ArMeme: Propagandistic Content in Arabic Memes

Anonymous ACL submission

This paper contains content that may be sensitive or disturbing to some readers.

Abstract

With the rise of digital communication memes have become a significant medium for cultural and political expression that is often used to mislead audience. Identification of such misleading and persuasive multimodal content become more important among various stakeholders, including social media platforms, policymakers, and the broader society as they often cause harm to the individuals, organizations and/or society. While there has been effort to develop AI based automatic system for resource rich languages (e.g., English), it is relatively little to none for medium to low resource languages. In this study, we focused on developing an Arabic memes dataset with manual annotations of propagandistic content.[1] We annotated $\sim 6K$ Arabic memes collected from various social media platforms, which is a first resource for Arabic multimodal research. We provide a comprehensive analysis aiming to develop computational tools for their detection. We will make them publicly available for the community.

Figure 1: Examples of Arabic memes representing different categories.

1 Introduction

Social media platforms have enabled people to post and share content online. A significant portion of this content provides valuable resources for initiatives such as citizen journalism, raising public awareness, and supporting political campaigns. However, a considerable amount is posted and shared to mislead social media users and to achieve social, economic, or political agendas. In addition, the freedom to post and share content online has facilitated negative uses, leading to an increase in online hostility, as evidenced by the spread of disinformation, hate speech, propaganda, and cyberbullying (Brooke, 2019; Joksimovic et al., 2019; Schmidt and Wiegand, 2017; Davidson et al., 2017; Da San Martino et al., 2019a; Van Hee et al., 2015). A lack of *media literacy*[2] is also a major factor contributing to the spread of misleading information on social media (Zannu et al., 2024). This can lead to the uncritical acceptance and sharing of false or misleading content, which can quickly disseminate through social networks. In their study, Zannu et al. (2024) highlight the crucial role of media literacy in mitigating the spread of fake news among users of platforms such as Instagram and Twitter.

Online content typically consists of different modalities, including text, images, and videos. Disinformation, misinformation, propaganda, and other harmful content are shared across all these modalities. Recently, the use of *Internet memes*

[1] Propaganda is a form of communication designed to influence people's opinions or actions toward a specific goal, employing well-defined rhetorical and psychological techniques(for Propaganda Analysis, 1938).

[2] Media literacy encompasses the ability to access, analyze, evaluate, and create media in various forms.

have become very popular on these platforms. A meme is defined as "a collection of digital items that share common characteristics in content, form, or stance, which are created through association and widely circulated, imitated, or transformed over the Internet by numerous users" (Shifman, 2013). Memes typically consist of one or more images accompanied by textual content (Shifman, 2013; Suryawanshi et al., 2020). While memes are primarily intended for humor, they can also convey persuasive narratives or content that may mislead audiences. To automatically identify such content, research efforts have focused on addressing offensive material (Gandhi et al., 2020), identifying hate speech across different modalities (Gomez et al., 2020; Wu and Bhandary, 2020), and detecting propaganda techniques in memes (Dimitrov et al., 2021a).

Among the various types of misleading and harmful content, the spread of propagandistic content can significantly distort public perception and hinder informed decision-making. To address this challenge, research efforts have been specifically directed towards defining techniques and tackling the issue in different types of content, including news articles (Da San Martino et al., 2019), tweets (Alam et al., 2022b), memes (Dimitrov et al., 2021a), and textual content in multiple languages (Piskorski et al., 2023a). Most of these efforts have focused on English, with relatively little attention given to Arabic. Prior research on Arabic textual content includes studies presented at WANLP-2022 and ArabicNLP-2023 (Alam et al., 2022b; Hasanain et al., 2023). However, for multimodal content, specifically memes, there are no available datasets or resources. To address this gap, we have collected and annotated a dataset consisting of approximately 6,000 memes, categorizing them into four categories (as shown in Figure 1) to identify propagandistic content. Below we briefly summarize the contribution of our work.

- The first Arabic meme dataset with manual annotations defining four categories.
- A detailed description of the data collection procedure, which can assist the community in future data collection efforts.
- An annotation guideline that will serve as a foundation for future research.
- Detailed experimental results, including:
  - Text modality: training classical models and fine-tuning monolingual vs. multilingual transformer models.
  - Image modality: fine-tuning CNN models with different architectures.
  - Multimodality: training an early fusion-based model.
  - Evaluating different LLMs in a zero-shot setup for all modalities.
- Releasing the dataset to the community.[3] The dataset and annotation guideline will be beneficial for research to develop automatic systems and enhance media literacy.

2 Related Work

The widespread use of social media has become one of the main ways of sharing information and is also responsible for creating and spreading misinformation and propaganda among users. Propagandistic techniques often utilize various types of content, such as fake news and doctored images, across multiple media platforms, frequently employing tools like bots. This information is distributed in diverse forms, including textual, visual, and multi-modal. To mitigate the impact of propaganda in online media, researchers have been developing resources and tools to identify and debunk such content.

2.1 Persuasion Techniques Detection

Early research on propaganda identification relies on the entire document to identify whether the content is propaganda, while recent studies focus on social media content (Dimitrov et al., 2021b), news articles (Da San Martino et al., 2019b), political speech (Partington and Taylor, 2017), arguments (Habernal et al., 2017, 2018), and multimodal content (Dimitrov et al., 2021a). Barrón-Cedeno et al. (2019) developed a binary classification (*propaganda* and *non-propaganda*) corpus to explore writing style and readability levels. An alternative approach followed by Habernal et al. (2017, 2018) to identify persuasion techniques within the texts constructing a corpus on arguments. Moreover, the study of Da San Martino et al. (2019b) developed a span-level propaganda detection corpus from news articles and annotated in eighteen propaganda techniques.

[3]Dataset will be released under CC-BY-NC-SA through `https://anonymous.com`.

Piskorski et al. (2023b) developed a dataset from online news articles into twenty-two persuasion techniques containing nine languages to address the multilingual research gap. Following the previous work, Piskorski et al. (2023a) and SemEval-2024 task 4 focus on resource development to facilitate the detection of multilingual persuasion techniques. Focusing on multimodal persuasion techniques for memes, Dimitrov et al. (2021a) created a corpus containing 950 memes and investigated pretrained models for both unimodal and multimodal memes. The study of Chen et al. (2024) proposed a multimodal visual-textual object graph attention network to detect persuasion techniques from multimodal content using the dataset described in (Piskorski et al., 2023b). In a recent shared task, Dimitrov et al. (2024) introduced a multilingual and multimodal propaganda detection task, which attracted many participants. The participants' systems included various models based on transformers, CNNs, and LLMs.

2.2 Multimodal Content

The study of multimodal content has gained popularity among researchers for propaganda detection due to the effectiveness of multimodal content in spreading propaganda information and creating positive impacts among the targeted audience. Sharma et al. (2022) presented propaganda can be used to cause several types of harm including hate, violence, exploitation, etc. while spreading mis- and dis-information is also one of the main reasons (Alam et al., 2022a). The study of Volkova et al. (2019) presented an in-depth analysis of multimodal content for predicting misleading information from news. Additionally, the deception and disinformation analysis on social media platforms using multimodal content in multilingual settings has been studied by Glenski et al. (2019). Moreover, hateful memes (Kiela et al., 2020), propaganda in visual content (Seo, 2014), emotions and propaganda (Abd Kadir et al., 2016) also studied by the researchers in the past few years.

Recent studies focusing on fine-tuning visual transformer models such as ViLBERT (Lu et al., 2019), Multimodal Bitransformers (Kiela et al., 2019), and VisualBERT (Li et al., 2019). Cao et al. (2022) study focuses on multimodal hateful meme identification using prompting strategies by adopting (Prakash et al., 2023). Hee et al. (2024) studied hate speech content moderation and discussed recent advancements leveraging large models.

Compared to previous studies, our work differs in that we provide the first resource for Arabic. Additionally, our annotation guidelines and data collection procedures for memes may be useful for other languages.

3 Dataset

3.1 Data Collection

Our data collection process involve several steps as highlighted in the Figure 2. We manually selected public groups from Facebook, Instagram, and Pinterest. In addition, we have also collected memes from Twitter using a set of keywords as listed in the Figure 3 (in Appendix). Our data curation consists of a series of steps as discussed below.

Manual selection of groups, links and keywords: Focusing on the mentioned sources we have manually selected public groups, which contains post on public figures, celebrity, and mentions about politics. In Table 1, we provide the sources of the dataset, number of groups and number of image we have collected.

Source	# of Group	# of Images
Facebook	19	5,453
Instagram	22	107,307
Pinterest	-	11,369
Twitter	-	5,369
Total		**129,498**

Table 1: Statistics of the initial data collection.

Crawling: Given that Facebook, Instagram and Pinterest do provide API or do not allow automatic crawling images, therefore, we developed a semi-automatic approach to crawl images from these platforms. The steps include manually loading images and then crawl the images that are loaded on the browser. For the Twitter (X-platform), we used the keywords to crawl tweets, which consists of media/image.

3.2 Filtering

Filtering duplicate images: Given that user might have posted same meme or a slight modification of it in multiple platforms, which is very common for social media, therefore, we applied an exact and near-duplicate image detection method

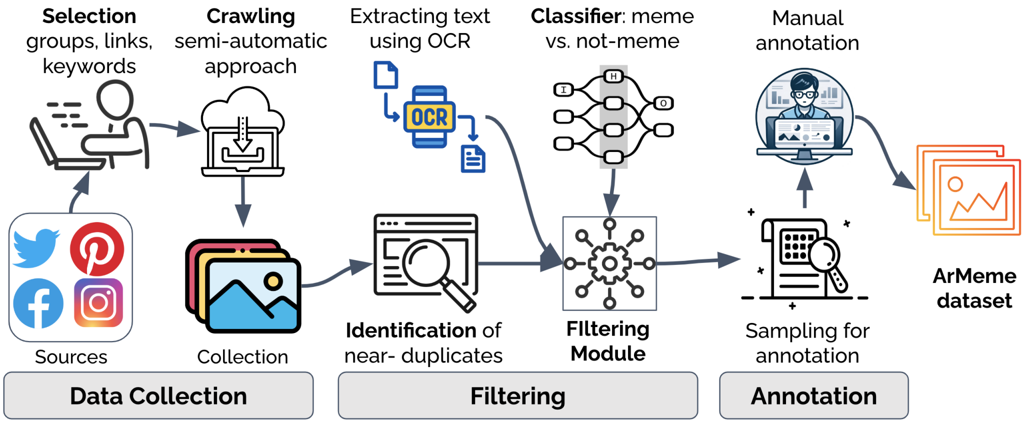

Figure 2: Data curation pipeline.

to remove them. This method consists of extracting features using a pre-trained deep learning model and compute similarity. Given a dataset $\mathcal{D} = \{x_1, x_2, \ldots, x_N\}$ consisting of N data points, we extracted features using a pre-trained deep learning model and used nearest neighbor based approach (Cunningham and Delany, 2007). The model is trained by fine-tuning ResNet18 (He et al., 2016) using the social media dataset discussed in (Alam et al., 2020). Let $f : \mathbb{R}^d \rightarrow \mathbb{R}^m$ be a pre-trained deep learning model that maps an input data point $x_i \in \mathbb{R}^d$ to a feature vector $f(x_i) \in \mathbb{R}^m$. For each data point $x_i \in \mathcal{D}$, the feature vector is extracted as: $\mathbf{z}_i = f(x_i), \quad \text{for } i = 1, 2, \ldots, N$ where $\mathbf{z}_i \in \mathbb{R}^m$ is the feature vector of the data point x_i. To compute the nearest neighbors between a data point x_i and the entire dataset $\mathcal{D}$, we use the euclidean distance. We then use a threshold of 3.6 to define the near-duplicate images as those with a euclidean distance less than or equal to this threshold value.

OCR Text: We used EasyOCR[4] to extract text from memes. Memes with no extracted OCR text were filtered out.

Classifier-Based Filtering: We employed an in-house meme vs. non-meme classifier to filter out images that were not classified as memes. The classifier was developed using a dataset of 3,935 images, consisting of 2,000 memes and 1,935 non-memes. Following the approach of (Hasnat et al., 2019), we developed a lightweight meme classifier to perform binary classification based on the extracted image features. The classifier achieved the best performance of 94.79% test set accuracy in classifying memes using a 256-dimensional normalized histogram extracted from gray-scale images as features, with a Multilayer Perceptron (MLP) as the classifier.

3.3 Annotation

Data Sampling: Due to budget constraints for manual annotation, we randomly sampled ~6K images.

Manual Annotation: For the manual annotation, we first prepared an annotation guideline to assist the annotators. To facilitate the annotation tasks, we developed an annotation platform as presented in Appendix D. The details of the annotation guidelines are reported in Appendix C. Note that we developed the annotation guidelines in English, (see Section C), which were then translated into the Arabic language. Translating the guideline in native language was indeed important and also inspired by prior work (Alam et al., 2021; Hasanain et al., 2024a). The idea is not only make the annotation task more convenient but also capture different linguistic aspects. The guidelines included several examples of memes. It was reviewed by several NLP experts who are also native Arabic speakers. The details of the Arabic annotation guideline can be found in https://shorturl.at/3z4CS.

In Figure 1, we provide examples of memes representing different categories. Figure 1(a) depicts a couple in what appears to be a therapy session. The therapist asks, "Do you feel your wife is controlling you?" The wife responds, "No, I don't feel so." It is evident that the question was directed towards the husband, yet the wife answers instead of him. The irony lies in her controlling the conversation when her control is the subject of discussion. This meme

[4] https://github.com/JaidedAI/EasyOCR

attempts to humorously portray the stereotypical notion that wives are controlling in marriages. Figure 1(b) employs a play on words to create humor but does not contain any propagandistic techniques. Figure 1(c) features a meme that uses an image of a scene with dialogue and added text to create humor. However, it was categorized as "other" because the dialogues were in English, rather than "not propagandistic" or "propagandistic." Figure 1(d) shows a picture of book covers, which might have been part of an advertisement.

The annotation tasks consist of two phases:

- Phase 1 (meme categorization): labeling memes as *(i)* not-meme, *(ii)* other, *(iii)* not propaganda, or *(iv)* propaganda. Each meme was annotated by three annotators and final label is decided based on majority agreement.
- Phase 2 (text editing): editing the text to fix OCR errors.

Annotation Team: The team in phase 1 consisted of three members, and in phase 2, it consisted of one member. All annotators are native Arabic speakers holding at least a bachelor's degree. Our in-house expert annotator provided them with several iterations of training, supervised and monitored their work, and handled quality control throughout the entire annotation process. This quality assurance included periodic checks of random annotation samples and providing feedback. Since the institute requires the signing of a Non-Disclosure Agreement (NDA), each annotator signed an NDA after being made aware of the institute's terms and conditions. They were compensated at the same rate as charged by external companies.

Annotation platform: We utilized our in-house annotation platform for the annotation task. Separate annotation interfaces were designed for each phase.

Annotation Agreement For the Phase 1 annotation, we computed annotation agreement using various evaluation measures, including Fleiss' kappa, Krippendorff's alpha, average observed agreement, and majority agreement. The resulting scores were 0.529, 0.528, 0.755, and 0.873, respectively. Based on the value of Krippendorff's alpha, we can conclude that our annotation agreement score indicates moderate agreement.[5] In the final label selection, we excluded the ∼200 memes on which the annotators disagreed. In the *second phase*, we mainly edited text to fix the OCR errors, which has been done by a single annotator. To ensure the quality of the *editing phase*, random samples were checked by an expert annotator and periodically provided feedback. Note that the post-editing has been done for only propagandistic and non-propagandistic memes. It is to reduce the cost of the annotation, and to further annotate them with span-level propaganda techniques.

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 2: Data split statistics.

3.4 Statistics

Table 2 shows the number of memes for each category. For the rest of the experiments, the data was split into train, dev, and test as shown in the table. The dataset comprises a total of 5,725 annotated samples, with "Not propaganda" covers over half of the dataset (∼66%), followed by "Propaganda." The "Not-meme" and "Other" classes are significantly smaller in comparison. The distribution indicates a significant class imbalance, particularly between "Not propaganda" and the other classes, which could affect model training and performance.

In Table 3, we report the distribution of the dataset across different sources. The annotated number of memes reflects the memes we collected from various sources, as detailed in Table 1. We have the highest number of memes collected and annotated from Instagram. A very small number from Twitter is due to different image filtering steps. As shown in Table 3 the prevalence of propagandistic memes is relatively higher on Facebook than that of non-propagandistic memes.

4 Experiments

4.1 Training and Evaluation Setup

For all experiments, except for those involving LLMs as detailed below, we trained the models using the training set, fine-tuned the parameters

[5] Note that Kappa values of 0.21–0.40, 0.41–0.60, 0.61–0.80, and 0.81–1.0 correspond to fair, moderate, substantial, and perfect agreement, respectively (Landis and Koch, 1977).

Source	Not prop.	Prop.	Not-meme	Other	Total
Facebook	464	332	58	144	998
Instagram	2,052	637	46	60	2,795
Pinterest	1,245	414	147	78	1,884
Twitter	3	5	38	2	48
Total	**3,764**	**1,388**	**289**	**284**	**5,725**

Table 3: Number of annotated memes across different sources. Prop. - Propaganda.

with the development set, and assessed their performance on the test set. We use the model with the best weighted-F1 on the development set to evaluate its performance on the test set. For the LLMs, we accessed them through APIs.

Evaluation Measures For the performance measure for all different experimental settings, we compute accuracy, and weighted precision, recall and F_1 score. In addition, we also computed macro-F1.

4.2 Models

We conducted our experiments using classical models (e.g., SVM) as well as both small (e.g., ConvNeXt-T) and large language models. It is important to note that our definitions of 'small' and 'large' models are based on the criteria discussed in (Zhao et al., 2023).[6]

4.2.1 Baseline:

We adopted widely-used standard baseline methods, including the majority and random baselines.

4.2.2 Small Language Models (SLMs)

We implemented classical models across all modalities, consisting of *(i)* feature extraction followed by model training, and *(ii)* fine-tuning pre-trained models (PLMs). For fine-tuning PLMs, we used a task-specific classification head over the training subset.

Text-Based Models: For the text-based unimodal model, we transformed text into n-gram (n=1) format using a tf-idf representation, considering the top 5,000 tokens, and trained an SVM model with a parameter value of $C = 1$. Additionally, we fine-tuned several pre-trained transformer models (PLMs). These included the monolingual transformer model AraBERT (Antoun et al., 2020), Qarib (Abdelali et al., 2021) and multilingual transformers such as multilingual BERT (mBERT) (Devlin et al., 2019), and XLM-RoBERTa (XLM-r) (Conneau et al., 2019). We used the Transformer toolkit (Wolf et al., 2019) for the experiment. Following the guidelines outlined in (Devlin et al., 2019), we fine-tuned each model using the default settings over three epochs. Due to instability, we performed ten reruns for each experiment using different random seeds, and we picked the model that performed best on the development set. We provided the details of the parameters settings in Appendix B.

Image-Based Models: For the image-based unimodal model with feature-extraction approach, we extracted features using ConvNeXt-T (Liu et al., 2022),[7] and trained an SVM model. For fine-tuning image-based PLMs, we used ResNet18, ResNet50 (He et al., 2016), VGG16 (Simonyan and Zisserman, 2014), MobileNet (Howard et al., 2017), and EfficientNet (Tan and Le, 2019). We chose these diverse architectures to understand their relative performance. The models were trained using the Adam optimizer (Kingma and Ba, 2015) with an initial learning rate of 10^{-3}, which was decreased by a factor of 10 when accuracy on the development set stopped improving for 10 epochs. The training lasted for 150 epochs.

Multimodal Models: We developed a multimodal model by concatenating text features (extracted using AraBERT) and image features (extracted using ConvNeXt-T), which were then fed into an SVM.

4.2.3 LLMs for Text

For the LLMs, we investigate their performance with zero-shot learning settings without any specific training. It involves prompting and post-processing of output to extract the expected content. Therefore, for each task, we experimented with a number of prompts. We used GPT-4 (OpenAI, 2023). We set the temperatures to zero for all these models to ensure deterministic predictions. We used LLMeBench framework (Dalvi et al., 2024) for the experiments, which provides seamless access to the API end-points and followed prompting approach reported in (Abdelali et al., 2024).

[6]The term 'LLMs' specifically refers to models that encompass tens or hundreds of billions of parameters.

[7]The configuration of ConvNeXt-T includes $C = (96, 192, 384, 768)$ and $B = (3, 3, 9, 3)$, where C and B represent the number of channels and blocks, respectively.

4.2.4 Multimodal LLMs

For the multimodal models (Xu et al., 2023), we experimented with several well-known and top-performing commercial models. These included OpenAI's GPT models (GPT-4 Turbo and GPT-4o) (OpenAI, 2023), as well as Google's Gemini Pro models (versions 1.0 and 1.5) (Team et al., 2023).

Using these models, we tested *(i)* the meme/image only, *(ii)* text only (text extracted using OCR from the image), and *(iii)* multimodal (meme and OCR text) in a zero-shot learning setting. This means we did not provide any training examples within the prompts to the models.

We designed a prompt based on trial and error using the visual interfaces of OpenAI's GPT-4 user interface. The prompt instructs the models to perform a deeper analysis of the image and any text that they can read within the image before answering whether the meme can be classified as spreading propaganda. Additionally, it requests the models to provide the output in a valid JSON format. For the experiments, we used the default parameters for each multimodal model.

4.3 Prompting Strategy

LLMs produce varied responses depending on the prompt design, which is a complex and iterative process that presents challenges due to the unknown representation of information within different LLMs. The instructions expressed in our prompts include English language with the input text content in Arabic.

As mentioned earlier we employed zero-shot prompting, providing natural language instructions that describe the task and specify the expected output. This approach enables the LLMs to construct a context that refines the inference space, yielding a more accurate output. In Listing 1, we provide an example of a zero-shot prompt, emphasizing the instructions and placeholders for both input and label. Along with the instruction we provide the labels to guide the LLMs and provide information on how the LLMs should present their output, aiming to eliminate the need for post-processing.

```
Instructions:
prompt = (
"You are an expert social media image
    analyzer specializing in identifying
     propaganda in Arabic contexts. "
"I will provide you with Arabic memes
    and the text extracted from these
    images. Your task is to briefly
    analyze them. "
"To accurately perform this task, you
    will: (a) Explicitly focus on the
    image content to understand the
    context and provide a meaningful
    description and "
"(b) pay close attention to the
    extracted text to enrich your
    description and support your
    analysis. "
"Finally, provide response in valid JSON
     format with two fields with a
    format: {\"description\": \"text\",
    \"classification\": \"propaganda\"}.
     Output only json. "
"The \"description\" should be very
    short in maximum 100 words and \"
    classification\" label should be \"
    propaganda\" or \"not-propaganda\"
    or \"not-meme\" or \"other\". "
"Note, other is a category, which is
    used to label the image that does
    not fall in any of the previous
    category."
)
```

Listing 1: Zero-shot prompt example for GPT-4.

5 Results and Discussion

In Table 4, we report the detailed classification results for different modalities and models. All models outperform the majority and random baselines. Among the text-based models, the fine-tuned Qarib model outperforms all other models, achieving the best results (**0.690** weighted F1) across all modalities and models. AraBERT is the second-best fine-tuned model, with a weighted F1-score of 0.666 among the text-based models. The performance of multilingual transformer models is relatively worse than that of monolingual models.

For the image-based models, the fine-tuned ResNet50 shows the best result (**0.673** weighted F1) among all other fine-tuned models and GPT-4o model. The performance of MobileNet (v2) and **CNeXt + SVM** rank as the second and third best among the fine-tuned models. The results of VGG16 and EfficientNet (b7) are almost similar.

For the multimodal models, the model trained with *ConvNeXt + AraBERT + SVM* shows the highest performance (0.659 weighted F1) among the

Model	Acc	W-P	W-R	W-F1	M-F1
Baseline					
Majority	0.658	0.433	0.659	0.522	0.198
Random	0.479	0.518	0.479	0.479	0.239
Unimodal - Text					
Ngram	0.669	0.624	0.669	0.582	0.280
AraBERT	0.688	0.670	0.688	0.666	0.511
Qarib	0.697	0.688	0.697	**0.690**	**0.551**
mBERT	0.707	0.688	0.707	0.675	0.487
XLM-r	0.699	0.676	0.699	0.678	0.489
GPT-4v	0.664	0.620	0.664	0.624	0.384
GPT-4o	0.573	0.611	0.573	0.579	0.350
Unimodal - Image					
CNeXt + SVM	0.655	0.608	0.655	0.614	0.405
MobileNet (v2)	0.660	0.618	0.660	0.620	0.426
ResNet18	0.656	0.597	0.656	0.593	0.358
ResNet50	0.660	0.638	0.660	0.637	0.434
Vgg16	0.656	0.597	0.656	0.593	0.358
Eff (b7)	0.660	0.597	0.660	0.595	0.352
GPT-4v	0.565	0.551	0.565	0.545	0.223
GPT-4o	0.693	0.627	0.693	0.634	0.305
Multimodal					
CNeXt + ArB + SVM	0.683	0.655	0.683	0.659	0.513
Gemini	0.519	0.551	0.519	0.521	0.276
GPT-4v	0.681	0.461	0.330	0.619	0.340
GPT-4o	0.653	0.443	0.354	0.639	0.363

Table 4: Classification with different modalities. CNeXt: ConvNeXt, Eff (b7): Efficientnet (b7), Gemini: Gemini-1.5-flash-preview-0514l, GPT-4v: GPT-4-vision (gpt-4-vision-preview) W-*: weighted average; M-: Macro average. XLM-r: XLM-RoBERTa base.

multimodal LLMs. The performance of Gemini is significantly worse than that of the GPT-4 variants. GPT-4o demonstrates higher performance compared to GPT-4 Vision.

In our experiments all multimodal model are tested using zero-shot setting, therefore, such lower performance compared to the fine-tuned models are expected.

6 Additional Experiments

We further conducted experiments using the dataset released as part of the ArAIEval shared task 2 (Hasanain et al., 2024b), focusing on two labels: propaganda and not-propaganda. The dataset statistics are provided in Table **??**. The goal was to investigate model performance in a binary classification scenario and we benchmarked this dataset using multimodal models.

Table 6 presents the competitive results of four multimodal models with image-only input: GPT-4o, GPT-4 Turbo, and Gemini Pro 1.0. Among these models, GPT-4o significantly outperforms the others and demonstrates the highest performance across all evaluated metrics, achieving an accuracy of 85.17%, a precision of 84.80, a recall of 85.17, and a weighted F1-score of 84.87. In comparison, GPT-4 Turbo lags behind GPT-4o in all metrics, with an accuracy of 76.44%, indicating a significant performance drop compared to GPT-4o. Gemini Pro 1.0 shows lower performance than the GPT-4 models, with an accuracy of 72.47%.

Class labels	Train	Dev	Test	Total
Not propaganda	1,540	224	436	2,200
Propaganda	603	88	171	862
Total	2,143	312	607	3,062

Table 5: Distribution of dataset for ArAIEval shared task 2.

Model	Acc.	W-P	W-R	W-F1	M-F1
Gemini	0.725	0.685	0.725	0.663	0.345
GPT-4v	0.764	0.748	0.764	0.735	0.645
GPT-4o	0.852	0.848	0.852	**0.849**	0.810

Table 6: Results on ArAIEval dataset. Gemini: version Pro 1.0.

7 Conclusions and Future Work

In this study, we introduce a manually annotated dataset for detecting propaganda in Arabic memes. We have annotated $\sim$ 6K memes with four different categories, making it the first such resource for Arabic content. To facilitate future annotation efforts for this type of content, we developed annotation guidelines in both English and Arabic and are releasing them to the community. Our work provides an in-depth analysis of the dataset and includes extensive experiments focusing on different modalities and models, including pre-trained language models (PLMs), large language models (LLMs), and multimodal LLMs. Our results indicate that fine-tuned models significantly outperform LLMs.

In future work, we plan to extend the dataset with further annotations that include hateful, offensive, and propagandistic techniques.

8 Limitations

The dataset we have collected originates from various public groups on Facebook, Instagram, Pinterest, and Twitter. The annotated dataset is highly imbalanced, which may affect model performance. Therefore, it is important to develop models with this aspect in mind.

Ethics and Broader Impact

Our dataset solely comprises memes, and we have not collected any user information; therefore, the privacy risk is nonexistent. It is important to note that annotations are subjective, which inevitably introduces biases into our dataset. However, our clear annotation schema and instructions aim to minimize these biases. We urge researchers and users of this dataset to remain critical of its potential limitations when developing models or conducting further research. Models developed using this dataset could be invaluable to fact-checkers, journalists, and social media platforms.

Acknowledgments

The contributions of F. Alam, M. Hasanain, and F. Ahmed were funded by the NPRP grant 14C-0916-210015, which is provided by the Qatar National Research Fund (a member of Qatar Foundation).

References

Shamsiah Abd Kadir, Anitawati Lokman, and T. Tsuchiya. 2016. Emotion and techniques of propaganda in YouTube videos. *Indian Journal of Science and Technology*, Vol (9).

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

Ahmed Abdelali, Hamdy Mubarak, Shammur Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Samir Abdaljalil, Yassine El Kheir, Daniel Izham, Fahim Dalvi, Majd Hawasly, Nizi Nazar, Youssef Elshahawy, Ahmed Ali, Nadir Durrani, Natasa Milic-Frayling, and Firoj Alam. 2024. LAraBench: Benchmarking Arabic AI with large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 487–520, St. Julian's, Malta. Association for Computational Linguistics.

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

Firoj Alam, Fahim Dalvi, Shaden Shaar, Nadir Durrani, Hamdy Mubarak, Alex Nikolov, Giovanni Da San Martino, Ahmed Abdelali, Hassan Sajjad, Kareem Darwish, et al. 2021. Fighting the covid-19 infodemic in social media: a holistic perspective and a call to arms. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 913–922.

Firoj Alam, Hamdy Mubarak, Wajdi Zaghouani, Giovanni Da San Martino, and Preslav Nakov. 2022b. Overview of the WANLP 2022 shared task on propaganda detection in Arabic. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop*, Abu Dhabi, UAE.

Firoj Alam, Ferda Ofli, Muhammad Imran, Tanvirul Alam, and Umair Qazi. 2020. Deep learning benchmarks and datasets for social media image classification for disaster response. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 151–158. IEEE.

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.

Sian Brooke. 2019. "condescending, rude, assholes": Framing gender and hostility on stack overflow. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 172–180, Florence, Italy. Association for Computational Linguistics.

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.

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

Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv:1911.02116*.

Padraig Cunningham and Sarah Jane Delany. 2007. k-nearest neighbour classifiers. *Multiple Classifier Systems*, 34(8):1–17.

Giovanni Da San Martino, Alberto Barrón-Cedeño, and Preslav Nakov. 2019. Findings of the NLP4IF-2019 shared task on fine-grained propaganda detection. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, NLP4IF '19, pages 162–170, Hong Kong, China.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019a. Fine-grained analysis of propaganda in news articles. In *EMNLP*.

Giovanni Da San Martino, Seunghak Yu, Alberto Barron-Cedeno, Rostislav Petrov, and Preslav Nakov. 2019b. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, EMNLP '19, pages 5636–5646, Hong Kong, China.

Fahim Dalvi, Maram Hasanain, Sabri Boughorbel, Basel Mousi, Samir Abdaljalil, Nizi Nazar, Ahmed Abdelali, Shammur Absar Chowdhury, Hamdy Mubarak, Ahmed Ali, Majd Hawasly, Nadir Durrani, and Firoj Alam. 2024. LLMeBench: A flexible framework for accelerating LLMs benchmarking. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 214–222, St. Julians, Malta. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11 of *AAAI '17*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of the 2019 Conference of the NAACL*, NAACL-HLT '19, Minneapolis, MN, 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 *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, ACL-IJCNLP '21, pages 6603–6617.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021b. Task 6 at SemEval-2021: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th International Workshop on Semantic Evaluation*, SemEval '21, Bangkok, Thailand.

Institute for Propaganda Analysis. 1938. In Institute for Propaganda Analysis, editor, *Propaganda Analysis. Volume I of the Publications of the Institute for Propaganda Analysis*, chapter 2. New York, NY.

Shreyansh Gandhi, Samrat Kokkula, Abon Chaudhuri, Alessandro Magnani, Theban Stanley, Behzad Ahmadi, Venkatesh Kandaswamy, Omer Ovenc, and Shie Mannor. 2020. Scalable detection of offensive and non-compliant content / logo in product images. *WACV*, pages 2236–2245.

Maria Glenski, E. Ayton, J. Mendoza, and Svitlana Volkova. 2019. Multilingual multimodal digital deception detection and disinformation spread across social platforms. *ArXiv*, abs/1909.05838.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *WACV*, pages 1470–1478.

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, 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 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. 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.

Abul Hasnat, Nadiya Shvai, Assan Sanogo, Marouan Khata, Arcadi Llanza, Antoine Meicler, and Amir Nakib. 2019. Application guided image quality estimation based on classification. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 549–553. IEEE.

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, Shivam Sharma, Rui Cao, Palash Nandi, Preslav Nakov, Tanmoy Chakraborty, and Roy Ka-Wei Lee. 2024. Recent advances in hate speech moderation: Multimodality and the role of large models. *arXiv preprint arXiv:2401.16727*.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv:1704.04861*.

Srecko Joksimovic, Ryan S. Baker, Jaclyn Ocumpaugh, Juan Miguel L. Andres, Ivan Tot, Elle Yuan Wang, and Shane Dawson. 2019. Automated identification of verbally abusive behaviors in online discussions. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 36–45, Florence, Italy. Association for Computational Linguistics.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. In *Proceedings of the NeurIPS 2019 Workshop on Visually Grounded Interaction and Language*, ViGIL@NeurIPS '19.

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. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, NeurIPS '20.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the Conference on Neural Information Processing Systems*, NeurIPS '19, Vancouver, Canada.

R OpenAI. 2023. Gpt-4 technical report. *arXiv*, pages 2303–08774.

Alan Partington and Charlotte Taylor. 2017. *The language of persuasion in politics: An introduction*. Routledge.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023a. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.

Jakub Piskorski, Nicolas Stefanovitch, Nikolaos Nikolaidis, Giovanni Da San Martino, and Preslav Nakov. 2023b. 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.

Nirmalendu Prakash, Han Wang, Nguyen Khoi Hoang, Ming Shan Hee, and Roy Ka-Wei Lee. 2023. PromptMTopic: Unsupervised multimodal topic modeling of memes using large language models. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 621–631, New York, NY, USA. Association for Computing Machinery.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Hyunjin Seo. 2014. Visual propaganda in the age of social media: An empirical analysis of Twitter images during the 2012 Israeli–Hamas conflict. *Visual Communication Quarterly*, 21(3).

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.

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

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *TRAC*, pages 32–41.

Mingxing Tan and Quoc V Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv:1905.11946*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Veronique Hoste. 2015. Detection and fine-grained classification of cyberbullying events. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 672–680, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Svitlana Volkova, Ellyn Ayton, Dustin L. Arendt, Zhuanyi Huang, and Brian Hutchinson. 2019. Explaining multimodal deceptive news prediction models. In *Proceedings of the International Conference on Web and Social Media*, ICWSM '19, Munich, Germany.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art natural language processing. *arXiv:abs/1910.03771*.

Ching Seh Wu and Unnathi Bhandary. 2020. Detection of hate speech in videos using machine learning. In *CSCI*, pages 585–590.

Peng Xu, Xiatian Zhu, and David A Clifton. 2023. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Prosper Nunayon Zannu, Felix Olajide Talabi, Bernice Oluwalanu Sanusi, Oluwakemi A. Adesina, Adebola Adewunmi Aderibigbe, Omowale T. Adelabu, Oloyede Oyinloye, and Samson Adedapo Bello. 2024. Influence of media literacy on the dissemination of fake news among instagram and twitter users. *International Research Journal of Multidisciplinary Scope (IRJMS)*, 5(2):246–255. Original Article.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

A Additional Dataset Details

أخبار غير صحيحة, إشاعة مكذوبة, إمسح التغريدة بلاش أخبار كاذبة, اخبار كاذبه, اخبارك مضروبة, الخبر غير صحيح, الخبر كاذب, الصورة غير صحيحة, المعلومة غير صحيحة, المعلومة كاذبه, خبر كاذب, خبر غير صحيح, غير صحيح الخبر هذا, البيان مزور, هذه أخبار مكذوبة, إشاعة مكذوبة, الخبر عار تماما, الخبر غير دقيق, خبر مفبرك, الخبر مفبرك, ليس صحيح الخبر, عنف, تنمر, صارِم, تهديد, مفرط, قوي, عنيف, متوحشون, قاسِي

Figure 3: Keywords used to collect tweets.

B Details of the experiments

For the experiments with transformer models, we adhered to the following hyper-parameters during the fine-tuning process. Additionally, we have released all our scripts for the reproducibility.

- Batch size: 8;
- Learning rate (Adam): 2e-5;
- Number of epochs: 10;
- Max seq length: 256.

Models and Parameters:

- **AraBERT**: L=12, H=768, A=12; the total number of parameters is 371M.
- **XLM-RoBERTa** (xlm-roberta-base): L=24, H=1027, A=16; the total number of parameters is 355M.

C Annotation Task

We designed the annotation instructions through careful analysis and discussion, followed by iterative refinements based on observations and input from the annotators based on the pilot annotation. Our annotation schema is structured into two phases.

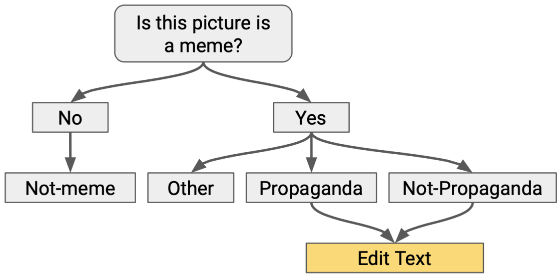

Figure 4: A visual representation of the annotation process. Block with yellow color represents phase 2.

C.1 Phases of Annotations

To ensure the quality of the annotation and facilitate the work of annotators, we conducted the annotation in two phases: *(i)* meme categorization and *(ii)* text editing. The *first phase* (see Section C.2) focuses primarily on categorization. In the second phase (see Section C.3), our goal is to edit the text only for memes labeled as propagandistic or not propagandistic. The motivation for editing the text for these categories is to further utilize them for other annotation tasks. For example, propagandistic memes can be further annotated with specific propagandistic techniques. In Figure , we illustrate the thought process of the meme annotation phases.

C.2 Meme Categorization

C.2.1 Definition of a Meme:

Memes typically consist of a background image, which could be a photograph, illustration, or screenshot, and a layer of text that adds context, humor, or commentary to the image. The text is usually placed at the top and/or bottom of the image but not always. The combination of the image and the text creates a specific message, joke, or commentary that is meant to be easily understood, relatable, and shareable. Some characteristics of memes as observed during analysis and discussion:

1. Text overlaid on image.
2. The text has humor in it.
3. The image *must* meet points 1 and 2.
4. Some contents of the image have been edited.
5. Text might be added to different locations of the image.
6. Mostly uses images of entities with facial expressions (human, animals, fictional characters), which are then used to construct meaning alongside the added text.
7. Uses an entity performing a certain action that might be used to construct meaning alongside the added text.
8. Uses an entity that represents an idea or culture, to construct meaning alongside the added text.
9. Mostly uses screenshots from movie scenes and dialogues with added comments, to create memes.
10. Most of the pictures used to make the meme can be re-edited and a new funny comment can be added to it.

Note: In points 6, 7, and 8, the removal of the entity from the images will affect the meaning. In other words, if the entity is removed, then the meaning will not be complete. This is what we mean by constructing meaning.

C.2.2 Defining Propaganda:

Propaganda is any communication that deliberately misrepresents symbols and/or entities, appealing to emotions and prejudices while bypassing rational thought, to influence its audience toward a specific goal. Memes are created to be humorous; therefore, it is natural that they lack rational discussion. Instead, they use content to appeal to emotions and prejudices. For our task, we defined the following four categories and annotated the memes accordingly.

(1) Not-Meme: For images that do not follow the definition of a meme, examples of images labeled as "not-meme" are shown in Figure 5.

(2) Other: For images that can be defined as memes but fall under any of the criteria listed below. Examples of images labeled as "not-meme" are shown in Figure 6. The criteria for "Other":

1. Memes that rely on nudity and offensive content, unless the target of the offense is a famous, political, or religious entity.
2. Memes that rely on numbers or figures to construct meaning.
3. Memes that show explicit nudity.
4. Memes that explicitly use offensive words.
5. Memes that are in a different language (not Arabic).

Figure 5: Examples of images labeled as *not-meme*.

6. Memes that you could not understand due to the dialect it was written in, poor font size, or for any other reason.

Note: Memes might contain words that have an implicitly offensive meaning, or the use of offensive words may be aimed at social, religious, or political groups. In these cases, the meme does not fall under this criterion.

(3) Not Propaganda: For memes that follow the definition of memes but do not contain any propaganda techniques, examples of images labeled as "not propagandistic" are shown in Figure 7.

(4) Propaganda: For memes that follow the definition of memes and contain propaganda techniques, examples of images labeled as "propagandistic" are shown in Figure 8.

C.3 Text Editing

The task is to edit the text to match the text shown in the image. The interface will show you the picture, alongside the text that is viewed in it. The text was extracted automatically, so it might contain errors. It might not reflect all you see in the picture. Some important guidelines to follow for editing the text are listed below:

1. Each part that is a standalone sentence and makes complete meaning should be written as one line.

2. Punctuation marks are considered a part of the text. They need to be edited/added.

3. If the text is in columns, put first all the text of the first column, then all the text of the next column. This task will specifically address memes in Arabic, so the first column should be considered from the right. However, this is not a rule, and memes might change this orientation, so it is up to the annotator to decide the order based on their understanding.

4. Rearrange the text so that there is one sentence per line, if possible.

5. If there are separate blocks of text in different locations of the image, start a new line from each block.

6. Leave a blank between two blocks of text if they were shown in two different locations on the picture.

7. Items that should be excluded from the text:
   - Usernames and social media account names (if visible in the image).
   - Websites, logos, and any text that is not a part of the meme, so that removing that part does not affect the meaning of the meme.
   - Any text that is hidden and is hard to read.

8. In special cases, a logo can be used in the meme to create meaning. In this case, add the text of the logo to the edited text, if needed.

Example 1: Figure 9 shows an example of a meme, for editing the text that can be viewed it, the following points are important:

- Each dialog box is one sentence

- Start a new line for each box (each box is a different block of text)

- Remove any elements that are not part of the meaning: account name and location

Figure 6: Examples of images labeled as *other*.

Figure 7: Examples of images labeled as *not propaganda*.

- Add or modify punctuation to suit what is presented in the text
- Text after modification (text translated to EN and read from the first speech bubble from right):

```
Get him ... Get him... corner him...
get him so we can give him his rights
come... aren't you coming??
come...take your rights you son of
a bastard
Wallah we gonna get you till...
we give you all your rights
you chick ....
```

Example 2: Figure 10 shows another example, for which the following points are important.

- Text written in red is difficult to understand and read, so it should not be included in the text.
- The text written on the hat and the text in black are each a different block of text. Start a line for each of them and leave a space for each new line.
- This example is for illustrative purposes only, and "memes" in English will not be shown in this task.
- Text after modification:

```
Bernie
Riding with Biden **2020**
Haha hey its the Obama guy
```

D Annotation Platform

D.1 Meme Categorization Task

In Figure 11, we provide a screenshot of the annotation platform for the meme categorization task. As shown in the figure, the platform displays the meme itself on the right, the extracted text on the left, a link to the annotation guidelines, and labels with buttons at the bottom for selecting a category for the meme. The task of the annotator was to label the meme as one of the below categories, according to the definitions detailed in the guideline (see Section C). To facilitate the work of annotators in the annotation process, we used the keywords 'meme' along with the labels 'other', 'propaganda', and 'not-propaganda'.

- Not Meme

Figure 8: Examples of images labeled as *propaganda*.

Figure 9: An example of a meme for editing text.

Figure 10: An example of a meme for editing text.

- Meme, Other
- Meme, Not Propaganda
- Meme, Propaganda

Given that the memes we collected were from different social media platforms, they may contain offensive content. Therefore, we added a note that *some pictures may contain offensive content, and that we apologize for any inconvenience that such content may cause. We appreciate your contribution to this project which will minimize the spread of such harmful content on the internet*.

To further guide the annotation process, we asked the annotators to follow the following steps.

1. Begin by determining whether the image presented is a "meme". If the image is not a meme, select "Not Meme", then click "Submit". The next image will then be loaded.

2. If the image is a "meme", assess whether it falls under the category of "Other". If so, select "Other", then click "Submit". The next image will then be loaded.

3. If the image does not fall under the category of "Other", choose one of the remaining two labels based on your interpretation of the meme's content. After selecting the appropriate label, edit the text as needed.

D.2 Text Editing Task

In this phase, the task was to edit the text based on the guidelines discussed in Section C.3. In Figure 12, we provide a screenshot demonstrating the text extracted from OCR, an editable text box, and the original meme. The task was to edit the text to match it with the original meme.

Figure 11: A screenshot of the annotation platform for the *meme categorization* task.

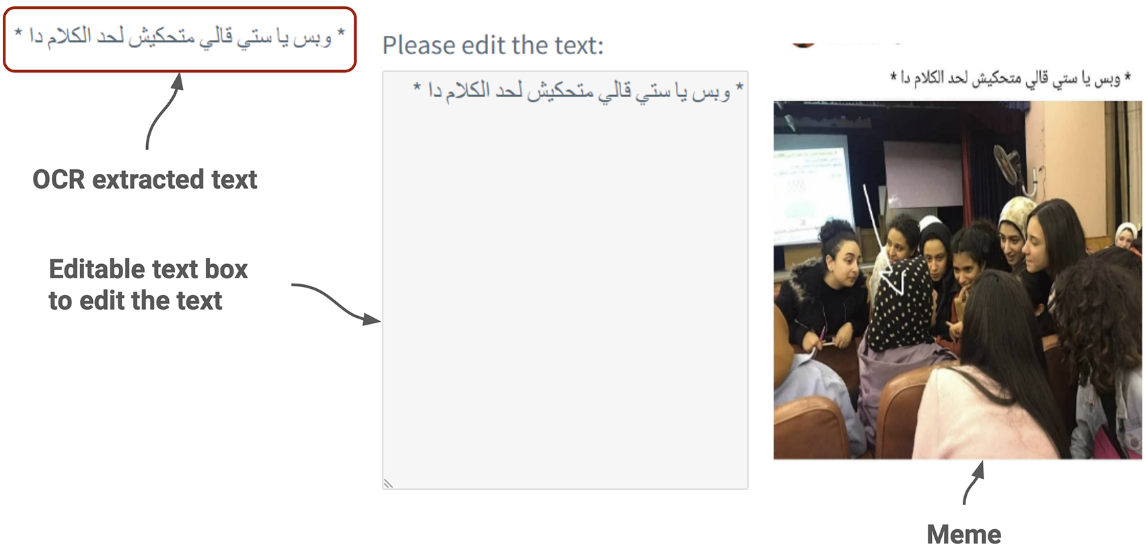

Figure 12: An screenshot of the annotation platform for the *text editing*.