# MODOS at ArAIEval Shared Task: Multimodal Propagandistic Memes Classification Using Weighted SAM, CLIP and ArabianGPT

**Abdelhamid Haouhat**
Lab. d'Informatique
et de Mathématiques
Université Amar Telidji
Laghouat, Algeria
a.haouhat
@lagh-univ.dz

**Hadda Cherroun**
Lab. d'Informatique
et de Mathématiques
Université Amar Telidji
Laghouat, Algeria
hadda_cherroun
@lagh-univ.dz

**Slimane Bellaouar**
Lab. Mathématiques
et Sciences Appliquées
Université de Ghardaia
Algeria
bellaouar.slimane
@univ-ghardaia.dz

**Attia Nehar**
Faculty of Exact Sciences
and Computer Science
*University of Djelfa*
Algeria
neharattia
@univ-djelfa.dz

## Abstract

Arabic social media platforms are increasingly using propaganda to deceive or influence people. This propaganda is often spread through multimodal content, such as memes. While substantial research has addressed the automatic detection of propaganda in English content, this paper presents the MODOS team's participation in the Arabic Multimodal Propagandistic Memes Classification shared task.[1] Our system deploys the Segment Anything Model (SAM) and CLIP for image representation and ARABIAN-GPT embeddings for text. Then, we employ LSTM encoders followed by a weighted fusion strategy to perform binary classification. Our system achieved competitive performance in distinguishing between propagandistic and non-propagandistic memes, scored 0.7290 macro F1, and ranked 6th among the participants.

## 1 Introduction

The term "meme" was introduced by Richard Dawkins in his renowned publication, The Selfish Gene (Dawkins, 2006). Dawkins derived the term from the Greek word "mimeme," modifying it to denote a concept signifying "something which is imitated." In the era of social media, memes have frequently become a common method of communication, often used to spread messages with varied degrees of propaganda (Shifman, 2013). Propaganda is defined as deliberate expressions to influence opinions or actions, often using rhetorical and psychological devices.3 Understanding the dynamics of online conversation and preventing the spread of false information depends on the ability to identify propagandistic content. Consequently, researchers in Natural Language Processing (NLP) are increasingly interested in automatically detecting the application of propaganda techniques in text, images, and multimodal information (Dimitrov et al., 2021; Hasanain et al., 2023a). Prior work in Arabic has focused primarily on detecting propaganda in a social media text, particularly tweets (Alam et al., 2022; Hasanain et al., 2023a; Alam et al., 2024). Several shared tasks have been launched to address the detection of propaganda in Arabic tweets and the identification of persuasion techniques and disinformation in Arabic text.[2],[3] These tasks have attracted many researchers (Ojo et al., 2023; Xiao and Alam, 2023; Lamsiyah et al., 2023; Mittal and Nakov, 2022; Refaee et al., 2022; Attieh and Hassan, 2022; Samir et al., 2022; Hasanain et al., 2023b), leading to significant advances in recognizing and extracting linguistic cues that indicate the use of propaganda techniques. In these tasks, researchers have achieved good results in detecting propaganda in Arabic text by identifying the propaganda techniques used, which often include playing on the audience's emotions (e.g., appealing to fear, using loaded language, etc.) (Miller, 1939). Despite significant progress conducted in multimodal learning for Arabic, which focused on tasks such as sentiment analysis, emotion recognition, summarization, and more, rather than propaganda detection. For instance, Haouhat et al. (Haouhat et al., 2023) developed an Arabic multimodal dataset for sentiment analysis that comprises transcripts, voice recordings and videos, enhancing the ability to capture nuanced content. These efforts highlight the potential of multimodal approaches in improving understanding across different domains. However, the specific challenge of detecting propaganda in multimodal formats remains less explored, indicating a critical gap in this area. To address the gap in Arabic multimodal propaganda detection research, Hasanain et al. (Hasanain et al., 2024b)

[1] https://araieval.gitlab.io/

[2] WANLP-2022 Shared Task on propaganda detection in Arabic. Available at https://sites.google.com/view/propaganda-detection-in-arabic/home, Accessed on 17/05/2024

[3] https://araieval.gitlab.io/2023/



*Proceedings of The Second Arabic Natural Language Processing Conference*, pages 483–488

introduced a subtask to the ArAIEval Shared Task at ArabicNLP 2024.[4] Previous research predominantly focused on unimodal textual detection, leaving a notable gap in understanding the propagandistic content presented in multimodal formats. This subtask specifically focused on the classification of memes, representing a significant advancement in the field. Our team actively participated in this initiative, specifically targeting the second subtask, *subtask2C*, where we aimed to contribute to the detection of propagandistic content in multimodal memes. Despite the complexity of the task, our system demonstrated promising results, securing the sixth position in the competition. Our contribution involved the development of a comprehensive approach that integrates various techniques, including image segmentation, feature extraction, multimodal weighted fusion, and classification.

## 2 Related Work

(Haouhat et al., 2023) This section summarizes the previous research on Arabic propaganda detection in both unimodal and multimodal contexts. Authors in (Khanday et al., 2021) address propaganda dissemination through social networking platforms. They focus on classifying propagandist text from non-propagandist text using supervised machine learning algorithms. Data collected from news sources is annotated, and feature engineering is conducted using techniques such as term frequency/inverse document frequency (TF/IDF) and Bag of Words (BOW). Support Vector Machine (SVM) and Multinomial Naïve Bayesian (MNB) classifiers are employed. the study yields encouraging results. Bodor et al. (Almotairy et al., 2024) tackle the challenge of detecting and characterizing Arab computational propaganda on Twitter by providing a dataset that includes 16 million tweets. However, only 2100 labeled propagandist tweets covering banking and sports are included in the dataset, and the unlabeled tweets are limited to Saudi Arabian Twitter data, necessitating generalization to diverse topics and regions within the Arab world. Despite these limitations, the dataset offers valuable insights into Arab computational propaganda and enables supervised and unsupervised machine learning and deep learning algorithm applications to classify the credibility of Arab tweets. Firoj et al. (Alam et al., 2022) highlight the importance of bridging the language gap in propaganda detection by focusing on Arabic, which previously received less attention. The authors organized a shared task conducted as part of the WANLP 2022 workshop, aiming to detect propaganda techniques in Arabic tweets through multi-label classification. The task attracted significant participation, with 63 teams.

Firoj et al. (Hasanain et al., 2023a) focus on the critical problem of identifying and labeling propagandistic spans within the textual content. The authors explore the potential of large language models (LLMs) like GPT-4 to function as both general and expert annotators for this task. The study looks at how well GPT-4 annotates spans and if it can take the place of consolidators during the annotation process. Notably, the study makes significant contributions by releasing annotations from multiple human annotators and GPT-4, aiming to benefit the research community and facilitate further advancements in propaganda detection and analysis. In (Hasanain et al., 2024a), the authors meticulously construct ArPro, the largest propaganda dataset in Arabic, comprising 8K paragraphs from newspaper articles labeled at the text span level across 23 propagandistic techniques. Additionally, the study delves into the potential of large language models (LLMs) for fine-grained propaganda detection from text. However, results reveal that while GPT-4 performs acceptably in classifying paragraphs as propagandistic or not.

## 3 Task Definition

The ArAIEval shared task is part of ArabicNLP 2024 conference, comprises two main challenges focused on the automated identification of propagandistic content within mainstream and social media. the first one deal with unimodal data by detecting propagandistic textual spans with persuasion techniques and second challenge focus on distinguishing between propagandistic and non-propagandistic multimodal memes. Participants are tasked with analyzing tweets, news articles, and memes to identify misleading information, aiding fact-checkers, journalists, and the general public in combating disinformation.

## 4 Data

In this work, we utilize the dataset released for the ArAIEval shared task (Hasanain et al., 2024b). The dataset consisted of a diverse collection of memes labeled as propaganda or non-propaganda. Table1

[4]https://arabicnlp2024.sigarab.org/

| Data | Training | Dev | Testing |
|---|---|---|---|
| Size | 2143 | 312 | 607 |
| # of P samples | 603 | 88 | 171 |
| # of NP samples | 1540 | 224 | 436 |
| # of Textual Tokens | 38490 | 5445 | 10590 |
| # of Visual Segments | 68624 | 9965 | 19709 |

Table 1: Overview of the dataset split where Propagandist and Non-propagandist classes are abbreviated as P and NP, respectively.

presents an overview of the dataset, including the size of the training, development (Dev), and testing sets, as well as the number of propagandist (P) and non-propagandist (NP) instances in each set. Additionally, we provide information about the textual tokens and visual segments generated by SAM present in the dataset. Visual segments represent distinct regions within images identified and segmented by the Segment Anything Model (SAM), facilitating multimodal analysis alongside textual data.

## 5 System

Our methodology consists of several key steps: image segmentation, feature extraction, multimodal fusion, and classification. A GitHub repository for our project is provided.[5] Figure 1 provides an overview of our approach.

### 5.1 Pre-trained Models

Before getting into the description of our approach, we start by describing the deployed pre-trained models.

Segment Anything Model (**SAM**), is a cutting-edge approach to image segmentation with outstanding performance and zero-shot capabilities. The three essential modules of SAM are the VisionEncoder, PromptEncoder, and MaskDecoder. They are all responsible for generating object masks. SAM is well-versed in both Automatic-Mask-Generation, which enables the system to automatically generate masks for every object in an image, and Prompted-Mask-Generation. It was trained on a dataset consisting of 11 million images and 1.1 billion masks (Kirillov et al., 2023). Figure 2 illustrates the segmentation of all possible objects in the image with a high score of confidence and stability set these scores to 0.95 to ensure getting masks with only minor boundary errors.

**CLIP** is a state-of-the-art multimodal vision and language model that is both scalable and resilient. This neural network is quite versatile. In contrast to conventional models, CLIP has exceptional "zero-shot" capabilities similar to GPT-2 and GPT-3, allowing it to carry out tasks like anticipating the most relevant text excerpt given the names of the visual categories to be identified (picture), all without the need for direct optimization (Radford et al., 2021).

**ArabianGPT-08** is an improved version of ArabianGPT-01 with improved Arabic text processing capabilities. ArabianGPT-08 performs better in language generation tasks with enhanced accuracy and coherence thanks to adjusted parameters and a larger model size. Because of its improved tokenizer named "Aranizer", Arabic's special linguistic characteristics can be handled more effectively, leading to more accurate text processing (Koubaa et al., 2024).

We employed SAM to segment objects on all meme images. This allowed us to isolate relevant visual elements from each image for further analysis. For feature extraction, our team performed cleaning up and pre-processing using the steps applied in previous NLP tasks on Arabic. We then extracted word embeddings from the text within the memes using ArabianGPT08. Additionally, for each segmented image, we extract image embeddings using a CLIP image encoder. Once all the embeddings capture both visual and textual information, we input the features using two LSTM encoders. This ensures the alignment and the equivalence of feature lengths, addressing inherent differences due to the nature of visual and textual data. The LSTM encoders generated new hidden features with dimensions of 768 by 128. We then employed a weighted fusion strategy, by calculating importance scores for each modality to combine the hidden features into multimodal representations, as shown in Equations 1 and 2.

$$h_i = Concat(f_t, f_i) \quad (1)$$

$$h = h_i^T . W_\alpha \quad (2)$$

The multilayer perceptron (MLP) effectively categorizes the multimodal features obtained from the fusion step into binary classes, indicating whether a meme is propagandistic or not. In the training process, we utilized the cross-entropy loss function $L$

[5] https://github.com/belgats/Multimodal-Propaganda-Detection

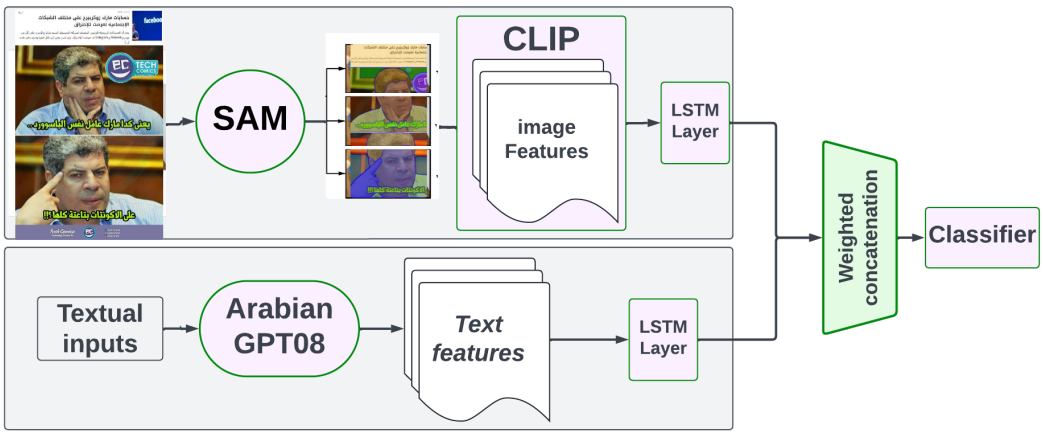


Figure 1: Overview of the approach

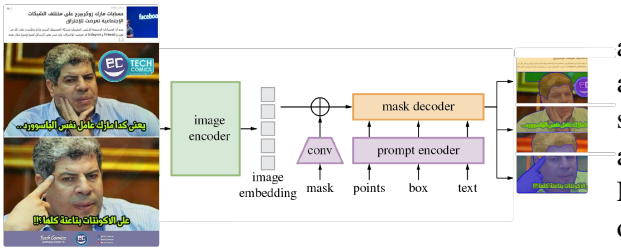


Figure 2: The image demonstrates the application of the Segment Anything Model (SAM) for segmenting objects within the image.

to quantify the discrepancy between predicted class probabilities and true labels. Equation 3 describes this function:

$$L(y, \bar{y}) = -\frac{1}{N}\sum_{n=1}^{N}\Big(y_n \log(\bar{y_n}) + (1-y_n)\log(1-\bar{y_n})\Big) \quad (3)$$

## 6 Results and Discussion

We evaluated our model against the baseline models provided by shared task organizers: Ngram, Random, ImgBert, and Majority. The metrics used are precision, recall, F1 score, and accuracy. The performance of the models is presented in Table 2.

| Model | Acc | F1 | Recall | Precision |
|---|---|---|---|---|
| Random | 0.5189 | 0.4923 | 0.5189 | 0.6094 |
| Majority | 0.7183 | 0.4180 | 0.7183 | 0.5159 |
| ImgBert | 0.7364 | 0.6563 | 0.7364 | 0.7245 |
| Ngram | 0.7628 | 0.6476 | 0.7628 | 0.7452 |
| Our's | **0.8040** | **0.7348** | **0.8040** | **0.7948** |

Table 2: Performance Metrics of the Multimodal Propagandistic Memes Classification Model and Baselines

Our model demonstrates superior performance across all metrics, achieving an accuracy of 0.8040 and an F1 score of 0.7348 on the test set, ranking sixth in the leaderboard, the best performance with an F1 score of 0.8051 achieved by the AlexUNLP-MZ team. Our finding indicate the effectiveness of our approach in correctly classifying the memes, with a good balance between precision and recall, efficiently handling both false positives and false negatives. However, there are observed limitations in model generalization due to not employing several propaganda techniques (e.g., name-calling, virtue words, deification, testimonial, fear, etc.) that could enhance the detection. Furthermore, The unbalanced class distribution in the dataset poses a challenge, potentially biasing the model towards the majority class and affecting its generalization capability.

### 6.1 Baselines

In this section, we define the baselines used for evaluating our approach:

The **Majority Baseline** classifies all instances with the most frequent class label found in the training dataset. It serves as a simple benchmark by predicting the most common class without considering any features of the data.

The **Random Baseline** predicts class labels by randomly selecting from the Propagandistic and non Propagandistic classes. This model also does not utilize any data features, instead provides a benchmark by guessing the labels randomly.

The **N-gram Baseline** employs a machine learning pipeline that includes a TF-IDF vectorizer to convert text data into numerical features and a Support Vector Machine (SVM) classifier for predict-

ing class labels. This method leverages n-gram features of the text, capturing patterns and dependencies in the textual data to inform classification.

The **ImgBert Baseline** integrates both image and text features for classification. It utilizes Preprocessed features from both modalities, combines them, and applies an SVM classifier to predict class labels. This approach is more complex, leveraging the advantages of multimodal data to enhance predictive accuracy.

## 7 Conclusion

In this study, we presented our approach to the Multimodal Propagandistic Memes Classification subtask at the ArAIEval shared task. By leveraging a combination of image segmentation, feature extraction, multimodal fusion, and fusing weighted strategy, we captured rich information from both modalities. We developed a model capable of effectively identifying propagandistic content in memes. Our experimental results demonstrate that the proposed model outperforms several baseline methods, achieving higher accuracy and F1 scores. Overall, Our findings indicate good participation. However, there is room for improvement, particularly in addressing overfitting and enhancing generalization capabilities. Future work will explore advanced multimodal fusion techniques, incorporate additional contextual information, and utilize larger and more diverse datasets.

## References


Firoj Alam, Abul Hasnat, Fatema Ahmed, Md Arid Hasan, and Maram Hasanain. 2024. Armeme: Propagandistic content in arabic memes.

Firoj Alam, Hamdy Mubarak, Wajdi Zaghouani, Giovanni Da San Martino, and Preslav Nakov. 2022. Overview of the WANLP 2022 shared task on propaganda detection in Arabic. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 108–118, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Bodor Moheel Almotairy, Manal Abdullah, and Dimah Hussein Alahmadi. 2024. Dataset for detecting and characterizing arab computation propaganda on x. *Data in Brief*, 53:110089.

Joseph Attieh and Fadi Hassan. 2022. Pythoneers at wanlp 2022 shared task: Monolingual arabert for arabic propaganda detection and span extraction. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 534–540.

Richard Dawkins. 2006. *The selfish gene*. Oxford university press.

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. *arXiv preprint arXiv:2109.08013*.

Abdelhamid Haouhat, Slimane Bellaouar, Attia Nehar, and Hadda Cherroun. 2023. Towards arabic multimodal dataset for sentiment analysis. In *2023 Fourth International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*, pages 126–133.

Maram Hasanain, Fatema Ahmed, and Firoj Alam. 2023a. Large language models for propaganda span annotation. *arXiv preprint arXiv:2311.09812*.

Maram Hasanain, Fatema Ahmed, and Firoj Alam. 2024a. Can gpt-4 identify propaganda? annotation and detection of propaganda spans in news articles. In *Proceedings of the 2024 Joint International Conference On Computational Linguistics, Language Resources And Evaluation*, LREC-COLING 2024, Torino, Italy.

Maram Hasanain, Firoj Alam, Hamdy Mubarak, Samir Abdaljalil, Wajdi Zaghouani, Preslav Nakov, Giovanni Da San Martino, and Abed Alhakim Freihat. 2023b. Araieval shared task: Persuasion techniques and disinformation detection in arabic text. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. Association for Computational Linguistics.

Maram Hasanain, Md. Arid Hasan, Fatema Ahmed, Reem Suwaileh, Md. Rafiul Biswas, Wajdi Zaghouani, and Firoj Alam. 2024b. ArAIEval Shared Task: Propagandistic techniques detection in unimodal and multimodal arabic content. In *Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*, Bangkok. Association for Computational Linguistics.

Akib Mohi Ud Din Khanday, Qamar Rayees Khan, and Syed Tanzeel Rabani. 2021. Detecting textual propaganda using machine learning techniques. *Baghdad Science Journal*, 18(1):0199.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment anything. *arXiv:2304.02643*.

Anis Koubaa, Adel Ammar, Lahouari Ghouti, Omar Najar, and Serry Sibaee. 2024. Arabiangpt: Native arabic gpt-based large language model. *Preprint*, arXiv:2402.15313.

Salima Lamsiyah, Abdelkader Mahdaouy, Hamza Alami, Ismail Berrada, and Christoph Schommer. 2023. Ul & um6p at araieval shared task: Transformer-based model for persuasion techniques

and disinformation detection in arabic. In *The 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics (ACL), Singapore, Singapore.

Clyde R Miller. 1939. The techniques of propaganda. from "how to detect and analyze propaganda," an address given at town hall. *The Center for learning*.

Shubham Mittal and Preslav Nakov. 2022. Iitd at the wanlp 2022 shared task: Multilingual multi-granularity network for propaganda detection. *arXiv preprint arXiv:2210.17190*.

Olumide E Ojo, Olaronke O Adebanji, Hiram Calvo, Damian O Dieke, Olumuyiwa E Ojo, Seye E Akinsanya, Tolulope O Abiola, and Anna Feldman. 2023. Legend at araieval shared task: Persuasion technique detection using a language-agnostic text representation model. *arXiv preprint arXiv:2310.09661*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Eshrag Ali Refaee, Basem Ahmed, and Motaz Saad. 2022. Arabem at wanlp 2022 shared task: Propaganda detection in arabic tweets. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 524–528.

Ahmed Samir, Abu Bakr Soliman, Mohamed Ibrahim, Laila Hesham, and Samhaa R El-Beltagy. 2022. Ngu_cnlp at wanlp 2022 shared task: Propaganda detection in arabic. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 545–550.

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

Yunze Xiao and Firoj Alam. 2023. Nexus at araieval shared task: Fine-tuning arabic language models for propaganda and disinformation detection. *arXiv preprint arXiv:2311.03184*.