The NakbaVirality Shared Task on MultimodalVirality Prediction in High-Stakes Discourse

Saad Ezzini[1], Salima Lamsiyah[2], Shadi Abudalfa[1],
Samir El-Amrany[2], Walid Alsafadi[3]
[1]King Fahd University of Petroleum & Minerals
[2]University of Luxembourg
[3]University College of Applied Sciences

Abstract

Social media virality significantly shapes public discourse during geopolitical conflicts, where emotionally charged and multimodal content can rapidly gain widespread attention. However, most prior approaches rely on retrospective engagement signals, limiting their usefulness for early prediction. Multimodal virality modeling in high-stakes Arabic discourse remains largely unexplored. We introduce *NakbaVirality*, a shared task on multimodal virality classification in conflict-related social media posts, organized as part of the NakbaNLP workshop at LREC 2026. The dataset consists of 2,600 anonymized posts from X and Reddit collected after October 7, 2023, each including text, an associated image, and normalized engagement labels. Participants must classify posts into low, medium, or high virality categories using only textual and visual inputs. The task provides standardized splits, baseline systems, and evaluation using macro-F1 and accuracy. *NakbaVirality* establishes the first benchmark for multimodal virality prediction in Arabic high-stakes discourse and promotes research on contextual and multimodal modeling for early impact prediction. The shared task attracted 18 participants, who contributed a total of 5 official test phase submissions.

1. Introduction

Social media platforms have become central to information diffusion during geopolitical conflicts, where emotionally charged narratives and symbolic imagery can rapidly influence public discourse. Understanding what makes content "go viral" has therefore attracted significant attention across computational social science and machine learning. Early studies showed that community structure and diffusion topology strongly influence cascade formation and meme spread (Doerr et al., 2012; Weng et al., 2013; Cheung et al., 2016), while psychological factors such as emotional intensity also contribute to online sharing dynamics (Berger and Milkman, 2012). More recent approaches incorporate multi-feature fusion and deep learning models to forecast popularity using temporal and engagement signals (Gao et al., 2021; Kowalczyk and Larsen, 2018; Xu and Qian, 2023; Zhang and Gao, 2024). However, these methods largely rely on retrospective interaction patterns (e.g., early likes, retweets, diffusion graphs), which are unavailable at publication time.

Advances in representation learning have significantly improved content modeling capabilities. Transformer-based language models such as BERT (Devlin, 2018) and multimodal architectures such as CLIP (Radford et al., 2021) enable joint reasoning over textual and visual signals. Multimodal fusion approaches have demonstrated improvements in related tasks including misinformation detection and popularity prediction (Singhal et al., 2019; Riis et al., 2020). Despite this progress, publicly available benchmarks for multimodal virality prediction in high-stakes discourse remain limited, particularly for Arabic. While Reddit-V (Elamrany et al., 2025) introduced a dataset for pre-engagement virality prediction, it did not specifically target conflict-driven Arabic discourse.

The discourse surrounding the Nakba and related regional discussions constitutes a high-impact setting marked by historical references, polarized sentiment, and multimodal communication. In such contexts, images often amplify textual rhetoric, making multimodal alignment critical for predicting reach. Modeling this content requires capturing contextual and symbolic cues in Arabic, which presents additional challenges including rich morphology, dialectal variation, and code-switching. Furthermore, viral posts typically form a minority class, requiring robust evaluation under class imbalance using metrics such as macro-F1 (Powers, 2011).

To address these challenges, we introduce *NakbaVirality*[1], the first shared task on multimodal virality classification in Arabic high-stakes discourse. This shared task formed part of the NakbaNLP Workshop at LREC 2026 (Jarrar et al., 2026). The presented dataset consists of 2,600 anonymized posts collected from X and Reddit after October 7, 2023. Each instance includes post text, an associated image, and normalized engagement labels. Participants are required to classify posts into low, medium, and high virality categories using only textual and visual inputs. The task provides standardized splits, baseline systems grounded in

[1]NakbaVirality URL: https://ezzini.github.io/NakbaVirality

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 98–103

transformer and convolutional architectures, and evaluation using macro-F1 and accuracy.

The shared task received strong engagement from the global NLP community, with 18 participating teams in total and 5 unique final submissions to the leaderboard. *NakbaVirality* establishes the first benchmark suite for multimodal virality prediction in Arabic high-impact discourse and promotes research on contextual modeling, multimodal alignment, and early impact prediction in real-world social media settings.

2. Related Work

Foundations of Virality and Diffusion Modeling. Early research on virality prediction has emphasized structural and psychological drivers of information diffusion. Prior studies have shown that cascade topology, community structure, and user susceptibility have strongly influenced meme spread and rumor propagation (Doerr et al., 2012; Weng et al., 2013; Cheung et al., 2016; Hoang and Lim, 2016). Complementary work has demonstrated that emotional intensity, sentiment, and social signaling have played a central role in driving online sharing behavior (Berger and Milkman, 2012; Tsugawa and Ohsaki, 2017; Mousavi et al., 2022). Subsequent research has incorporated temporal cascade features and engagement trajectories to forecast popularity at early stages (Shamma et al., 2011; Lu and Szymanski, 2018; Gao et al., 2021; Xu and Qian, 2023). More recently, graph-based and multi-task learning frameworks have modeled virality jointly with rumor detection and user vulnerability analysis (Zhang and Gao, 2024; Esteban-Bravo et al., 2024). However, these approaches have largely relied on early interaction signals (e.g., retweets, likes, diffusion graphs), which have not been available at publication time and therefore have limited purely content-based early prediction.

Content-Based and Multimodal Virality Prediction. To overcome the limitations of diffusion-dependent models, a growing body of work has focused on predicting virality from content alone. Early NLP-based studies have explored lexical, stylistic, and affective cues associated with viral text (Guerini et al., 2011). Subsequent research has proposed scalable content-based prediction models for news and social media posts (Lu and Szymanski, 2018; Kowalczyk and Larsen, 2018). With advances in deep learning, transformer-based approaches have been applied to virality prediction, including RoBERTa-based models for news tweets (Maldonado-Sifuentes et al., 2021) and user-aware architectures such as ViralBERT (Rameez et al., 2022). In the visual domain, attention mechanisms and deep visual representations have been leveraged to model video and image popularity (Bielski and Trzcinski, 2018; Riis et al., 2020), while multimodal frameworks have combined textual and visual signals to improve performance in related prediction settings (Singhal et al., 2019). More recently, large-scale vision–language pretraining models such as CLIP have enabled joint multimodal representation learning applicable to popularity and virality modeling (Radford et al., 2021). Nevertheless, multimodal virality benchmarks have remained limited, particularly beyond English and outside general-domain social media contexts.

Large Language Models and Arabic Virality Benchmarks. Recent work has begun to explore large language models for content-based virality prediction. In particular, Reddit-V has introduced a benchmark for pre-engagement virality classification and has evaluated zero-shot large language models for predicting popularity without relying on diffusion signals (El-amrany et al., 2025). Beyond this effort, existing virality datasets and evaluations have largely focused on English or general-domain social media content, and have not specifically addressed multimodal Arabic discourse in conflict-driven settings. The proposed *NakbaVirality* shared task has aimed to fill this gap by establishing the first benchmark dedicated to multimodal, content-based virality prediction in Arabic high-stakes discourse.

3. Data Collection and Selection (will be enhanced)

NakbaVirality targets conflict-related posts published after October 7, 2023 and pairs each text with an associated image and engagement-derived labels for multimodal virality classification. The data collection pipeline is shown in Figure 1.

3.1. Sources and acquisition

Reddit platform. Data from Reddit were collected from subreddits that include regular discussion of Gaza / Palestine and related topics. Data from each subreddit was collected according to the same general procedure: the subreddit was obtained (either from a curated list, or from a predefined list as a fallback); the posts within the subreddit that included one or more keywords from a multilingual list were searched; it was confirmed that each post was within the scope of the collection; and then the data from each relevant post that included at least one image was collected from the post. For each collected post, its identifier, name of its subreddit, date/time of posting, and various basic statistics associated with the post were collected. The image itself and its metadata were also collected.

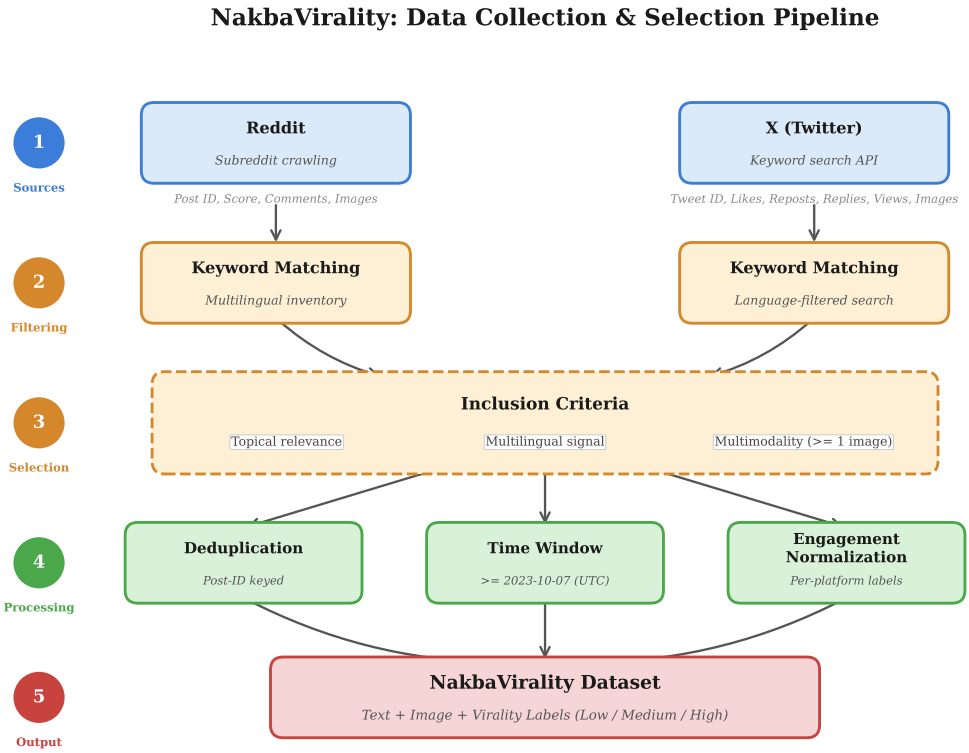

Figure 1: Data collection pipeline.

X platform. X data was collected through keyword search on the platform. The X search platform allows searches to be filtered by language. Consequently, searches were performed with the keyword list in each relevant language; search results were collected and posted in separate datasets for each language; and duplicates were removed to ensure that every post is represented only once in the dataset. The searches were filtered to include only posts published between October 7, 2023 and the present date. For each collected post, its identifier, text, date/time of posting, various statistics associated with the post, and any media linked to the post were collected.

3.2. Inclusion criteria

To make the dataset more relevant to multilingual and multimodal analysis, posts must meet the following inclusion criteria:

1. **Topical relevance :** The title and body of the post on Reddit, or the tweet text on X, must match at least one keyword from our inventory.

2. **Multilingual signal :** A post is retained if it matches at least one term from our multilingual keyword inventory. The corpus consists mostly of English (60%) and Arabic (10%) posts, with other languages including Farsi, Hebrew, Spanish, and Turkish comprising the remainder. Figure 2 reports the language distribution of the retained dataset.

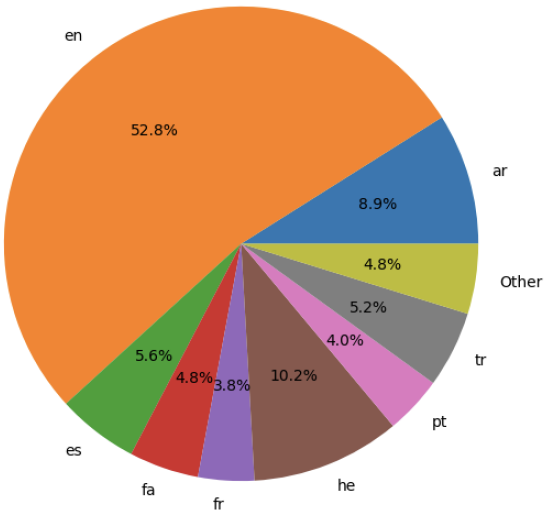

Figure 2: Language distribution of the collected posts in NakbaVirality.

3. Multimodality: the post has at least one image. For Reddit, this can be directly linked images or those uploaded to Reddit. For X, this requires

parsing the tweet card to find image URLs.

3.3. Deduplication and normalization

We remove duplicates at the *post* level using platform-native identifiers: (i) Reddit submissions are keyed by `Post ID`, and (ii) X posts are keyed by `tweet_id`. During crawling, seen identifiers are tracked to avoid collecting the same post multiple times across queries or sources.

We preserve the raw engagement fields provided by each platform to keep label construction reproducible. For Reddit, we keep the submission score and discussion volume (e.g., comment count), together with basic post metadata. For X, we keep the observed interaction counters (likes, reposts, replies, and views when available) and we additionally compute an engagement score as

$$E = \text{likes} + \text{reposts} + \text{replies}.$$

Because engagement magnitudes differ substantially across platforms, virality is not compared using raw counts directly. Instead, labels are derived by normalizing engagement *within each platform* and mapping posts into the three shared-task categories (low / medium / high).

3.4. Time window selection

All instances are restricted to a shared-task time window with a fixed lower bound of **2023–10–07** (UTC). For X, this lower bound is enforced at query time; for Reddit, the initial crawl may return older content, so the time constraint is applied during the final filtering step.

In the current data snapshot, the retained timestamps span: (i) **X:** 2023–10–07 to 2025–12–13 (UTC), (ii) **Reddit:** 2023–10–07 to 2025–11–04 (UTC). Posts outside the time window are excluded from the final release.

4. Evaluation Track

A total of 4 teams submitted their predictions to the shared task. Table 1 outlines the official leaderboard, evaluating the submissions primarily on Macro F1-Score, supplemented by Accuracy. The teams employed a wide range of analytical methodologies ranging from simplistic machine learning workflows to foundational multimodal models and recursive prompt-tuning techniques.

4.1. Participant Systems

HCMUS_TheFangs: The winning system deliberately avoided the complexities of deep multimodal learning by formulating virality prediction strongly on sociological structure. They extracted the source community structure and encoded it utilizing a sophisticated Bayesian smoothed target-encoding feature. Combining this single powerful feature with text-level descriptors (Character count, Word count, Hashtag count) and simple textual TF-IDF tokenization, they trained an XGBoost classifier. This streamlined approach surprisingly overshadowed all deep learning methods, emphasizing that virality during conflict heavily depends on community amplification factors and user biases, far more than on semantic properties.

Digilians: The Digilians team approached the task through a complex multimodal neural architecture, integrating both XLM-RoBERTa for multilingual textual processing and a Vision Transformer (ViT-B/16) mechanism. They significantly minimized semantic gaps by injecting a bidirectional cross-attention fusion layer that mutually conditioned text and image features against each other. Their implementation addressed severe class imbalances by leveraging selective Flan-T5 text paraphrasing on minority samples and applying a Focal Loss function inversely weighted to class frequencies.

xin1212: This anonymous submission implemented a frozen multimodal paradigm. Retaining the parameters of the base LAION CLIP-ViT, they integrated a custom built, lightweight residual adapter directly positioned over the concatenated text-image embeddings. Followed by a three-way MLP classifier layer, this frozen approach mitigated overfitting typical of data-constrained tasks and stabilized optimization in the multimodal learning domain.

ashhadulislam (Pushing_Boundaries): In contrast to explicitly engineering architectural models, this submission pioneered Recursive Prompt Improvement (RPI) in a parameter-free manner using prominent large language models (DeepSeek, GPT-5.2, and Qwen2.5-VL). They utilized diagnostic signals from top misclassified texts in each iteration and iteratively revised classification meta-prompts. They demonstrated that prompt tuning progressively clarifies separation boundaries among the different viral classes with zero neural weight updates.

5. Discussion

The differing outcomes of the submitted systems bring significant insights concerning multimodal virality prediction in emotionally charged and polarizing contexts. Most notably, standard state-of-the-art vision-language fusion models generally underperformed relative to the non-neural system pro-

Rank	Team	Macro F1	Accuracy
1	HCMUS_TheFangs	**0.7062**	**0.7305**
2	Digilians (noureldeen)	0.4983	0.6009
3	xin1212	0.4559	0.6089
4	ashhadulislam	0.3199	0.4392

Table 1: Official NakbaVirality System Leaderboard.

vided by HCMUS_TheFangs, suggesting an inherent "visual homogeneity" where explicit visual content does not directly determine virality. Rather, historical community context and audience ideological affiliations dictate how rapidly information disseminates. Furthermore, complex representations and explicit hashtags were found to decrease organic viral propensity, reaffirming the "Less is More" phenomenon during high-stakes conflict propagation.

6. Conclusion and Future Work

The *NakbaVirality* shared task successfully introduced the first benchmark specifically dedicated to multimodal virality prediction within Arabic high-stakes conflict discourse. With participation from multiple international teams, the evaluation process revealed a crucial insight: in politically charged contexts, historical community behavior and audience dynamics often exhibit greater predictive power than the semantic or visual content of the posts themselves. Deep multimodal architectures struggled with the inherent visual homogeneity of conflict imagery, demonstrating the vital necessity of integrating sociological factors into virality models. Future iterations of this task will focus on expanding the dataset across additional social media platforms, capturing temporal sentiment shifts, and incorporating zero-shot prompt-based evaluation for larger foundation models to further explore early-stage virality detection.

Limitations

While this shared task provides a foundational benchmark, several limitations must be acknowledged. First, the dataset is restricted to a specific time window following October 7, 2023, and may not fully capture the evolving linguistic and visual strategies used in long-term geopolitical conflicts. Second, to ensure strict ethical compliance and protect user privacy given the highly sensitive nature of the discourse, critical network diffusion features such as explicit user identities and follower graphs were anonymized or excluded. This restricts the ability to perform deep cascade or topology-based virality analysis. Finally, the inherent rarity of truly viral content results in a severe class imbalance, significantly complicating the training process and constraining the generalization capabilities of standard predictive models.

7. Bibliographical References

Jonah Berger and Katherine L Milkman. 2012. What makes online content viral? *Journal of marketing research*, 49(2):192–205.

Adam Bielski and Tomasz Trzcinski. 2018. Pay attention to virality: understanding popularity of social media videos with the attention mechanism. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2335–2337.

Ming Cheung, James She, Alvin Junus, and Lei Cao. 2016. Prediction of virality timing using cascades in social media. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 13(1):1–23.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Benjamin Doerr, Mahmoud Fouz, and Tobias Friedrich. 2012. Why rumors spread so quickly in social networks. *Communications of the ACM*, 55(6):70–75.

Samir El-amrany, Matthias R. Brust, Salima Lamsiyah, and Pascal Bouvry. 2025. Reddit-V: A virality prediction dataset and zero-shot evaluation with large language models. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 334–341, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Mercedes Esteban-Bravo, Jose M Vidal-Sanz, et al. 2024. Predicting the virality of fake news at the early stage of dissemination. *Expert Systems with Applications*, 248:123390.

Liqun Gao, Yujia Liu, Hongwu Zhuang, Haiyang Wang, Bin Zhou, and Aiping Li. 2021. Public opinion early warning agent model: A deep learning cascade virality prediction model based on

multi-feature fusion. *Frontiers in Neurorobotics*, 15:674322.

Marco Guerini, Carlo Strapparava, and Gozde Ozbal. 2011. Exploring text virality in social networks. In *proceedings of the international AAAI conference on web and social media*, volume 5, pages 506–509.

Tuan-Anh Hoang and Ee-Peng Lim. 2016. Tracking virality and susceptibility in social media. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 1059–1068.

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.

Damian Konrad Kowalczyk and Jan Larsen. 2018. Scalable privacy-compliant virality prediction on twitter. *arXiv preprint arXiv:1812.06034*.

Xiaoyan Lu and Boleslaw K Szymanski. 2018. Scalable prediction of global online media news virality. *IEEE Transactions on Computational Social Systems*, 5(3):858–870.

Christian E Maldonado-Sifuentes, Jason Angel, Grigori Sidorov, Olga Kolesnikova, and Alexander Gelbukh. 2021. Virality prediction for news tweets using roberta. In *Mexican International Conference on Artificial Intelligence*, pages 81–95. Springer.

Maryam Mousavi, Hasan Davulcu, Mohsen Ahmadi, Robert Axelrod, Richard Davis, and Scott Atran. 2022. Effective messaging on social media: What makes online content go viral? In *Proceedings of the ACM web conference 2022*, pages 2957–2966.

David Powers. 2011. Evaluation: From precision, recall and f-measure to roc, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1):37–63.

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. arXiv preprint arXiv:2103.00020.

Rikaz Rameez, Hossein A Rahmani, and Emine Yilmaz. 2022. Viralbert: A user focused bert-based approach to virality prediction. In *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, pages 85–89.

Christoffer Riis, Damian Konrad Kowalczyk, and Lars Kai Hansen. 2020. On the limits to multimodal popularity prediction on instagram–a new robust, efficient and explainable baseline. *arXiv preprint arXiv:2004.12482*.

David Shamma, Jude Yew, Lyndon Kennedy, and Elizabeth Churchill. 2011. Viral actions: Predicting video view counts using synchronous sharing behaviors. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, pages 618–621.

Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE fifth international conference on multimedia big data (BigMM)*, pages 39–47. IEEE.

Sho Tsugawa and Hiroyuki Ohsaki. 2017. On the relation between message sentiment and its virality on social media. *Social network analysis and mining*, 7(1):19.

Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. 2013. Virality prediction and community structure in social networks. *Scientific reports*, 3(1):1–6.

Zhixuan Xu and Minghui Qian. 2023. Predicting popularity of viral content in social media through a temporal-spatial cascade convolutional learning framework. *Mathematics*, 11(14):3059.

Xuan Zhang and Wei Gao. 2024. Predicting viral rumors and vulnerable users with graph-based neural multi-task learning for infodemic surveillance. *Information Processing & Management*, 61(1):103520.