MemeIntel: Explainable Detection of Propagandistic and Hateful Memes

Anonymous ACL submission

Abstract

The proliferation of multimodal content on social media presents significant challenges in understanding and moderating complex, context-dependent issues such as misinformation, hate speech, and propaganda. While efforts have been made to develop resources and propose new methods for automatic detection, limited attention has been given to label detection and the generation of explanation-based rationales for predicted labels. To address this challenge, we introduce *MemeXplain*, an explanation-enhanced dataset for **propagandistic memes in Arabic and hateful memes in English**, making it the *first* large-scale resource for these tasks. To solve these tasks, we propose a novel **multi-stage optimization approach** and train **Vision-Language Models (VLMs)**. Our results demonstrate that this approach significantly improves performance over the base model for both **label detection** and **explanation generation**, outperforming the current state-of-the-art with an **absolute improvement of $\sim 3\%$ on ArMeme and $\sim 7\%$ on Hateful Memes**. For reproducibility and future research, we aim to make the *MemeXplain* dataset and scripts publicly available.[1]

1 Introduction

Despite the rapid growth of multimodal content—integrating images, text, and sometimes video—the automated detection of harmful and false information on online news and social media platforms has become increasingly critical. In particular, identifying propaganda and hate in memes is essential for combating misinformation and minimizing online harm. While most research has focused on textual analysis, multimodal approaches have received comparatively less attention. In propaganda detection, text-based methods have evolved from monolingual to multilingual setups (Piskorski et al., 2023; Hasanain et al., 2023), from binary to fine-grained span-level tasks (Barrón-Cedeno et al., 2019; Habernal et al., 2017, 2018; Da San Martino et al., 2019). Hate speech detection has similarly progressed from text-based to multimodal approaches that integrate both textual and visual elements (Kiela et al., 2020; Alam et al., 2022; Sharma et al., 2022).

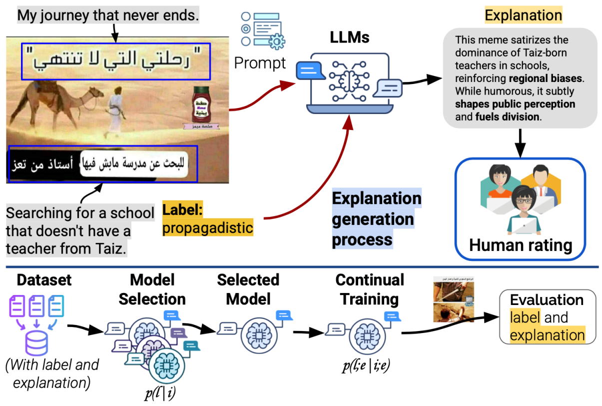


Figure 1: Experimental steps for explanation generation and training.

The emergence of LLMs has demonstrated significant capabilities across various disciplines. To capture implicit hate and propaganda, LLMs utilize different techniques (Cao et al., 2022; Kumar and Nandakumar, 2022), such as prompt-based methods, contrastive learning, and cross-modal attention. Consequently, efforts have been made to leverage VLMs (Zhang et al., 2024) to classify the harmful and propagandistic memes (Cao et al., 2023). Despite progress, detecting implicit hate, especially with sarcasm or irony, remains challenging. Propagandistic memes add complexity through emotional appeals, humor, cultural cues, and manipulative language. To address these nuances, it is crucial for a system to provide not only accurate predictions but also clear, relatable explanations that align with the meme's visual context and aid user understanding (Hee et al., 2023; Yang et al., 2023; Huang et al., 2023; Sun et al., 2023).

An explanation-based approach offers numer-

[1] anonymous.com

ous advantages and improves performance in various tasks (Li et al., 2022; Magister et al., 2022; Nandi et al., 2024; Kumari et al., 2024). Although most studies have focused on textual content (Li et al., 2022; Magister et al., 2022), some recent approaches (Nandi et al., 2024; Kumari et al., 2024) have applied explainability to images. However, these methods rely on QA-based explanations that lack naturalness, use multiple inference with custom models, thus increasing computational complexity, and employ explanations only during training rather than as an inference output. To address these limitations, we propose a simplified method that achieves state-of-the-art meme classification and explanation generation across two datasets.

We summarize our main contributions below:

- We developed an explanation-enhanced datasets, *MemeXplain*, using a rapid and low-cost annotation procedure;
- We investigated state-of-the-art VLMs to identify an appropriate model for meme classification and explanation generation;
- We proposed a novel and efficient *multi-stage optimization procedure* that mitigates gradient conflicts and avoids catastrophic forgetting by formulating the optimization problem from the perspective of *domain adaptation* and *task-incremental learning*.
- We achieved state-of-the-art performance on two types of datasets related to propaganda and hateful content detection.

Our findings are as follows: *(a)* A higher human evaluation score suggests that explanations from stronger models (e.g., GPT-4o) are reliable and can serve as gold-standard explanations for training smaller models; *(b)* Task-specific fine-tuning improves performance over the base model and *(c)* Our multi-stage optimization improves label detection and explanation quality by reducing gradient conflicts and avoiding catastrophic forgetting, achieving state-of-the-art performance. Overall, our work is the ***first*** to enhance VLMs for simultaneous propaganda and hateful content detection while providing natural reasoning to end users.

2 Related Work

The widespread use of social networks has become a major channel for spreading misinformation, propaganda, and harmful content. Significant research efforts have been directed toward addressing these challenges, particularly in multimodal disinformation detection (Alam et al., 2022), harmful memes (Sharma et al., 2022), and propagandistic content (Dimitrov et al., 2021a). However, most studies have focused on detection, while less attention has been given to generating natural explanations/reasons behind the predicted labels.

Multimodal Propagandistic Content. Following the previous research for propaganda detection using textual content (Da San Martino et al., 2019), Dimitrov et al. (2021b) introduced SemEval-2021 Task 6 focusing on persuasion techniques detection in both textual and visual memes. Subsequently, the focus has extended to the detection of multilingual and multimodal propagandistic memes (Dimitrov et al., 2024). Similar multimodal work on Arabic involves the development of datasets and shared task for propaganda detection (Alam et al., 2024b; Hasanain et al., 2024). For the detection problem, typical approaches include a fusion of textual and visual embedding and a classification head on top them (Hasanain et al., 2024; Shah et al., 2024), graph attention network based approach for multimodal objects Chen et al. (2024).

Multimodal Hate speech. Similarly, there has been growing interest in detecting multimodal hate speech (Kiela et al., 2020; Velioglu and Rose, 2020; Hee et al., 2022). Due to the lack of resources, Kiela et al. (2020) developed a large-scale dataset for multimodal hate identification. This study advanced research in this area and emphasized the importance of integrating textual and visual features for effective detection. To further progress in this field, efforts have been made to develop resources for multiple languages, including Arabic (Alam et al., 2024a), Bangla (Hossain et al., 2022), and English (Hee et al., 2023). A more detailed summary can be found in Sharma et al. (2022), which also highlights key challenges and outlines future research directions.

Training with Explanations. Integrating reasoning or explainability capabilities to enhance LLM/VLM performance has been shown to be highly beneficial for various tasks across multiple domains (Plaat et al., 2024). This approach has also proven effective for knowledge distillation and model compression (Li et al., 2022; Magister et al., 2022), where explanations generated by large LLMs improve the performance and capabilities of smaller LLMs. For example, in the hateful speech detection task, Hare (Yang et al., 2023) employs Chain-of-Thought (CoT) reasoning, while (Huang et al., 2023) utilizes Chain of Explanation (CoE). Their aim is to improve the effectiveness of LLM-based sentiment classifiers by leveraging reasoning capabilities. Sun et al. (2023) proposed Clue and Reasoning Prompting (CARP), which uses keywords and reasoning to guide explanation.

CoT-Based Approaches. CoT is a widely recognized prompting technique that generates a chain of reasoning to derive answers. Kumari et al. (2024) proposed a CoT-based framework for meme analysis using scene graphs to capture text- and image-based entity-object relationships. Their three-step prompting strategy guides the LLM to identify Emotion, Target, and Context for effective meme interpretation. SAFE-MEME (Nandi et al., 2024) introduced two multimodal datasets and a CoT-based reasoning framework for meme hate speech detection, using Q&A-style prompts and hierarchical labels. However, their method was not evaluated on the Hateful Memes dataset, limiting direct comparison.

Comparison with Prior Work. A key limitation of CoT-based methods is their reliance on multi-step reasoning, requiring multiple VLM inferences. In contrast, our approach: *(a)* avoids complex CoT steps, improving efficiency and reducing cost; and *(b)* generates explanations alongside classifications to enhance transparency and reliability. (Hee et al., 2023) constructed a dataset providing explanations for hateful memes. However, unlike us, they focused solely on evaluating explanation generation and did not perform classification tasks. Despite the availability of their data, we do not use it due to the lack of *naturalness*. In particular, their explanations do not fully account for image content or image-centric contextual perspectives.

3 Dataset

3.1 ArMeme

The ArMeme dataset comprises approximately ∼6k manually annotated Arabic memes collected from various social media platforms (Alam et al., 2024b). This dataset has been manually annotated with four labels such as *Not propaganda*, *Propaganda*, *Not-meme* and *Other*. Table 1 provides the distribution of the data splits. The memes with "Not propaganda" category covers over half of the dataset (∼66%), followed by "Propaganda" and the distribution of "Not-meme" and "Other" classes are significantly smaller. This distribution highlights a substantial class imbalance, particularly between "Not propaganda" and the other categories.

Class label	Train	Dev	Test	Total
Not propaganda	2,634	384	746	3,764
Propaganda	972	141	275	1,388
Not-meme	199	30	57	286
Other	202	29	56	287
Total	**4,007**	**584**	**1,134**	**5,725**

Table 1: Data splits for ArMeme datasets.

3.2 Hateful Meme

The Hateful Memes dataset (Kiela et al., 2020), is a benchmark designed to evaluate multimodal hate speech detection. It consists of ∼12k memes, combining both text and images, carefully curated to ensure that effective classification requires an understanding of both modalities. The dataset was created using a mix of synthetically generated memes and real-world examples, sourced from social media, while ensuring a balanced distribution of hateful and non-hateful content. In Table 2, we report the distribution of hateful meme dataset used for this study. Note that hateful meme dataset consists of two other splits (dev-seen and test-seen), here, we used unseen versions.

Class Label	Train	Dev-unseen	Test-unseen	Total
Not Hateful	5,481	340	1,250	7,071
Hateful	3,019	200	750	3,969
Total	**8,500**	**540**	**2,000**	**11,040**

Table 2: Distribution of hateful-meme dataset.

4 *MemeXplain*: Explanation Generation

The outcomes of an automatic system become more reliable for users if it provides decisions with adequate and interpretable natural explanations, which help users better understand the underlying reason behind the system's decision (Hee et al., 2023; Yang et al., 2023; Huang et al., 2023; Sun et al., 2023). Technically, this approach provides numerous advantages in terms of knowledge distillation, model compression, and enhancing the performance of target tasks in different domains (Li et al., 2022; Magister et al., 2022; Nandi et al., 2024; Kumari et al., 2024). This motivates us to adopt the explanation-based approach in our research. However, we also aim to improve its efficiency, partic-

Data	Total Words	Avg. Words	Total Expl. Words		Avg. Expl. Words	
			Ar	En	Ar	En
ArMeme						
Train	58,688	15	280,341	375,843	70	94
Dev	8,583	15	40,756	55,336	70	95
Test	16,653	15	79,360	105,476	70	93
Total	**83,924**	**15**	**400,457**	**536,655**	**70**	**94**
Hateful Meme						
Train	99,812	12	–	740,624	–	87
Dev	4,904	9	–	43,956	–	81
Test	18,079	9	–	173,982	–	87
Total	**122,795**	**10**	–	**958,562**	–	**85**

Table 3: Descriptive statistics of the dataset. *Total Words* and *Avg.* refer to the total and average number of words in the text. The last two columns represent the corresponding values for the explanations.

ularly with respect to dataset generation, model training, and system inference procedures.

In this research, we generate explanations for two different stages: (a) during existing dataset enhancement, which leverages an expert VLM (such as GPT) to generate high-quality explanations and (b) during training/inference with a smaller VLM (such as Llama-3.2 11b). Figure 1 illustrates these different stages. Mathematically, these two stages can be described by the functions $f(i, l) = e$ and $g(i) = (l, e)$, where e denotes the explanation, l is the label, and i is the input image or meme. Specifically, $f(i, l)$ returns an explanation e given both i and l, whereas $g(i)$ generates both the label l and the explanation e from only the input i.

This research enhances two existing datasets with explanations, see Section 3 and Table 3 for the details and statistics. For the explanation generation task, it first uses a VLM for $f(i, l)$ and then involves human experts, which significantly accelerates high-quality explanation generation and lowers the overall cost and time. The following subsections provide step-by-step details.

4.1 VLMs for Explanation Generation

Figure 1 illustrates an example of an Arabic meme along with its explanation-generation process using a VLM. We leverage GPT-4o (version 2024-11-20) for automated explanation generation. The choice of this model is motivated by prior studies Wang et al. (2023), which show that advanced GPT models can produce fluent, informative, persuasive, and logically sound explanations when properly prompted. In Listing 1, we present the *prompts* used for generating explanations for **ArMeme** and **Hateful Memes**. To refine the prompt, we iteratively tested several memes in both English and Arabic, selecting the one that produced the most reasonable explanations.

For Arabic memes, we generate two sets of explanations—one *in English* and one *in Arabic*. The motivation behind this approach is to assess the multilingual capability and quality of smaller VLMs, such as Llama-3.2 11b, in generating explanations and labels in both languages.

Size of the Explanation Determining the optimal length for the explanations is important to balance informativeness and cognitive load (Herm, 2023). Shen et al. (2022) explored the relationship between the length of the explanation and human understanding, finding that the shortest rationales are often ineffective. Recently, Wang et al. (2023) also studied the effect of explanation size and found that human evaluators are less willing to read longer explanations. To achieve an optimal balance, we iteratively tested various explanation lengths and ultimately set a limit of 100 words.

Model and its Parameters To utilize GPT-4o (OpenAI, 2023), we accessed the OpenAI API via Azure services. Although recently released o1 models have shown promising directions for complex reasoning, they were not accessible to us. For explanation generation, we used zero-shot learning. To ensure reproducibility, we set the temperature value to zero.

4.2 Manual Annotation of Explanation

Given that our idea is to use the generated explanation as gold data for further training and evaluation, therefore, we intended to go through human evaluation process. Following the prior studies (Wang et al., 2023; Huang et al., 2023; Agarwal et al., 2024) we adopted four metrics discussed below. For each metric we use 5-point Likert scale.

Informativeness. Measures the extent to which the explanation provides relevant and meaningful information for understanding the reasoning behind the label. A highly informative explanation offers detailed insights that directly contribute to the justification, while a low-informative explanation may be vague, incomplete, or lacking key details.

Clarity. Assesses how clearly the explanation conveys its meaning. A clear explanation is well-structured, concise, and easy to understand without requiring additional effort. It should be free from ambiguity, overly complex language, or poor phrasing that might hinder comprehension.

Plausibility. Refers to the extent to which an explanation logically supports the assigned label and appears reasonable given the meme's content. A plausible explanation should be coherent, factually consistent, and align with the expected reasoning behind the label.

Faithfulness. Measures how accurately an explanation reflects the reasoning behind the assigned label. A faithful explanation correctly represents the key factors and logical steps that justify the label, without adding misleading or unrelated details.

Annotation Setup. For manual annotation, we first prepared an annotation guideline and a platform (see Appendix B and A, respectively) for the annotators. For the Arabic memes, we recruited annotators who are native Arabic speakers and fluent in English, all holding at least a bachelor's degree. Because of their fluency, they also handled the hateful meme. We provided necessary training and consultation, and all had prior experience with similar tasks. A total of six annotators participated in the evaluation. In line with institutional requirements, each signed a Non-Disclosure Agreement (NDA), and a third-party company managed their compensation at standard hourly rates based on their location.

Annotation Agreement for Explanation. In Table 4, we summarize the annotation agreement scores of the explanations. We used 5-point Likert scale for various human evaluation metrics, including informativeness, clarity, plausibility, and faithfulness. We compute the average Likert scale value (from three annotators) for all evaluation metrics. We manually evaluated complete test sets for both ArMeme and Hateful meme except ArMeme with English explanation. For the later case we randomly selected 200 random examples. This decision was made to reduce the human annotation cost. The average agreement scores for the ArMeme dataset with Arabic explanations are > 4.5 out of 5 indicating high agreement in all evaluation metrics. However, for the English explanations of ArMeme, the faithfulness and plausibility scores are relatively lower. To better understand this issue, we plan to conduct further annotation on the complete test set of explanations. For the Hateful Memes dataset, the average Likert scale agreement scores range from 4.04 to 4.10 out of 5.

In addition, we also computed the annotation agreement on ordinal scales by adopting the agreement index $r^*_{wg(j)}$ (James et al., 1984), which compares observed variance in ratings to the maximum possible variance under complete disagreement. For each item, the agreement score is computed as: $r^*_{wg(j)} = 1 - \frac{S_X^2}{\sigma_{\text{mv}}^2}$ where S_X^2 is the observed variance across annotators and σ_{mv}^2 is the maximum variance possible given the scale (computed as $\sigma_{\text{mv}}^2 = 0.5(X_U^2 + X_L^2) - [0.5(X_U + X_L)]^2$, with $X_U = 5$ and $X_L = 1$ for a 5-point scale). The average agreement scores for ArMeme and Hateful memes are above 0.83 and 0.92, respectively, for all metrics. These values indicate strong agreement (O'Neill, 2017).

Dataset	Faithfulness	Clarity	Plausiblity	Informative
ArMeme Ar expl.	4.64	4.69	4.69	4.74
*ArMeme En expl.	3.91	4.50	3.81	4.13
Hateful meme	4.01	4.18	4.04	4.10

Table 4: *Average* Likert scale value for each human evaluation (annotation) metric across different sets of explanations.

4.3 Basic Statistics

Table 3 presents the basic statistics for both datasets. The average explanation length is 94 words for Arabic and 85 words for English. Notably, we instructed GPT-4o to generate explanations with fewer than 100 words.

5 Methodology

5.1 Instructions Dataset

Our approach follows the standard pipeline for aligning LLMs with user intentions and specific tasks through fine-tuning on representative data (Zhang et al., 2023; Kmainasi et al., 2024). This process typically involves curating and constructing instruction datasets that guide the model's behavior, ensuring it generates responses that align with the desired objectives. For our study, the responses include label and explanation. Hence, we created instruction format for both datasets. For the ArMeme dataset, we replicated the experiments for both Arabic and English explanations.

5.2 Model Selection

As shown in Figure 1, our first experimental phase involves model selection among several recent VLMs, including Llama-3.2 (11b) (Dubey et al., 2024), Paligemma 2 (3b) (Steiner et al., 2024), Qwen2-vl (Wang et al., 2024), and Pixtral (12b) (Agrawal et al., 2024).

We evaluate the base models in a zero-shot setting and fine-tune them using an instruction-following paradigm. The instructions prompt the model to generate responses in the format "`Label: (class_label)`". We use and a regex-based function to extract the predicted labels.

Note that this stage fine-tunes the models to predict class labels only, allowing us to verify whether they can handle multilingual inputs—especially in understanding Arabic text, cultural nuances, and image context. We do not ask the model to generate explanations here, as that is a more complex task and could affect their performance.

Based on the results reported in Tables 5 and 6, we selected Llama-3.2-vision-instruct (11b) for further training with explanations.

5.3 *Multi-Stage (MS)* Optimization Procedure

To emphasize our contribution, we introduce a novel optimization procedure to train VLM with *MemeXplain*, which decouples the classification and explanation generation tasks. While training both tasks in a single step may seem like an obvious choice, in practice, the two objectives produce *conflicting gradient signals* due to the fundamental differences between the learning objectives. Classification requires precise mapping of multimodal cues to discrete labels, whereas explanation generation demands fluency and coherence in free-form natural language. Merging them too early can compromise the model's performance on both tasks. Conversely, training them completely separately risks overwriting knowledge of one task while learning the other. To address these challenges, we propose a two-stage optimization procedure that *decouples* the learning objectives by first optimizing in the classification domain in *stage-1*, followed by augmentation through explanation generation in *stage-2*.

Problem Formulation. We formulate the joint optimization problem as follows: $L_{\text{total}} = L_{\text{classif}} + W_{\text{expl}} \cdot L_{\text{expl}}$, where L_{classif} and L_{expl} are the classification and explanation losses, respectively. W_{exp} is a step-function weight that switches from 0 to 1 between the stages. This formulation draws on the principles from *Domain Adaptation* and *Task-Incremental Learning* (Van de Ven et al., 2022) with the aim to: (a) *Isolate classification learning* and avoid conflicting updates and (b) *Prevent catastrophic forgetting* by gradually integrating the explanation objective.

Stage 1 - Classification Fine-Tuning. We set $W_{\text{exp}} = 0$ and optimize exclusively on L_{classif}. It adapts the pretrained VLM to the domain of hateful/propagandistic content, establishing a strong feature backbone for accurate label prediction.

Stage 2 - Joint Classification & Explanation Fine-Tuning. We set $W_{\text{exp}} = 1$ to optimize $L_{\text{classif}} + L_{\text{explanation}}$. After obtaining the domain-adapted backbone from Stage 1, the model proceeds to learn how to generate coherent, contextually grounded explanations alongside accurate classifications. This stepwise integration ensures that the model preserves its classification capabilities and avoids catastrophic forgetting, while developing proficiency in natural language reasoning.

By decoupling and then recombining these objectives in a straightforward two-phase procedure, our multi-stage (MS) optimization presents an easily implementable extension to standard VLM training pipelines. To validate its effectiveness, we compare it against a single-stage (SS) fine-tuning baseline, where the model is directly trained on the label-with-explanation dataset. Our ablation studies (detailed in Section 6) demonstrate that the proposed multi-stage approach significantly outperforms the single-stage strategy.

Model	Setup	Acc (%)	W-F1	M-F1
(Alam et al., 2024b)	Qarib	69.7	0.690	0.551
(Alam et al., 2024b)	mBERT	70.7	0.675	0.487
Llama-3.2 (11b)	Base	13.4	0.172	0.113
Llama-3.2 (11b)	FT	68.0	0.665	**0.452**
Paligemma2 (3b)	Base	15.3	0.090	0.080
Paligemma2 (3b)	FT	65.9	0.524	0.200
Qwen2 (7b)	Base	63.1	0.550	0.242
Qwen2 (7b)	FT	27.0	0.149	0.195
Pixtral (12b)	Base	14.6	0.177	0.133
Pixtral (12b)	FT	70.8	0.636	0.377

Table 5: Results for ArMeme. FT: Fine-tuned. Qarib (Abdelali et al., 2021) is a Arabic BERT (text only). mBERT - multilingual BERT (text only).

5.4 Training Setup

Our fine-tuning experiments utilize QLoRA (Dettmers et al., 2023), which combines INT4 quantization with parameter-efficient fine-tuning through Low-Rank Adaptation (LoRA) (Hu et al., 2022). In our setup, the base model is quantized to 4-bit precision, with

Model	Setup	Acc (%)	W-F1	M-F1
(Kiela et al., 2020)		69.47±2.06		
(Cao et al., 2022)		72.98±1.09		
Llama-3.2 (11b)	Base	66.1	0.650	0.618
Llama-3.2 (11b)	FT	77.7	0.770	**0.748**
Paligemma2 (3b)	Base	35.2	0.277	0.217
Paligemma2 (3b)	FT	69.2	0.664	0.623
Qwen2 (7b)	Base	66.4	0.669	0.442
Qwen2 (7b)	FT	77.9	0.773	0.753
Pixtral (12b)	Base	66.7	0.667	0.430
Pixtral (12b)	FT	77.2	0.766	0.746

Table 6: Results for Hateful meme. FT: Fine-tuned

LoRA updates applied to a subset of the model parameters. This approach was selected to address computational and memory resource constraints.

We adapted all relevant submodules (vision, language, attention, and MLP layers) with a LoRA rank of 16, an alpha of 16, and no dropout. For training, we used a per-device batch size of 2 with gradient accumulation over 4 steps, optimizing with AdamW at a learning rate of 2×10^{-4}, a weight decay of 0.01, and a linear scheduler with 5 warmup steps. For the second stage experiments (label-with-explanation), the learning rate was reduced to 1×10^{-5}. All experiments were run for a total of 5 epochs for fine-tuning. However, to ensure a fair comparison, we allocated 3 epochs to Stage 1 and 2 epochs to Stage 2 in the MS setup.

5.5 Evaluation Setup and Metrics

We train the models using the training set, fine-tune the parameters with the development set, and evaluate their performance on the test set as reported in Tables 5 and 6. For performance measurement across different experimental settings, we compute accuracy, weighted F_1, and macro-F_1. We evaluate the model's explanation performance on the test set using semantic similarity-based metric, measured by the F_1 score within BERTScore (Zhang et al., 2020). This score is computed using contextual embeddings extracted from pre-trained BERT models. To enhance accuracy, we utilize language-specific transformer models for embedding extraction. For Arabic we use AraBERT (v2) (Antoun et al., 2020) model and for English we use bert-base-uncased model (Devlin et al., 2019). In addition, we also compute BLEU and METEOR scores.

6 Experimental Results and Discussion

This section first presents competitive results among our proposed method and the state-of-the-art approaches. Next, it briefly analyzes and investigates the proposed method to validate and highlight the core contributions of this research.

Model	Setup	Acc(%)	W-F1	M-F1
ArMeme				
(Alam et al., 2024b)	Qarib	69.7	0.690	0.551
(Alam et al., 2024b)	mBERT	70.7	0.675	0.487
(Alam et al., 2024b)	ResNet50	66.0	0.637	0.434
Llama MS	FT	72.1	0.699	0.536
Llama (Ar-Exp) MS	FT	72.0	0.696	0.499
Hateful Meme				
(Kiela et al., 2020)		69.47±2.06		
(Cao et al., 2022)		72.98±1.09		
Llama MS	FT	79.9	0.802	0.792

Table 7: Comparison with SOTA and our results (*Llama MS and Llama (Ar-Exp) MS*). ResNet50 (He et al., 2016) is an image only model. MS: Multi-stage *(Our approach)*.

Comparison with State-of-the-Art. Table 7 presents the comparison. On the ArMeme dataset, it achieves the best accuracy at 72.1% and the best weighted F1 at 0.699, with Qarib and mBERT following behind. Although Qarib attains the highest macro F1 (0.551), our model remains competitive with a macro F1 of 0.536. Importantly, our method stands out because it provides explanations that add significant value. On the Hateful Meme dataset, our approach clearly outperforms the state-of-the-art by achieving the best performance with an accuracy of 79.9%, a weighted F1 of 0.802, and a macro F1 of 0.792. These results clearly highlight the advantages of our explainability-enhanced dataset and our proposed optimization procedure for both classification and explanation-generation tasks.

Model	Setup	Acc (%)	W-F1	M-F1	BS	BL	M
ArMeme							
Llama	Base	12.7	0.165	0.105	0.61	0.24	0.17
Llama SS	FT	68.2	0.584	0.257	0.70	0.56	0.36
Llama MS	FT	72.1	0.699	0.536	0.70	0.57	0.35
Llama Ar-Exp	Base	19.0	0.246	0.125	0.58	0.12	0.09
Llama MS Ar-Exp	FT	72.0	0.696	0.499	0.72	0.55	0.29
Hateful Meme							
Llama	Base	65.2	0.615	0.567	0.661	0.35	0.23
Llama SS	FT	75.9	0.760	0.745	0.767	0.65	0.47
Llama MS	FT	79.9	0.802	0.792	0.777	0.67	0.49

Table 8: Results with ArMeme and Hateful meme classification and explanation generation. Llama: Llama-3.2 (11b), BS: BERTScore, BL: BLEU, M: METEOR. SS: Single-stage, MS: Multi-stage. Ar-Exp: Model trained with Arabic explanation. W-F1: Weighted F1, M-F1: macro-F1

Ablation Study on Different Model Training Settings. Table 8 provides classification and explanation-generation results on the *ArMeme* and *Hateful Meme* datasets from several perspectives:
(a) **Base vs. FT**: demonstrates the performance difference between the same model with and without fine-tuning (FT);
(b) **Single-stage (SS) vs. Multi-stage (MS)**: highlights the necessity and benefits of the proposed optimization procedure, and
(c) **Eng-Exp vs. Ar-Exp**: showcases the multilingual capability of the selected VLM.

First, we compare the **Base vs. FT** setup, from which it is evident that the FT model significantly outperforms the baseline. For example, on the ArMeme dataset, while the baseline achieves an accuracy of 12.7%, the proposed fine-tuning boosts it to 72.1%. Similarly, on the Hateful Meme dataset, fine-tuning improves the base accuracy from 65.2% to 79.9%. We observe similar improvements in the F1 metrics for classification and BERTScore for explanation quality. These significant performance gains *validate our approach of fine-tuning the base models with the explainability enhanced dataset*, demonstrating its efficacy for the meme classification and explanation generation tasks.

Next, we compare the **SS vs. MS** setup, which reveals that multi-stage (MS) fine-tuning further enhances performance over the single-stage (SS) approach. For example, on the ArMeme dataset, the accuracy increased from 68.2% to 72.1%, the weighted F1 increased from 0.584 to 0.699, the macro F1 increased significantly from 0.257 to 0.536, and the BERTScore for Arabic explanation increased significantly from 0.58 to 0.72. A similar trend is observed on the Hateful Meme dataset, where additional fine-tuning iterations yield more robust classification (approximately 4% improvement) and enhanced explanation quality. These performance gains *validate our proposed multi-stage optimization procedure* to further refine the VLMs.

Assessment of Multilingual Capability. We compare ***Llama MS - FT*** with ***Llama MS Ar-Exp*** in Table 8, which shows that fine-tuning using explanations generated in both languages yields comparable outcomes. This *validates the multilingual capability of our empirically chosen VLM* for the target task and helps users understand multilingual content without fluency in that language. For example, our model allows an English speaker to analyze Arabic memes and receive explanations in English.

Error Analysis. To better understand the model's reasoning capabilities and failure modes, we conduct an error analysis contrasting correct and incorrect label predictions, as well as comparing single-stage and multistage training paradigms. Our analysis reveals that while the model often aligns well with human explanations in correctly classified instances, it struggles with nuanced cases involving humor or implicit context. Furthermore, multistage training consistently outperforms single-stage training in label accuracy by better integrating textual and visual cues while providing explanation. See Section E for detailed examples and discussion.

7 Conclusions and Future Work

In this study, we introduce a *MemeXplain* dataset for propagandistic and hateful meme detection and natural explanation generation, making it the *first* resource of its kind. To address both detection and explanation generation tasks and ensure efficient VLMs model training on this dataset, we also propose a novel multi-stage optimization procedure. To evaluate the multilingual capability of the model, we developed Arabic and English explanations for Arabic memes. The inclusion of English explanations benefits non-Arabic speakers, whereas providing explanations in the native language ensures that cultural nuances are accurately conveyed. With our novel multi-stage training procedure, we demonstrate improved detection performance for both *ArMeme* and *hateful* memes. The higher performance of explanation generation further demonstrates the efficacy of our multi-stage training approach.

Moreover, the proposed multi-stage scheme is agnostic to specific VLM architectures and applicable in contexts where tasks impose divergent gradient demands, thereby offering a versatile framework for multi-task vision–language learning.

We foresee several future directions to extend this research and explore the following: *(a)* training the model with additional data through data augmentation, which could help it become an instruction-generalized model and potentially enhance its performance further; *(b)* incorporating pseudo and self-labeled data using an active learning procedure to incrementally improve the model's capabilities; and *(c)* developing a task-generalized model that addresses multiple tasks.

8 Limitations

Due to the complex nature of manual explanation creation, we have relied on GPT-4o for explanation generation. To ensure the reliability of the explanation we have manually evaluated in four criteria such as informativeness, clarity, plausiblity, and faithfulness on a small sample for each set of explanation. The preliminary evaluation scores suggest that we can rely on the gold explanation as the reference. As a part of ongoing work we plan to conduct manual evaluation on a larger set. An important aspect of the ArMeme dataset is that it is highly imbalanced, which affects overall performance. One possible approach to address this issue is to increase the number of memes labeled as propaganda, other, and not-meme. This can be achieved through data augmentation or by collecting additional memes.

Ethics and Broader Impact

We extended existing datasets by adding explanations. To the best of our knowledge, the dataset does not contain any personally identifiable information, making privacy risks nonexistent. Regarding the explanations, we provided clear annotation instructions and cautioned annotators that some memes might be offensive. It is important to note that annotations are inherently subjective, which can introduce biases into the overall evaluation results. We encourage researchers and users of this dataset to remain critical when developing models or conducting further research. Models built using this dataset could be highly valuable for fact-checkers, journalists, and social media platforms.

References

Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. 2021. Pre-training bert on arabic tweets: Practical considerations.

Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. 2024. Faithfulness vs. plausibility: On the (un)reliability of explanations from large language models. *arXiv preprint arXiv:2402.04614*, 2402.04614.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. 2024. Pixtral 12b. *arXiv preprint arXiv:2410.07073*.

Firoj Alam, Md Rafiul Biswas, Uzair Shah, Wajdi Zaghouani, and Georgios Mikros. 2024a. Propaganda to hate: A multimodal analysis of arabic memes with multi-agent llms. In *International Conference on Web Information Systems Engineering*, pages 380–390. Springer.

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

Firoj Alam, Abul Hasnat, Fatema Ahmad, Md. Arid Hasan, and Maram Hasanain. 2024b. ArMeme: Propagandistic content in Arabic memes. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21071–21090, Miami, Florida, USA. Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15.

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

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*, MM '23, page 5244–5252, New York, NY, USA. Association for Computing Machinery.

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

Pengyuan Chen, Lei Zhao, Yangheran Piao, Hongwei Ding, and Xiaohui Cui. 2024. Multimodal visual-textual object graph attention network for propaganda detection in memes. *Multimedia Tools and Applications*, 83(12):36629–36644.

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)*, Hong Kong, China.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

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*, NAACL-HLT '19, Minneapolis, Minnesota, USA.

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 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

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 *ACL-IJCNLP*.

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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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

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

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

Maram Hasanain, Md. Arid Hasan, Fatema Ahmed, Reem Suwaileh, Md. Rafiul Biswas, Wajdi Zaghouani, and Firoj Alam. 2024. 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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, CVPR '16, pages 770–778. IEEE.

Ming Shan Hee, Wen-Haw Chong, and Roy Ka-Wei Lee. 2023. Decoding the underlying meaning of multimodal hateful memes. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 5995–6003.

Ming Shan Hee, Roy Ka-Wei Lee, and Wen-Haw Chong. 2022. On explaining multimodal hateful meme detection models. In *Proceedings of the ACM Web Conference 2022*, pages 3651–3655.

Lukas-Valentin Herm. 2023. Impact of explainable ai on cognitive load: Insights from an empirical study. *arXiv preprint arXiv:2304.08861*.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. MUTE: A multimodal dataset for detecting hateful memes. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 32–39, Online. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Chain of explanation: New prompting method to generate quality natural language explanation for implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*, pages 90–93.

Lawrence R James, Robert G Demaree, and Gerrit Wolf. 1984. Estimating within-group interrater reliability with and without response bias. *Journal of applied psychology*, 69(1):85.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624.

Mohamed Bayan Kmainasi, Ali Ezzat Shahroor, Maram Hasanain, Sahinur Rahman Laskar, Naeemul Hassan, and Firoj Alam. 2024. LlamaLens: Specialized multilingual llm for analyzing news and social media content. *arXiv preprint arXiv:2410.15308*.

Gokul Karthik Kumar and Karthik Nandakumar. 2022. Hate-CLIPper: Multimodal hateful meme classification based on cross-modal interaction of clip features. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 171–183.

Gitanjali Kumari, Kirtan Jain, and Asif Ekbal. 2024. M3hop-cot: Misogynous meme identification with multimodal multi-hop chain-of-thought. *arXiv preprint arXiv:2410.09220*.

Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, et al. 2022. Explanations from large language models make small reasoners better. *arXiv preprint arXiv:2210.06726*.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2022. Teaching small language models to reason. *arXiv preprint arXiv:2212.08410*.

Palash Nandi, Shivam Sharma, and Tanmoy Chakraborty. 2024. SAFE-MEME: Structured reasoning framework for robust hate speech detection in memes. *arXiv preprint arXiv:2412.20541*.

Thomas A O'Neill. 2017. An overview of interrater agreement on likert scales for researchers and practitioners. *Frontiers in psychology*, 8:777.

OpenAI. 2023. GPT-4 technical report. Technical report, OpenAI.

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

Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Back. 2024. Reasoning with large language models, a survey. *arXiv e-prints*, pages arXiv–2407.

Uzair Shah, Md. Rafiul Biswas, Marco Agus, Mowafa Househ, and Wajdi Zaghouani. 2024. MemeMind at ArAIEval shared task: Generative augmentation and feature fusion for multimodal propaganda detection in Arabic memes through advanced language and vision models. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 467–472, Bangkok, Thailand. 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, Vienna, Austria. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Hua Shen, Tongshuang Wu, Wenbo Guo, and Ting-Hao Huang. 2022. Are shortest rationales the best explanations for human understanding? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 10–19.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, et al. 2024. Paligemma 2: A family of versatile vlms for transfer. *arXiv preprint arXiv:2412.03555*.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8990–9005. Association for Computational Linguistics.

Gido M Van de Ven, Tinne Tuytelaars, and Andreas S Tolias. 2022. Three types of incremental learning. *Nature Machine Intelligence*, 4(12):1185–1197.

Riza Velioglu and Jewgeni Rose. 2020. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge. *arXiv preprint arXiv:2012.12975*.

Han Wang, Ming Shan Hee, Md Rabiul Awal, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. 2023. Evaluating GPT-3 generated explanations for hateful content moderation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 6255–6263.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se-Young Yun. 2023. HARE: Explainable hate speech detection with step-by-step reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5490–5505.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

A Annotation Guideline

You will be shown a meme, a label assigned to it, and an explanation for the assigned label. As an annotator, your task is to carefully examine each meme, label, and explanation. Then assess the quality of the explanation provided for the assigned label. Follow the steps below to ensure a thorough evaluation:

Analyze the Meme

- Observe the image and read the accompanying text.
- Understand the overall message and the potential implications of the meme.

Check the Assigned Label

- Check the given label. The label is the result of annotation done by multiple human annotators.

Evaluate the Explanation

- Read the explanation provided for why the meme has been assigned its label.
- Assess the explanation based on the metrics below. Each metric is scored on a Likert scale from 1-5.

Kindly note that to evaluate the explanation, you do not have to agree or disagree with the given label.

A.1 Metrics

A.1.1 Informativeness

Measures the extent to which the explanation provides relevant and meaningful information for understanding the reasoning behind the label. A highly informative explanation offers detailed insights that directly contribute to the justification, while a low-informative explanation may be vague, incomplete, or lacking key details.

As an annotator, you are judging if the explanation provides enough information to explain the label assigned to the meme.

- 1 = Not informative: The explanation lacks relevant details and does not help understand why the meme is labeled as such.
- 2 = Slightly informative: The explanation provides minimal information, but key details are missing or unclear.
- 3 = Moderately informative: The explanation contains some useful details but lacks depth or supporting reasoning.
- 4 = Informative: The explanation is well-detailed, providing a clear and meaningful justification for the label.
- 5 = Very informative: The explanation is thorough, insightful, and fully justifies the label with strong supporting details.

A.1.2 Clarity

Assesses how clearly the explanation conveys its meaning. A clear explanation is well-structured, concise, and easy to understand without requiring additional effort. It should be free from ambiguity, overly complex language, or poor phrasing that might hinder comprehension.

As an annotator, you are judging the language and structure of the explanation. Spelling mistakes, awkward use of language, and incorrect translations will negatively impact this metric.

- 1 = Very unclear: The explanation is confusing, vague, or difficult to understand.
- 2 = Somewhat unclear: The explanation has some clarity but includes ambiguous or poorly structured statements.
- 3 = Neutral: The explanation is somewhat clear but may require effort to fully grasp.
- 4 = Clear: The explanation is well-structured and easy to understand with minimal ambiguity.
- 5 = Very clear: The explanation is highly readable, precise, and effortlessly understandable.

A.1.3 Plausibility

Refers to the extent to which an explanation logically supports the assigned label and appears reasonable given the meme's content. A plausible explanation should be coherent, factually consistent, and align with the expected reasoning behind the label. While it does not require absolute correctness, it should not contain obvious contradictions or illogical claims.

As an annotator, you are judging if the explanation actually supports the label assigned to the meme. For example, if a meme is labeled as Not Propaganda, the explanation given should justify that label.

- 1 = Not plausible at all: The explanation does not align with the label and seems completely incorrect.
- 2 = Weakly plausible: The explanation has some relevance but lacks strong justification or contains logical inconsistencies.

- 3 = Moderately plausible: The explanation somewhat supports the label but may be incomplete or partially flawed.
- 4 = Plausible: The explanation logically supports the label and is mostly reasonable.
- 5 = Highly plausible: The explanation is fully aligned with the label and presents a strong, logical justification.

A.1.4 Faithfulness

Measures how accurately an explanation reflects the reasoning behind the assigned label. A faithful explanation correctly represents the key factors and logical steps that justify the label, without adding misleading or unrelated details. High faithfulness means the explanation stays true to the actual reasoning used for classification, ensuring reliability and consistency.

As an annotator, you are judging how well the explanation reflects the logic behind the label. For example, if the explanation claims an implication of the meme, it should also present the logical reasoning behind it.

- 1 = Not faithful at all: The explanation is completely unrelated to the given label and does not reflect a valid reasoning process.
- 2 = Weakly faithful: Some elements of the explanation are relevant, but much of it is misleading, inconsistent, or lacks proper justification.
- 3 = Moderately faithful: The explanation captures parts of the reasoning but includes unrelated, unclear, or unnecessary justifications.
- 4 = Faithful: The explanation aligns well with the reasoning behind the label and includes relevant, logical details.
- 5 = Highly faithful: The explanation fully and accurately reflects the correct reasoning, without any misleading or irrelevant information.

B Annotation Platform

In Figure 2, we present the screenshot of the interface designed for the explanation evaluation of hateful meme, which consisted of an image, respective label, and explanation for the label, annotation guidelines, and four different evaluation metrics. We used 5-point Likert scale for each evaluation metric. Annotators select one of the Likert scale value following the annotation guideline for each metric and submit.

C Prompt for Explanation Generation

In Listings 1 and 2, we provide the prompts used to generate explanations for ArMeme and Hateful Meme. The prompt in Listing 1 is specifically for generating Arabic explanations for ArMeme. To generate English explanations, the same prompt was used, except it was adapted for English.

```
You are a Propaganda Image Detection Expert. A
    human expert has already classified the
    image as {class_label}.
Do not change or re-identify the classified
    label of the image.
First, analyze the image's visual elements (
    objects, symbols, color usage, composition)
and provide a concise description. Next, read
    and paraphrase any text in the image-
    especially if
it is in non-Arabic-so that a Arabic speaker
    could understand its literal or intended
    meaning.
Explain how that text, in conjunction with the
    visual elements, supports or illustrates the
     known
propaganda technique. Think carefully about
    emotional appeals, humor, cultural
    references, manipulative
language, or other rhetorical strategies.
Finally, provide a thoughtful explanation in
    Arabic (up to {explanation_length} words) of
     why these elements led the human to
    classify
the image as propaganda. Be specific about how
    the text and visuals interact, referencing
    the
Arabic context or cultural nuances if relevant.
    Your answer must be valid JSON with a single
     field:
{{
"explanation": "Your explanation here (up to {
    explanation_length} words) in Arabic."
}}
```

Listing 1: Prompt for generating Arabic explanation for ArMeme dataset. The place-holders refers to label, and number of words, respectively.

```
You are a Hateful Image Detection Expert. A
    human expert has already classified the
    image as {}. Do not change or re-identify
    the classified label of the image.
First, analyze the image's visual elements (
    objects, symbols, color usage, composition)
    and provide a concise description. Next,
    read and paraphrase any text in the image so
     anyone speaking English can understand its
    literal or intended meaning.
Explain how that text, in conjunction with the
    visual elements, supports or illustrates the
     known hateful content.
Finally, provide a thoughtful explanation in
    English (up to {} words) of why these
    elements led the human to classify the image
     as hateful. Be specific about how the text
    and visuals interact, referencing the
```

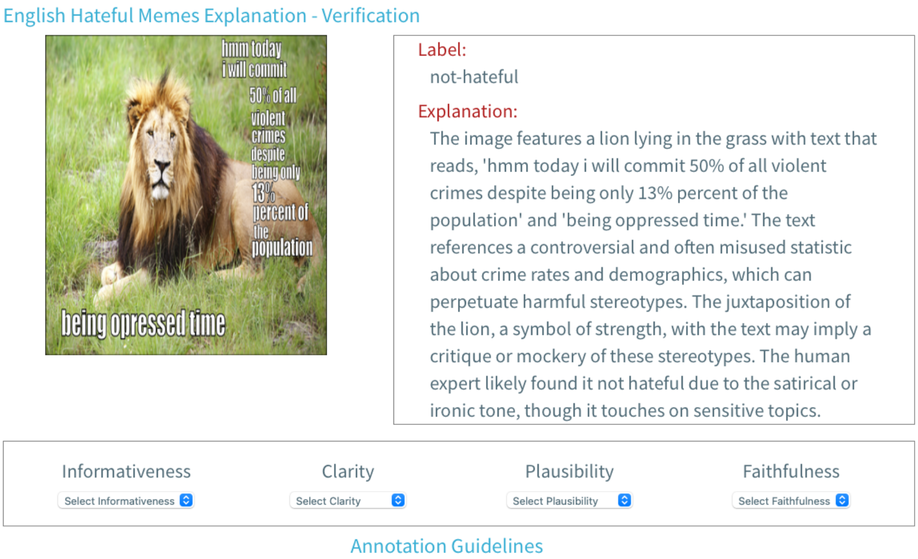


Figure 2: A screenshot of the annotation platform for the explanation evaluation of hateful meme.

```
    context or cultural nuances if relevant.
    Your answer must be valid JSON with a single
     field:
{{
"explanation": "Your explanation here (up to {}
    words) in English."
}}
```

Listing 2: Prompt for generating explanation. The place-holders refers to label, and number of words, respectively.

D Annotation Agreement: Additional Analysis

We report human evaluations of the generated explanations from two models: (i) Llama MS (ArMeme) and (ii) Llama MS (Hateful Memes), as detailed in Table 9. Due to resource constraints, the manual evaluation was conducted on a representative sample of 100 instances. For the ArMeme dataset, the scores across all evaluation metrics range from 4.15 to 4.74 out of 5, while for the Hateful Memes dataset, the scores range from 4.41 to 4.54, indicating consistently strong performance across both datasets.

Dataset	Faithfulness	Clarity	Plausiblity	Informative
ArMeme	4.63	4.74	4.56	4.15
Hateful meme	4.41	4.44	4.43	4.54

Table 9: *Average* Likert scale value for each human evaluation (annotation) metric for the explanations of different datasets.

E Error Analysis

E.1 Correct vs. Incorrect Label Prediction

In Figure 3, we present examples from the Hateful Memes dataset, showcasing cases where the model made both correct and incorrect predictions.

In Figure 3a, the Gold explanation describes the image as reinforcing a harmful racial stereotype by juxtaposing a joyful scene of Asian individuals eating with offensive text. The Predicted explanation correctly identifies the derogatory language and its racist implications, aligning with the gold annotation. The model's BERT-F1 score of 0.873 shows the high confidence in associating textual and visual elements to detect hate speech effectively.

In Figure 3, the Gold explanation interprets the image as a humorous juxtaposition, using wordplay between nationality and species without targeting any group. However, the Predicted explanation classifies it as hateful. This missclassification suggests that the model struggled to distinguish linguistic humor from implicit hate speech, as reflected in its BERT-F1 score of 0.6259. This highlights the challenge of detecting context-dependent content, where intent and interpretation play a crucial role in classification.

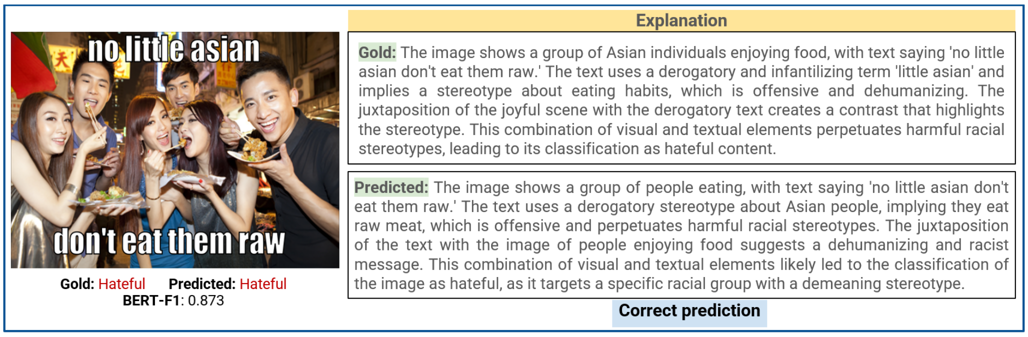
E.2 Effect of Single vs. Multistage Training

In Figure 4 we notice that multi-stage model correctly predicts the label and generates a faithful explanation, identifying how the combination of the image and text reinforces gender stereotypes. In contrast, the single-stage model label the meme with incorrect label – not-propaganda. Figure 5 shows another example from the hateful memes dataset. The multi-stage model correctly labels the image as hateful and grounds its explanation in the offensive combination of the *swastika* and *the mocking caption*. Whereas the single-stage model misclassifies the image as not-hateful. This highlights the effectiveness of multistage optimization in aligning predictions with grounded reasoning.

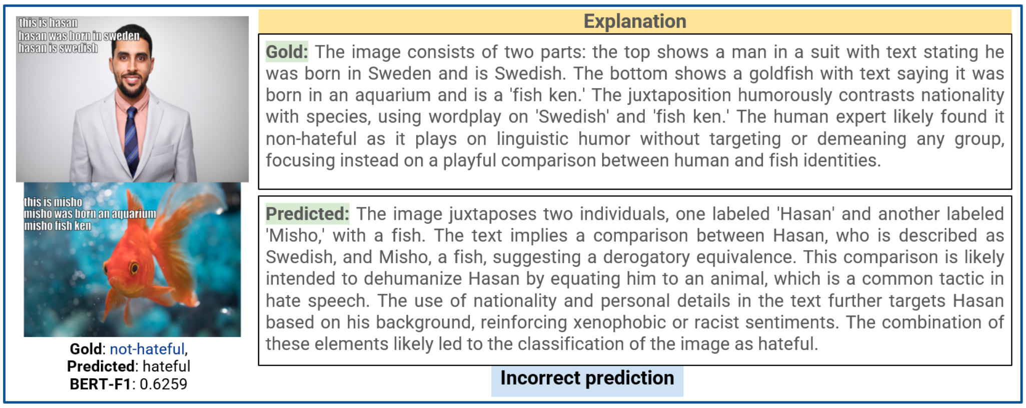
F Data Release

The *MemeXplain* dataset[2] will be released under the CC BY-NC-SA 4.0 – Creative Commons Attribution 4.0 International License: https://creativecommons.org/licenses/by-nc-sa/4.0/.

[2] anonymous.com



(a) Correct prediction.



(b) Incorrect prediction.

Figure 3: Example of correct and incorrect label prediction with explanation.

Figure 4: Single- vs. Multi-Stage prediction for *ArMeme*.

Ground Truth

Label: Hateful

Explanation: The image shows a person holding a sign with a swastika surrounded by stars, resembling the EU flag, which is a provocative and offensive symbol. The text 'when you find out spongebob is jewish' is nonsensical and derogatory, using a popular cartoon character to mock Jewish identity. The combination of a hate symbol and the text trivializes and targets Jewish people, contributing to antisemitic stereotypes. This interaction of visuals and text promotes hate, leading to its classification as hateful content due to its offensive and harmful implications.

Single Stage

Label: Not-hateful

Multi-Stage

Label: Hateful

Explanation: The image shows a person holding a flag with a swastika, a symbol associated with Nazi Germany and anti-Semitism, at a public gathering. The text 'when you find out spongebob is jewish' is a play on words, using a popular cartoon character to make a false and offensive claim about Jewish identity. The combination of the swastika and the text implies a hateful message, as it trivializes and mocks Jewish identity by associating it with a fictional character in a derogatory manner. This interaction of text and imagery is why the image is classified as hateful.

Figure 5: Single- vs. Multi-Stage prediction for *hateful memes*.