrated in Figure 1 uses the prompt template illustrated in Figure 4. The technique list consists of the names of the twenty-two techniques available in the dataset.

```
You are a helpful meme propaganda analyst designed to output JSON  without any markdown formatting. The
JSON you output must adhere to JSON rules such as case sensitivity for boolean values, and not having
double quotes inside of strings.

Consider the attached meme with the extracted meme text shown below:
{meme_text}

Provide a propaganda/persuasion analysis of the attached meme image alongside the extracted meme text.
Think of propaganda techniques such as:
{technique_list}

Return a JSON that has these attributes:
- description: a generic English description of the meme. Please don't use double quotes.
- techniques: a list of the techniques that are 100% without a doubt present in the meme
- is_propagandistic: True if meme is propagandistic, False otherwise

Only return the JSON, without any explanation. Make sure the JSON is properly formatted. For example,
to quote something inside of a JSON string, use single quotes. Never use double quotes. Also note that
JSON boolean is case sensitive.

{meme_image}
```

Figure 4: **Prompt template:** This is the template used in the binary classification setting illustrated in Figure 1