The NakbaArchiveClassifier Shared Task on Nakba Image Classification

Alexei Abrahams[1], Shadi Abudalfa[2], Mustafa Jarrar[1], George Mikros[1]
[1]Hamad bin Khalifa University
[2]King Fahd University of Petroleum & Minerals

Abstract

The proliferation of social media platforms has significantly reshaped how conflicts are documented, generating large-scale visual records that must be structured to enable meaningful analysis. In this paper, we present the NakbaArchiveClassifier shared task, which focuses on binary classification of infrastructure damage in images from Gaza. This task formed part of the Nakba-NLP Workshop at LREC 2026 and is grounded in an ongoing initiative focused on humanitarian archiving. It utilizes a carefully curated dataset of 2,001 images sourced from Palestinian journalists and content creators on Instagram, spanning the period from October 7, 2023 to December 15, 2025. The objective for participants was to classify whether an image depicts damaged or destroyed infrastructure versus intact structures. This task poses multiple challenges, such as the complexity of real-world conflict imagery, imbalance between classes, and the inherent ambiguity present in many visual scenes. The NakbaArchiveClassifier shared task introduces a new benchmark for analyzing conflict-related visual data and provides valuable resources for advancing research in humanitarian AI, crisis analytics, and Arabic digital humanities.

1. Introduction

The ongoing documentation of the Palestinian experience in Gaza has taken on new urgency since October 7, 2023, when a large-scale military campaign dramatically intensified the destruction of civilian infrastructure. Social media platforms (and Instagram in particular) have become primary channels through which Palestinian content creators and journalists bear witness to this destruction in near real time. The sheer volume of this visual documentation, with millions of images archived by the Tech for Palestine incubator project SaltPillar, presents both an extraordinary opportunity and a formidable challenge for researchers, humanitarian organizations, and archivists seeking to make these data accessible and analytically meaningful.

Automated image classification offers a promising pathway toward organizing such large-scale visual archives. By distinguishing images of destroyed or damaged infrastructure from images of intact infrastructure, a reliable classifier can help prioritize humanitarian assessments, support journalistic verification workflows, and enable systematic research into patterns of destruction. The NakbaArchiveClassifier shared task was designed with precisely these goals in mind, offering the research community a carefully curated dataset of 2,001 images drawn from Palestinian sources in Gaza and a well-defined binary classification objective. This shared task formed part of the Nakba-NLP Workshop at LREC 2026 (Jarrar et al., 2026).

Motivation

The term *Nakba* (Arabic for “catastrophe,” referring historically to the mass displacement of Palestinians in 1948 and more broadly to the ongoing dispossession of the Palestinian people) continues to unfold in visual form across social media. The SaltPillar archive captures this unfolding record, but raw archival scale alone is insufficient for research utility. Systematic categorization is a prerequisite for any downstream analysis, whether humanitarian, journalistic, or scholarly. Infrastructure damage is among the most analytically significant categories, as it speaks directly to the material conditions of life and survival in Gaza. Automating the detection of such damage in social media images would substantially reduce the manual annotation burden and accelerate the usability of the archive.

Challenges

NakbaArchiveClassifier presents unique challenges:

- **Domain specificity and visual complexity.** Images of conflict-affected infrastructure differ substantially from standard benchmark datasets. Destruction may be partial or obscured, lighting conditions are often poor, and images are frequently captured under duress, resulting in motion blur, occlusion, or unconventional framing that complicates automated analysis.

- **Class imbalance.** The dataset contains 706 images of destroyed or damaged infrastructure and 1,295 images of intact infrastructure, an imbalance that classifiers must address through appropriate sampling strategies or loss weighting.

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 91–97
11 May 2026.

- **Contextual ambiguity.** Some images are visually ambiguous (for example, rubble visible in the background of an otherwise ordinary scene, or intact buildings photographed in ways that imply surrounding damage), making binary labeling a non-trivial judgment even for human annotators.

- **Ethical and representational sensitivity.** Working with imagery of conflict and human suffering requires careful attention to the dignity of those depicted and the responsible use of automated systems trained on such material.

- **Out-of-distribution generalization.** The held-out test set of 500 images may capture temporal or thematic variation not fully represented in the training data, challenging models to generalize robustly across the full scope of the archive.

NakbaArchiveClassifier

NakbaArchiveClassifier[1] offers a binary image classification benchmark grounded in a real-world humanitarian archiving use case. Participants were provided with 2,001 labeled training images sourced from Palestinian content creators and journalists active on Instagram in Gaza between October 7, 2023 and December 15, 2025, and were evaluated on a held-out test set of 500 images. The classification objective is to distinguish images depicting destroyed or damaged infrastructure from those depicting infrastructure that remains intact.

The task received strong engagement from the global NLP and computer vision community:

- 30 teams registered.
- 16 unique submissions to the leaderboard.
- 14 countries are represented among the registered teams, including: Saudi Arabia, Pakistan, China, India, Finland, Egypt, Nigeria, Vietnam, Qatar, South Korea, Bangladesh, Morocco, Algeria, and Turkey.

NakbaArchiveClassifier contributes the first benchmark suite tailored for automated classification of conflict-related infrastructure damage in social media imagery from Gaza, establishing a replicable evaluation framework and a labeled dataset that can support future work in humanitarian AI, conflict documentation, and Arabic digital humanities.

[1]NakbaArchiveClassifier URL: https://nakba-nlp-2026.pillarofsalt.net/

2. Related Work

The NakbaArchiveClassifier task sits at the intersection of several active research areas: visual damage detection, humanitarian and crisis-related image analysis, social media content classification, and Arabic digital humanities. We review the most relevant prior work across these areas.

Damage and Disaster Detection in Images. Building-damage assessment from satellite and aerial imagery has been studied extensively in the context of natural disasters. A landmark contribution in this space is the xBD dataset (Gupta et al., 2019), which provides pre- and post-event satellite imagery across diverse disaster types, paired with ordinal damage labels and building polygons. With over 850,000 building annotations, xBD enabled large-scale evaluation of deep learning models for damage assessment. Transfer learning from architectures such as ResNet (He et al., 2016) and EfficientNet (Tan and Le, 2019) has become the dominant paradigm for these tasks, with models fine-tuned on disaster imagery achieving strong performance on held-out benchmarks (Alam et al., 2023a). More recently, transformer-based vision models (in particular the Vision Transformer (ViT) (Dosovitskiy et al., 2021)) have demonstrated competitive results on image classification tasks and are increasingly applied to specialized domains. However, satellite and aerial imagery differs markedly from the ground-level, social-media-sourced content that characterizes the NakbaArchiveClassifier dataset, with the latter exhibiting greater perspectival variability, visual noise, and contextual ambiguity.

Crisis and Conflict Image Analysis. A dedicated line of research addresses the automated analysis of crisis imagery shared on social media. The CrisisMMD dataset (Alam et al., 2018) introduced multimodal Twitter data from natural disasters with three annotation layers, enabling joint text-and-image classification for humanitarian response. Building on this foundation, Alam et al. (2021) consolidated eight human-annotated crisis-related datasets to facilitate comparative evaluation across informativeness and humanitarian classification tasks, while Alam et al. (2023b) benchmarked ten deep learning architectures for real-time crisis image classification across four task types. A comprehensive extension of this effort is the MEDIC dataset (Alam et al., 2023a), the largest social media image classification dataset for humanitarian response to date, covering 71,198 images across four tasks in a multi-task learning framework. NakbaArchiveClassifier extends this line of work to the specific domain of conflict and occupation in Gaza, with imagery drawn from a live archiving infrastructure and a binary classification objective centered on infrastructure damage.

Social Media Image Classification and Transfer Learning. General-purpose image classification has advanced rapidly with the development of large pretrained architectures. ResNet (He et al., 2016) established residual connections as a foundational design pattern that enabled the training of much deeper convolutional networks. EfficientNet (Tan and Le, 2019) later demonstrated that compound scaling of network depth, width, and input resolution can achieve strong accuracy with substantially fewer parameters. The Vision Transformer (Dosovitskiy et al., 2021) introduced a pure attention-based architecture that, when pretrained at scale, matches or surpasses convolutional baselines on standard image benchmarks. Contrastive pretraining on large image-text corpora, as in CLIP (Radford et al., 2021), has further expanded the repertoire of readily available representations, enabling effective zero-shot and few-shot transfer to downstream classification tasks. Fine-tuning these pretrained models on domain-specific data is now standard practice and has proven effective even in low-resource settings, making it directly relevant to the NakbaArchiveClassifier challenge, where the training set is modest in size and the imagery is domain-specific.

Arabic Digital Humanities and Arabic NLP. Computational approaches to Arabic-language materials have grown substantially in recent years. The development of AraBERT (Antoun et al., 2020), a BERT-based model pretrained specifically on large-scale Arabic text, demonstrated that language-specific pretraining substantially outperforms multilingual models on Arabic NLP benchmarks including sentiment analysis, named entity recognition, and question answering. Recent shared tasks have further pushed the state of the art in Arabic NLU: the ArabicNLU 2024 shared task (Khalilia et al., 2024) targeted word sense disambiguation and location mention disambiguation in Arabic, providing novel annotated resources alongside competitive evaluation. Work at the intersection of computational methods and Palestinian studies remains limited, however, making NakbaArchiveClassifier a contribution to an underserved area of Arabic digital humanities. The broader Nakba-NLP 2026 Workshop, of which this shared task is a part, represents a deliberate effort to develop language and vision resources attentive to Palestinian experience and narrative.

3. Data Collection and Selection

The SaltPillar archive consists of Instagram content, including millions of images and videos uploaded by content creators in Gaza (such as @wizard_bisan1 or @saleh.jafarawi) or shared by second-hand aggregators and observers (such as @eye.on.palestine or @palestine.pixel). The data were retrieved from Instagram using adversarial archiving techniques detailed in (Abrahams, 2026). These data constitute part of the collective memory of the Palestinian people, an archive for scholarly inquiry, and an evidence bank for war crimes investigation.

To be useful in these respects, however, the content of the archive must be sensibly labeled according to what they pertain to—airstrike, starvation, children, protest, resilience, medicine, and so forth. But due to the sheer size of the archive, it is infeasible to assign a team of humans the task of labeling every piece of content by hand. Moreover, the graphic and demoralizing nature of the content might plausibly have long-term consequences for the mental health of the labelers themselves. In lieu of this, machine learning classifiers can emulate the judgment of human expert labelers by training on a comparatively small sample of labeled content, then extrapolate to predict the labels of the remaining content.

We proceeded by randomly sampling 5,000 images from the SaltPillar archive for the period October 7, 2023 to December 15, 2025. In particular, we restricted our attention to a list of over 600 content creators either based in Gaza, or whose Instagram activity centered on sharing content originating from Gaza.

By hand, we then labeled these images according to whether they exhibited signs of infrastructure damage or not. We filtered out sensitive images, such as highly graphic content or content that too clearly identified individuals. We also filtered out irrelevant content, such as screenshots of tweets. In all, 60% of the sample was filtered out, leaving a sample of 2,001 images: 706 images of destroyed or damaged infrastructure, and 1,295 images of intact infrastructure.

4. Task Description and Evaluation

In this section, we describe in detail the shared. The shared task was conducted in two stages: a development phase and a final evaluation phase. During the development phase, participants tested and refined their models using a validation set. In contrast, the final phase relied on a held-out test set with concealed labels. All submissions were assessed via the **CodaBench** platform, an open-source infrastructure for benchmarking AI systems.

CodaBench enables the definition of tasks, including datasets, submission workflows, and evaluation criteria, and offers an automated and transparent environment for assessing participant models.

In this shared task, participants are required to develop a system that performs binary image classification to determine whether an image depicts destroyed or damaged infrastructure or intact infrastructure. The task focuses on real-world imagery collected from social media in Gaza, presenting unique challenges related to visual complexity, ambiguity, and domain specificity.

Formally, the objective is to design an image classifier C such that, given an input image x, the system predicts a label y where:

$$y = C(x), \quad y \in \{\text{Destruction}, \text{Not\_destruction}\}$$

- **Input**: An image collected from social media depicting infrastructure in Gaza.
- **Output**: A binary label indicating whether the image shows {Damaged, Intact} infrastructure.
- **Evaluation Metric**: Systems are evaluated using standard classification metrics, including Accuracy and F1-score, with the primary metric being the **Macro F1-score**. This metric is particularly appropriate due to class imbalance and ensures equal weighting of both classes.

5. Participant Systems

A total of 16 participants submitted systems to the NakbaArchiveClassifier shared task. Below, we provide a summary of selected participating systems, linking participant names to their leaderboard identifiers, modeling approaches, and reported performance:

- **HCMUS_TheFangs**: The top-ranked team proposed a comprehensive evaluation of multiple modern architectures, including EfficientNet, Vision Transformers, and MaxViT. Their best-performing system employed a MaxViT-Base backbone combined with focal loss, MixUp augmentation, and a tailored augmentation pipeline to address dataset bias and imbalance. Their approach achieved a Macro F1-score of 0.899, securing **1st place** on the leaderboard.
- **rahaf_jaber**: This system utilized a ConvNeXt-Tiny backbone trained using stratified 5-fold cross-validation, combined with weighted loss functions and test-time augmentation. An ensemble of fold-specific models was used during inference. The system achieved a Macro F1-score of 0.8952, ranking **2nd**.
- **zahira_blr**: The authors proposed a ConvNeXt-Tiny–based ensemble with focal loss, exponential moving average (EMA), and threshold calibration. Their approach incorporated out-of-fold validation and test-time augmentation. The system achieved a Macro F1-score of 0.8893, placing **3rd**.
- **salmakh**: This team employed a CLIP-based transfer learning approach, fine-tuning a Vision Transformer encoder with a lightweight classification head. Their method addressed class imbalance using focal loss and achieved a Macro F1-score of 0.877, ranking **4th**.
- **mohamedfathy**: Their system relied on a Swin Transformer architecture combined with transfer learning and semi-supervised data expansion to improve generalization. The model achieved strong performance with a Macro F1-score of approximately 0.865, ranking **5th**.
- **KvochurHegel**: This team used a ConvNeXt-V2 Nano backbone with extensive data augmentation and label smoothing. A multi-view test-time augmentation strategy was applied during inference. Their system achieved a Macro F1-score of 0.8593, ranking **6th**.
- **AyahVerse** (Username: *ibad-ur-rehman-546335*): The team evaluated multiple architectures, including ResNet50, MobileNetV2, and EfficientNet-B0, alongside threshold optimization strategies. Their findings highlighted the effectiveness of lightweight models and calibration techniques, achieving an F1-score of approximately 0.85.
- **men_na_aly**: This system adopted a ResNet-based transfer learning approach with weighted cross-entropy loss to address class imbalance. Their final model achieved a Macro F1-score of 0.8323 and demonstrated the effectiveness of lightweight CNN baselines.
- **kannanrrk**: The authors utilized an EfficientNet-B3 backbone with transfer learning, achieving a weighted F1-score of 0.8387 and demonstrating strong performance with relatively efficient architectures.
- **nw_123**: This submission proposed an ensemble of EfficientNet models combined with an SVM meta-classifier for decision fusion. The approach emphasized computational efficiency and achieved competitive performance under constrained settings.
- **heba13**: The team evaluated multiple CNN architectures, including ResNet and EfficientNet

variants, using transfer learning and standard preprocessing. Their results highlighted the effectiveness of moderate-depth models in low-resource settings.

- **m_elhussieni**: This submission used a ResNet-50 backbone with aggressive grayscale augmentation and test-time augmentation to address the challenges of archival imagery. The system achieved a Macro F1-score of approximately 0.45, serving as a baseline under challenging conditions.

6. Results

Table 1 summarizes the final standings of the NakbaArchiveClassifier shared task, where systems are ordered based on their Macro F1-score evaluated on the held-out test data. The results demonstrate generally strong performance across participants, with HCMUS_TheFangs achieving the highest score of 0.8991, followed closely by rahaf_jaber and zahira_blr. Most systems obtained F1-scores exceeding 0.80, indicating that current image classification methods can effectively address this task despite its complexity. However, a clear drop in performance is observed among the lower-ranked entries, suggesting differences in model generalization and robustness when dealing with challenging, real-world conflict imagery. Overall, these results highlight both the competitive nature of the task and the advancements made by participants in tackling specialized visual classification problems.

#	Participant Name	Macro F1-Score
1	HCMUS_TheFangs	0.8991
2	rahaf_jaber	0.8952
3	zahira_blr	0.8893
4	salmakh	0.8767
5	mohamedfathy	0.8655
6	KvochurHegel	0.8593
7	AyahVerse	0.8545
8	mohamedabdallah	0.8465
9	nw_123	0.8416
10	kannanrrk	0.8389
11	men_na_aly	0.8323
12	Md.Ajwad Hossain	0.8312
13	mealper	0.8236
14	heba13	0.8142
15	faisalm93	0.4506
16	m_elhussieni	0.3965

Table 1: Leaderboard for NakbaArchiveClassifier Shared Task on Nakba Image Classification.

7. Discussion

The outcomes of the NakbaArchiveClassifier shared task indicate that contemporary computer vision methods can effectively tackle the problem of identifying infrastructure damage in conflict imagery, even when operating under difficult real-world conditions. Leading submissions achieved Macro F1-scores close to 0.90, reflecting strong classification performance despite the dataset's limited size, class imbalance, and the inherently ambiguous nature of the images.

An important trend observed across the leaderboard is the strong performance of transformer-based and hybrid architectures, including MaxViT, ConvNeXt, and Vision Transformers. These models appear particularly capable of capturing intricate spatial relationships and contextual signals in complex scenes, where evidence of damage may be partial, subtle, or obscured. Moreover, the extensive reliance on transfer learning emphasizes the value of pretrained representations, particularly in specialized domains where labeled data are scarce.

Handling class imbalance also emerged as a critical factor in achieving competitive performance. Many high-ranking systems incorporated focal loss, class weighting, or tailored sampling approaches, which improved detection of the minority class representing damaged infrastructure. This is evident in the consistently high Macro F1-scores, suggesting that participants were able to reduce bias toward the more frequent class.

In addition, training strategies such as data augmentation and model ensembling proved highly beneficial. Methods like MixUp, test-time augmentation, and cross-validation-based ensembles contributed to improved robustness and generalization, especially given the noisy and heterogeneous nature of social media images. These results highlight the necessity of carefully designed training pipelines when working with real-world, unstructured datasets.

Nevertheless, the noticeable gap between top-performing and lower-performing systems points to unresolved challenges. Some approaches achieved substantially lower scores, in certain cases below 0.50 Macro F1, indicating difficulties in generalizing across the dataset. These shortcomings may be attributed to sensitivity to domain variation, limited capacity to resolve ambiguous cases, or insufficient optimization. This disparity suggests that, while feasible, the task remains challenging, particularly for simpler or less refined models.

The task also underscores the ambiguity inherent in the labeling scheme. Differentiating between damaged and intact infrastructure is not always straightforward, even for human annotators. As a result, models may rely on indirect cues or contextual patterns that do not perfectly align with human interpretation. This raises broader concerns regarding annotation consistency, potential label noise, and the adequacy of binary labels for representing complex humanitarian scenarios.

More broadly, the shared task illustrates both the potential benefits and the constraints of applying AI techniques to humanitarian archives. Automated classification can substantially reduce the burden of manual annotation and improve access to large-scale datasets. However, such systems must be deployed responsibly, taking into account ethical considerations such as misclassification risks, bias in representation, and the sensitive nature of the underlying content.

8. Limitations

Despite the contributions of the NakbaArchiveClassifier shared task, several limitations should be considered when interpreting its findings and potential applications.

To begin with, the dataset size is relatively limited. Although the 2,001 annotated images were carefully selected and curated, this number is small compared to large-scale benchmarks commonly used in computer vision. Consequently, model performance may rely substantially on transfer learning, data augmentation, and regularization techniques, and the dataset may not fully represent the variability of infrastructure-related imagery in Gaza.

In addition, the data collection is confined to Instagram and a specific subset of Palestinian content creators and aggregators. As a result, models trained on this dataset may face challenges when applied to images from different platforms, geographic contexts, or documentation environments.

Another limitation arises from the binary nature of the classification task. In practice, infrastructure damage is not strictly dichotomous but exists along a continuum, from minor deterioration to complete destruction. Some images may also present overlapping or unclear visual evidence. Although the binary setup simplifies annotation and evaluation, it reduces the ability to capture fine-grained distinctions that could be valuable for more detailed analysis.

Furthermore, the annotation process inherently involves a degree of subjectivity. Even trained annotators may differ in their judgments when evaluating ambiguous or borderline cases, particularly when damage is partially visible, obscured, or context-dependent. This introduces the possibility of label noise within the dataset, which should be taken into account when interpreting system performance.

The filtering stage also imposes constraints on the dataset's representativeness. Images deemed highly graphic or those clearly identifying individuals were excluded to ensure ethical compliance and protect human dignity. While this is a necessary and responsible decision, it limits the extent to which the dataset reflects the full spectrum of visual material present in the original archive.

Finally, system evaluation is primarily based on aggregate metrics such as Macro F1-score. Although suitable for benchmarking and ranking submissions, these metrics do not fully capture aspects that are crucial for real-world deployment, including model calibration, resilience to distribution shifts, interpretability, and fairness across different image types. Future work could benefit from more comprehensive evaluation frameworks that better align with humanitarian and archival requirements.

9. Conclusion

This overview paper introduced the NakbaArchiveClassifier shared task, which focuses on the problem of classifying social media images based on the presence of infrastructure damage. Built on real-world imagery from Gaza, the task emphasizes several inherent challenges, including ambiguity in visual cues, domain-specific characteristics, and imbalanced class distributions. By evaluating participating systems within this setting, the task aims to promote progress in analyzing conflict-related visual data and strengthening the reliability of computer vision approaches in humanitarian applications.

We anticipate that NakbaArchiveClassifier will provide a valuable benchmark for future research in areas such as digital archiving, crisis-focused AI, and humanitarian computing, while fostering the development of more robust, context-aware, and ethically responsible image classification models.

10. Bibliographical References

Alexei Abrahams. 2026. *Social Media Exposed: Track Disinformation, Archive War Crimes, Restore Oversight*. No Starch Press.

Firoj Alam, Tanvirul Alam, Md Arid Hasan, Abul Hasnat, Muhammad Imran, and Ferda Ofli. 2023a. MEDIC: a multi-task learning dataset for disaster image classification. *Neural Computing and Applications*, 35(3):2609–2632.

Firoj Alam, Tanvirul Alam, Ferda Ofli, and Muhammad Imran. 2023b. Robust training of social media image classification models. *IEEE Transactions on Computational Social Systems*, 10(5):2507–2518.

Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018. CrisisMMD: Multimodal twitter datasets from natural disasters. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, pages 465–473. AAAI Press.

Firoj Alam, Hassan Sajjad, Muhammad Imran, and Ferda Ofli. 2021. CrisisBench: Benchmarking crisis-related social media datasets for humanitarian information processing. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, pages 923–932.

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, Marseille, France. European Language Resources Association.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.

Ritwik Gupta, Bryce Goodman, Nirav Patel, Ricky Hosfelt, Sandra Sajeev, Eric Heim, Jigar Doshi, Keane Lucas, Howie Choset, and Matthew Gaston. 2019. Creating xBD: A dataset for assessing building damage from satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 10–17.

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)*, pages 770–778.

Mustafa Jarrar, Mo El-Haj, Amal Haddad, Serin Atiani, Shadi Abudalfa, Khalil Sima'an, Paul Rayson, and Camille Mansour, editors. 2026. *Proceedings of the second International Workshop on Nakba Narratives as Language Resources*. Association for Computational Linguistics, Spain.

Mohammed Khalilia, Sanad Malaysha, Reem Suwaileh, Mustafa Jarrar, Alaa Aljabari, Tamer Elsayed, and Imed Zitouni. 2024. ArabicNLU 2024: The first arabic natural language understanding shared task. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 361–371, Bangkok, Thailand. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Mingxing Tan and Quoc V. Le. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR.