# Propaganda to Hate: A Multimodal Analysis of Arabic Memes with Multi-Agent LLMs

Firoj Alam[1][0000−0001−7172−1997], Md. Rafiul Biswas[2][0000−0002−5145−1990], Uzair Shah[2][0000−0002−6729−5654], Wajdi Zaghouani[3][0000−0003−1521−5568], and Georgios Mikros[2][0000−0002−4093−5973]

[1]Qatar Computing Research Institute, Qatar, [2]Hamad bin Khalifa University, Qatar, [3]Northwestern University in Qatar, Qatar
fialam@hbku.edu.qa
[✉]Corresponding author

Abstract. In the past decade, social media platforms have been used for information dissemination and consumption. While a major portion of the content is posted to promote citizen journalism and public awareness, some content is posted to mislead users. Among different content types such as text, images, and videos, memes (text overlaid on images) are particularly prevalent and can serve as powerful vehicles for propaganda, hate, and humor. In the current literature, there have been efforts to individually detect such content in memes. However, the study of their *intersection* is very limited. In this study, we explore the *intersection* between propaganda and hate in memes using a multi-agent LLM-based approach. We extend the propagandistic meme dataset with coarse and fine-grained hate labels. Our finding suggests that there is an association between propaganda and hate in memes. We provide detailed experimental results that can serve as a baseline for future studies. We will make the experimental resources publicly available to the community.[1]

Keywords: Propaganda· Hateful Meme· Multimodality· LLMs·

1 Introduction

Social media has emerged as a primary channel for freely sharing content online. Its exponential growth has significantly transformed the landscape of information dissemination. However, misuse of these platforms has made them fertile grounds for the spread of inappropriate content, misinformation, and disinformation [2]. While interactions on social media facilitate public discussions on a range of topics, from local issues to politics, they also harbor and propagate hate speech and offensive content through various content types, text, images, and videos [26,16,31,2].

---

[1] https://github.com/firojalam/propaganda-and-hateful-memes.git

To address such problems across different modalities, there have been efforts to automatically detect them using both mono- and multimodal modeling approaches [9]. For propagandistic content detection, research efforts have specifically focused on defining techniques and tackling the issue across various types of content, including news articles, tweets, memes, and textual content in multiple languages [13,4,28]. Similarly, significant efforts have been made in hate-speech detection [29,10]. A notable initiative in meme research is the Hateful Memes Challenge [23], which has inspired many subsequent studies.

Our research lies at the intersection of multimodal content analysis, propaganda detection, and hate speech identification. While progress has been made in these fields for English and other high-resource languages, the research for Arabic remains underexplored. Building on the work of [5] and [18] in Arabic propaganda detection, our study analyzes Arabic memes, addressing a gap in the literature. The findings can assist social media platforms, policymakers, and civil society organizations in combating harmful online content.

The key aspects of our work are as follows: *(i)* we present a novel multi-agent LLM-based approach to analyze the association between propaganda and hateful memes in Arabic social media content; *(ii)* we demonstrate the application of multi-agent LLM systems for automated annotation of complex, multimodal data, offering a scalable solution for processing large volumes of data in low-resource settings; *(iii)* in addition to coarse-grained hateful categories, we also explore fine-grained categorization of hateful and non-hateful memes; *(iv)* we provide experimental results on both coarse and fine-grained categories; *(v)* finally, we will make the dataset of hateful memes available to the community.

2   Related Work

2.1   Propagandistic Content

Textual Content: The study of propagandistic content has attracted significant attention in recent years. Da et al. (2020) introduced a large-scale dataset for fine-grained propaganda detection in news articles, presenting a corpus of 350K sentences annotated with 18 propaganda techniques [8]. The annotation schema has been extended to include 23 techniques and a multilingual corpus has been proposed in [28]. Following the same annotation schema, datasets have been developed for Arabic and shared tasks has been organized [4,20,19].
Multimodal Content: Building on previous research in textual content, Dimitrov et al. (2021) introduced SemEval-2021 Task 6, which focuses on detecting persuasion techniques in texts and images within memes [14]. Subsequently, the focus has expanded to include the detection of multilingual and multimodal propagandistic memes [12]. Similarly, related work on Arabic involves the development of datasets and a shared task for propaganda detection [3,21]. Fang et al. (2022) used separate networks to embed text and images, fusing these multimodal embeddings. A split-and-share module with multi-level representations was employed to improve persuasive technique detection [15].

2.2   Hateful Memes

The study of hateful memes presents unique challenges due to their multimodal nature. Kiela et al. (2020) introduced the Hateful Memes Challenge, a large-scale dataset and benchmark for multimodal hate detection. This work highlighted the importance of integrating both textual and visual elements to identify hate speech in memes [23]. Addressing the challenge of low-resource languages, datasets have also been developed in various languages [22]. In [31], the authors provide a detailed survey of multimodal and harmful memes, highlighting the significance of the problem and proposing future research avenues.

2.3   Multi-Agent Systems in Content Analysis

The application of multi-agent systems to content analysis is an emerging field, which could be an effective approach in analyzing complex narratives across various media [17]. Chen et al. (2021) introduced a dynamic content moderation system using multi-agent reinforcement learning, which adapts based on user interactions and content patterns for improved detection of harmful content [7]. These studies emphasize the value of multi-agent systems in analyzing complex, often propagandistic or hateful content in memes. Building on this, our work focuses on Arabic memes, employing a multi-agent LLM approach. We explore the association between propaganda and hateful memes in low-resource settings. We employ LLMs as multiple agents to automate the data annotation process, demonstrating the utility of LLMs as data annotators in detecting hateful memes.

3   Dataset

3.1   Propagandistic Memes

For this study, we used ArAIEval-2024 dataset [21], which consists of approximately ∼3k memes, each annotated with labels as either propagandistic or not-propagandistic. These memes were collected from various social media platforms, including Facebook, Twitter, Instagram, and Pinterest. We have annotated each meme by three annotators, with the final label determined by majority vote. The text from the memes was extracted using an off-the-shelf OCR tool[2], followed by manual corrections for propagandistic memes. The distribution of propagandistic and not-propagandistic labels is 40% and 60%, respectively. Further details about this dataset are available in [21], and the comprehensive annotation guidelines are provided in [3]. For the experiments, the dataset is split into 70%, 10%, and 20% for training, development, and testing, respectively.
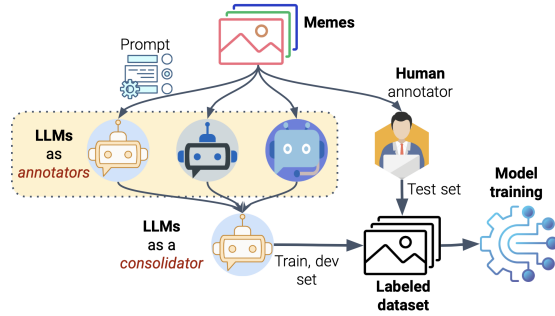
---

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

Fig. 1. Experimental pipeline with LLM agents as annotators and consolidator.

3.2   Hatefulness and Fine-grained Categories

For the hatefulness and their fine-grained categorization we used ArAIEval-2024 dataset, mentioned earlier. Our motivation to use ArAIEval-2024 dataset is that this is the only meme dataset currently available for Arabic, which has already been annotated for propagandistic content. Another motivation was to understand the association between propagandistic and hateful memes. In Figure 1, we provide full pipeline for the data preparation to classification experiments.

3.2.1   LLM Agents as Annotators To employ LLM agents as annotators, we selected three well-known and top-performing commercial models: OpenAI's GPT-4o [27], Google's Gemini Pro (version 1.5) [32], and Claude 3.5 (Sonnet).[3]. For the annotation process, we use the same manual procedure discussed in [18], which involves a two-phase approach. In the first phase, known as the *annotation phase*, three annotators independently annotate memes following the guidelines outlined in 3.2.2. In the second phase, known as the *consolidation phase*, we review and resolve any disagreements from the annotations received during the first phase. As illustrated in the figure, highlighted in dark red, we employ LLM agents as *annotators* in the first phase and as a *consolidator* in the second phase. For each phase, we use a specific prompt in a zero-shot setup for the LLM agent. Following the annotation guidelines discussed below, we ask an LLM agent to perform two tasks: *(Task 1)* label each meme as hateful or not-hateful, and *(Task 2)* based on the label from Task 1, provide a fine-grained categorization. For example, if a meme is categorized as hateful in Task 1, it should then provide a fine-grained label from one of the eight categories mentioned below. The prompt in the second phase is slightly different. Here, the task also involves considering the labels obtained from the first phase to make a final decision. For this phase, we have experimented with using GPT-4o as the consolidator.

We used this LLM-based multi-agent approach for the training and development (dev) sets. To validate the quality of the multi-agent approach, we quan-

---

[3] https://www.anthropic.com/news/claude-3-5-sonnet

tified the labels provided by each LLM agent by comparing them with human-annotated labels on the test dataset.

3.2.2   Manual Annotation To verify the LLM-based multi-agent approach, we manually annotated a test set from the ArAIEval-2024 dataset, as shown in Figure 1. For the annotation process, we developed a set of instructions, which are discussed below. The typical approach to annotation involves two to three annotators. However, for this study, we relied on a single annotator who had prior experience with similar annotation tasks.

3.2.3   Annotation Instructions The purpose of this annotation is to identify whether a meme is hateful or not-hateful. A hateful meme can attack different individuals, organizations or entities. Therefore, another task is identifying the attack types. A non-hateful meme can be humorous or sarcastic. Therefore, the idea to also identify the sub-categories within non-hateful memes. We adopted the annotation definition and instructions from prior work [23,25]. Below we provide the definitions:

Hateful: A direct or indirect attack on individuals based on characteristics such as ethnicity, race, nationality, immigration status, religion, caste, sex, gender identity, sexual orientation, disability, or disease. We define an attack as hate-speech that is violent, dehumanizing (such as comparing individuals to non-human entities like animals), involves statements of inferiority, or calls for exclusion or segregation. Mocking hate crimes is also classified as hate speech. However, attacks directed at groups that perpetuate hate (e.g., terrorist organizations) are not considered hate speech.[4]

    Fine-grained categories hatefulness:
- **Dehumanizing:** Explicitly or implicitly portraying or describing a group as subhuman.
- **Inferiority:** Asserting that a group is inferior, less worthy, or less important than society as a whole or compared to another group.
- **Inciting violence:** Explicitly or implicitly advocating for harm to be inflicted on a group, including physical violence.
- **Mocking:** Joking about, ridiculing, demeaning, or disparaging a group.
- **Contempt:** Expressing strong negative emotions or feelings toward a group.
- **Slurs:** Using biased or derogatory terms to refer to, describe, or characterize a group.
- **Exclusion:** Advocating for, planning, or justifying the exclusion or segregation of a group from society as a whole or from specific areas.
- **Other:** None of the above.

Not-Hateful: The content is humorous, neutral, or positive, without targeting or harming specific individuals or groups. It is light-hearted and intended for en-

---

[4] https://transparency.meta.com/en-gb/policies/community-standards/hate-speech/

tertainment without being offensive. Additionally, the content does not promote or incite violence, hatred, or discrimination.

Fine-grained not-hateful categories:
– **Humor:** The purpose of humor is to entertain, amuse, or bring joy to the audience. Often characterized by jokes, puns, or playful language. Humor can vary widely in style, including wit, slapstick, parody, and satire.
– **Sarcasm:** Typically involves saying the opposite of what one means. Sarcasm is a form of irony that always occurs with a deliberate mismatch between what is said and what is meant, intentionally to ridicule or mock a specific target.
– **Other:** None of the above

3.2.4   Annotated Dataset As discussed in Section 3.2.1, we annotated the test data into 'Hateful' and 'Not-Hateful' categories using GPT-4o, Sonnet (Claude 3.5), and Gemini (Vertex). We then provided the three annotated labels (obtained from GPT-4o, Sonnet, and Gemini) as prompts to GPT-4o and asked it to choose the best label that matches the data. The generated output label is termed *GPT-4o consolidation*. Table 1 shows the inter-annotator agreement (IAA) among the annotators. We computed the annotation agreement using pairwise Cohen's kappa score in different setups: *(i)* LLMs as annotators *vs.* an LLM as a consolidator, *(ii)* LLMs as annotators *vs.* human annotation, and *(iii)* pairwise between LLMs as annotators.

It shows that the IAA between *GPT-4o* and *GPT-4o consolidation* is high (0.786), representing substantial agreement. It is reasonable because the *GPT-4o consolidation* is derived from the GPT-4o, which means that the consolidated label inherently aligns closely with the labels initially provided by GPT-4o. Interestingly, we observe that the IAA between Sonnet and *GPT-4o consolidation* is significant and high (0.701), which denotes Sonnet's capability to understand hateful content and memes.

Table 1 shows that the IAA between Sonnet and the human annotator achieved a higher score (0.405) compared to other annotation labels. We also performed pairwise agreements among LLM annotators and found that the agreement between Sonnet and GPT-4o is higher (0.528). Finally, we measured the annotation agreement among all three annotators and obtained a score of 0.369.

The annotation agreements between the three different LLMs and the human annotator suggest that Sonnet has a fair capability of understanding 'Hateful' content compared to a human annotator. Therefore, we used Sonnet to annotate the training and development datasets.

Data Stat: Hateful Meme Table 2 presents the distribution of class labels for training, development, and testing datasets, categorized into "Hate/Not-hate" and further labeled into fine-grained categories. There is a significantly larger number of "Not-Hateful" (N=1931) category instances compared to the "Hateful" (N=212) category. In the fine-grained label, "Mocking" has a notable presence (N=133) in the 'Hateful' category. Similarly, in the fine-grained "Not-

Table 1. Annotation agreement for different setups.

Anno. 1	Anno. 2	Kappa	Anno. 1	Anno. 2	Kappa
Agreement: LLMs vs. LLM as a Consolidator			**Agreement: LLMs vs Human**		
GPT-4o	GPT-4o	0.786	GPT-4o	Human	0.233
Sonnet (Claude)	GPT-4o	0.701	Claude-3.5 (Sonnet)	Human	0.405
Gemini-1.5 (Vertex)	GPT-4o	0.236	Gemini-1.5 (Vertex)	Human	0.300
Agreement: LLMs (Pairwise)			GPO-4o Consolidation	Human	0.300
Gemini-1.5 (Vertex)	Claude-3.5 (Sonnet)	0.266			
GPT-4o	Gemini-1.5 (Vertex)	0.142			
Claude-3.5 (Sonnet)	GPT-4o	0.528			

Hateful" categories, "Humor" (N=1815) overwhelmingly dominates, followed by "Sarcasm". This distribution highlights the imbalance in the data.

Table 2. Distribution of annotated data: The training and development sets were labeled using Sonnet, while the test set was labeled by a human.

Hate/Not-hate				**Hate: Fine-grained categories**			
Label	**Train**	**Dev**	**Test**	**Label**	**Train**	**Dev**	**Test**
Hateful	212	32	154	Contempt	38	7	25
Not-Hateful	1,931	280	452	Dehumanizing	12	3	2
Total	2,143	312	606	Mocking	133	19	49
Non-Hate: Fine-grained categories				Inferiority	5	1	14
Label	Train	Dev	Test	Exclusion	6	7	3
Sarcasm	105	19	118	Inciting violence	13	2	12
Humor	1,815	260	334	Slurs	6	1	29
Total	1,920	279	452	Other	10	1	20
				Total	223	41	154

Propaganda and Hateful Meme: To understand the correlation between propaganda and hateful memes, we observe that out of 171 propagandistic memes in the test set, 56 memes are hateful (30%) and 70% are not hateful. This is possible because propagandistic memes may not always instigate hate or harm.

4   Experiments

Our experiments consist of three setups: *(i)* hate vs. not-hate, *(ii)* fine-grained categories for hateful memes, and *(iii)* fine-grained categories for non-hateful memes. These classification experiments involve unimodal (text and image) and multimodal classifications.

Text classification: We extracted text from propagandistic memes and applied various text classification techniques such as AraBERT, mBERT, CAMel-BERT, and Qarib-BERT ([6,11,1]). The original dataset is imbalanced, and so

we implemented a class weighting scheme during the fine-tuning process. More-over, we optimized the model by adjusting the dropout rate. This approach led to significant improvements over using the original dataset alone. We embed-ded the LoRa to fine-tune the model in an efficient way that does not require fine-tuning all the parameters of the model. However, embedding LoRA did not improve the performance. We then fine-tuned AraBERT model.

Image classification: We fine-tuned ResNet50 and ConvNeXt-tiny [24]. To ensure stable weight adjustments, we froze the feature extraction layers and fine-tuned the classification layer. We also adjusted the dropout rate during the model training.

Multimodal classification: To extract visual and text features, we applied ConvNext tiny and AraBERT models respectively. We combine the features using a fusion layer. We froze the visual models and trained the classification layer with textual data. We also used a dropout rate to improve the performance.

Experimental Setup: We performed all of our experiments and trained our models on an Nvidia-RTX 2080 GPU. We employed the Adam optimizer with an initial learning rate of 1e-5 and 1e-4 for text and image, respectively. We used a batch size of 32, a sequence length of 128. We set the dropout rate to 0.25 for text data and trained for 50 epochs. For image data, we set a dropout rate of 0.5 and trained for 30 epochs. For multimodal data, we trained the model for 100 epochs with a stochastic drop rate of 0.2. Note that our choice of the models and parameters was inspired by prior studies [3,30].

Table 3. Classification results on the test set for coarse and fine-grained hate labels across different modality setups.

Modality	Model	Acc	M-F1	Modality	Model	Acc	M-F1
Hate vs. Not-hate				**Balanced dataset: Hate vs. Not-hate**			
Text	AraBERT	0.819	0.705	Balanced	Fusion	0.817	0.709
Image	ConvNxT	0.779	0.669	**Fine-grained label**			
Text+Image	Fusion	0.764	0.709	Hateful	Fusion	0.224	0.166
				Not-Hateful	Fusion	0.622	0.537

5   Results and Discussion

In Table 3, we present the results for different classification setups. For hate vs. not-hate classification across different modalities, considering macro-F1, the text and multimodal models exhibit similar performance. For the multimodal model, we obtained model with a macro-F1 of 0.709 and an accuracy of 0.764. For the text modality, we obtained a macro-F1 of 0.705 and an accuracy of 0.818. For the image modality, we obtained a macro-F1 of 0.669 and an accuracy of 0.775.

To understand the class imbalance issue, we selected 500 propaganda labels and 500 non-propaganda labels from the dataset and applied the fusion model.

This yielded a macro-F1 of 0.709 and an accuracy of 0.818. The hateful memes were further fine-grained into eight labels, while the non-hateful memes were fine-grained into two labels, as shown in Table 2. We applied the fusion model individually to the hateful and non-hateful fine-grained labels. The F1-score for the hateful fine-grained memes is 0.224, with an accuracy of 0.166, which is very low. This occurs because there are multiple labels within the hateful category, and the dataset is imbalanced for fine-grained hateful memes. In contrast, with only two labels in the non-hateful memes, the model performs better in classification, achieving a macro-F1 of 0.537 and an accuracy of 0.622.

6   Conclusion and Future Work

In this study, we investigate whether the content in propagandistic memes may contain hate. To do so, we used a multi-agent LLM-based approach to label propagandistic memes with coarse and fine-grained hate categories. We observed that there is a moderate agreement between an LLM agent (Claude 3.5 Sonnet) and human annotation. This led us to label propagandistic memes with coarse and fine-grained hate categories. We further used the dataset to train the model and evaluate its performance on the test set. The developed dataset is skewed in nature, which also reflects its classification performance. It is important to note that this attempt can enable the development of a large-scale dataset in a cost-effective manner. The issue of label imbalance can be resolved with an increase in data size. Future study will investigate this direction further by increasing data size and exploring open-sourced LLMs.

7   Acknowledgments

References

1. Abdelali, A., Hassan, S., Mubarak, H., Darwish, K., Samih, Y.: Pre-Training BERT on Arabic tweets: Practical considerations (2021)
2. Alam, F., Cresci, S., Chakraborty, T., Silvestri, F., Dimitrov, D., Martino, G.D.S., Shaar, S., Firooz, H., Nakov, P.: A survey on multimodal disinformation detection. In: Proceedings of the 29th International Conference on Computational Linguistics. pp. 6625–6643 (Oct 2022)
3. Alam, F., Hasnat, A., Ahmed, F., Hasan, M.A., Hasanain, M.: ArMeme: Propagandistic content in arabic memes. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (Dec 2024)
4. Alam, F., Mubarak, H., Zaghouani, W., Da San Martino, G., Nakov, P.: Overview of the WANLP 2022 shared task on propaganda detection in Arabic. In: Proceedings of the Seventh Arabic Natural Language Processing Workshop (Dec 2022)

5. Alam, F., Mubarak, H., Zaghouani, W., Da San Martino, G., Nakov, P.: Overview of the wanlp 2022 shared task on propaganda detection in arabic. In: Proceedings of the Seventh Arabic Natural Language Processing Workshop. pp. 108–115 (2022)

6. Antoun, W., Baly, F., Hajj, H.: AraBERT: Transformer-based model for Arabic language understanding. In: Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools. OSAC '20, Marseille, France (2020)

7. Chen, J., Wu, Y., Yang, X., Liu, J., Yang, H., Wang, Z.: Multi-agent reinforcement learning for dynamic content moderation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 4536–4544 (2021)

8. Da San Martino, G., Barrón-Cedeño, A., Wachsmuth, H., Petrov, R., Nakov, P.: Semeval-2020 task 11: Detection of propaganda techniques in news articles. In: Proceedings of the 14th International Workshop on Semantic Evaluation. pp. 1377–1414 (2020)

9. Da San Martino, G., Yu, S., Barrón-Cedeño, A., Petrov, R., Nakov, P.: 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. pp. 5636–5646. ACL (Nov 2019)

10. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Proceedings of the International AAAI Conference on Web and Social Media. AAAI '17, vol. 11 (2017)

11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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. pp. 4171–4186. ACL (Jun 2019)

12. Dimitrov, D., Alam, F., Hasanain, M., Hasnat, A., Silvestri, F., Nakov, P., Da San Martino, G.: SemEval-2024 task 4: Multilingual detection of persuasion techniques in memes. In: Proceedings of the 18th International Workshop on Semantic Evaluation. pp. 2009–2026. ACL, Mexico City, Mexico (Jun 2024)

13. Dimitrov, D., Ali, B.B., Shaar, S., Alam, F., Silvestri, F., Firooz, H., Nakov, P., Da San Martino, G.: 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. pp. 6603–6617. ACL-IJCNLP '21 (2021)

14. Dimitrov, D., Bin Ali, B., Shaar, S., Alam, F., Silvestri, F., Firooz, H., Nakov, P., Da San Martino, G.: Semeval-2021 task 6: Detection of persuasion techniques in texts and images. In: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021). pp. 70–98 (2021)

15. Fang, T., Xu, B., Zong, L.: Emotion-enhanced multi-modal persuasive techniques detection using split features. In: Fuzzy Systems and Data Mining VIII, pp. 172–180. IOS Press (2022)

16. Fortuna, P., Nunes, S.: A survey on automatic detection of hate speech in text. ACM Computing Surveys (CSUR) **51**(4), 1–30 (2018)

17. Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N.V., Wiest, O., Zhang, X.: Large language model based multi-agents: A survey of progress and challenges. arXiv preprint arXiv:2402.01680 (2024)

18. Hasanain, M., Ahmed, F., Alam, F.: Large language models for propaganda span annotation. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (Dec 2024)

19. Hasanain, M., Ahmed, F., Alam, F.: 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 (2024)

20. Hasanain, M., Alam, F., Mubarak, H., Abdaljalil, S., Zaghouani, W., Nakov, P., Da San Martino, G., Freihat, A.: ArAIEval shared task: Persuasion techniques and disinformation detection in Arabic text. In: Sawaf, H., El-Beltagy, S., Zaghouani, W., Magdy, W., Abdelali, A., Tomeh, N., Abu Farha, I., Habash, N., Khalifa, S., Keleg, A., Haddad, H., Zitouni, I., Mrini, K., Almatham, R. (eds.) Proceedings of ArabicNLP 2023. pp. 483–493. ACL, Singapore (Hybrid) (Dec 2023)

21. Hasanain, M., Hasan, M.A., Ahmed, F., Suwaileh, R., Biswas, M.R., Zaghouani, W., Alam, F.: ArAIEval Shared Task: Propagandistic techniques detection in uni-modal and multimodal arabic content. In: Proceedings of the Second Arabic Natural Language Processing Conference. ACL, Bangkok (Aug 2024)

22. Hossain, E., Sharif, O., Hoque, M.M., Preum, S.M.: Deciphering hate: Identifying hateful memes and their targets. arXiv preprint arXiv:2403.10829 (2024)

23. Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., Testuggine, D.: The hateful memes challenge: Detecting hate speech in multimodal memes. In: Advances in Neural Information Processing Systems. vol. 33, pp. 2611–2624 (2020)

24. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)

25. Mathias, L., Nie, S., Mostafazadeh Davani, A., Kiela, D., Prabhakaran, V., Vidgen, B., Waseem, Z.: Findings of the WOAH 5 shared task on fine grained hateful memes detection. In: Mostafazadeh Davani, A., Kiela, D., Lambert, M., Vidgen, B., Prabhakaran, V., Waseem, Z. (eds.) Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021). pp. 201–206. ACL, Online (Aug 2021)

26. Mubarak, H., Abdaljalil, S., Nassar, A., Alam, F.: Detecting and identifying the reasons for deleted tweets before they are posted. Frontiers in Artificial Intelligence **6**, 1219767 (2023)

27. OpenAI, R.: GPT-4 technical report. arXiv pp. 2303–08774 (2023)

28. Piskorski, J., Stefanovitch, N., Da San Martino, G., Nakov, P.: 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. pp. 2343–2361. ACL, Toronto, Canada (Jul 2023)

29. Schmidt, A., Wiegand, M.: A survey on hate speech detection using natural language processing. In: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. pp. 1–10. ACL, Valencia, Spain (Apr 2017)

30. Shah, U., Biswas, M.R., Agus, M., Househ, M., Zaghouani, W.: 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. pp. 467–472 (Aug 2024)

31. Sharma, S., Alam, F., Akhtar, M.S., Dimitrov, D., Da San Martino, G., Firooz, H., Halevy, A., Silvestri, F., Nakov, P., Chakraborty, T.: Detecting and understanding harmful memes: A survey. In: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence. pp. 5597–5606. IJCAI (7 2022)

32. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)