Nakba Discourse 2025: A Bilingual Social Media Dataset for Collective Trauma Analysis

Wajdi Zaghouani[1], **Mabrouka Bessghaier**[1], **Kais Attia**[2]
[1]Northwestern University in Qatar
[2] Independent Researcher, Qatar
wajdi.zaghouani@northwestern.edu, mabrouka.bessghaier@northwestern.edu
kais.attia.w@gmail.com

Abstract

We introduce Nakba Discourse 2025, a bilingual full-year social media dataset capturing Arabic and English discourse about the 1948 Palestinian Nakba across Twitter/X and Facebook from January to December 2025. The corpus contains 70,312 unique posts organized into intersecting sub-corpora by language, sentiment, gender, geography, and platform, with engagement metadata and automatically extracted rhetorical features. Analyses reveal systematic variation in engagement and framing across communities. Per-post engagement is highest in Israel and UK subsets (50.62 and 49.08 average likes respectively), while Arabic-language discourse shows markedly lower per-post engagement. Sentiment distribution is strongly skewed, with negative sentiment posts outnumbering positive ones at an 11:1 ratio (54,424 vs. 4,827 posts). Despite dramatic variation in absolute engagement levels, virality rates remain structurally constant at approximately 10% across all Twitter/X sub-corpora, regardless of language, gender, or geography, pointing to platform-level amplification regularities. Gender analysis reveals that women achieve proportional virality equal to men despite producing roughly one-third the volume of posts. Temporal patterns align with cultural calendars, including Thursday peaks associated with Jumu'ah across Arabic and female subsets, and Sunday peaks in English-language subsets reflecting Western media cycles. The dataset will be released for research use and supports multilingual stance detection, virality modeling, rhetorical analysis, and computational studies of digital political memory.

Keywords: Nakba, social media dataset, political memory, rhetorical framing, virality, cross-lingual NLP, digital humanities

1. Introduction

The Nakba (النكبة), meaning "catastrophe" in Arabic, refers to the mass displacement of approximately 750,000 to 800,000 Palestinians during the 1948 Arab-Israeli war, resulting in the destruction of over 531 villages and the transformation of Palestinian society (Khalidi, 2020; Sa'di and Abu-Lughod, 2007; Pappé, 2006). This foundational trauma of Palestinian collective identity is transmitted intergenerationally through oral histories, literature, and increasingly through digital platforms. By 2025, marking the 77th anniversary commemorated under the hashtag #Nakba77, social media had become the primary arena for commemoration, contestation, and counter-narrative resistance amid continued regional violence.

Digital rights organizations documented intense platform dynamics throughout 2025, including over 15 million posts containing violent content directed at Palestinians (7amleh, 2025), systematic content moderation biases affecting Arabic-language political content (Sada Social, 2025), and algorithmic demotion of pro-Palestinian material despite its numerical dominance (Corsi, 2023). Yet computational linguistics has produced few dedicated resources for studying Nakba discourse. Existing datasets typically focus on acute escalation periods rather than sustained memory work, remain limited to single languages or narrow time windows, or lack the engagement metadata necessary for studying virality and amplification dynamics.

We construct the Nakba Discourse 2025 dataset to address these gaps, contributing five key innovations. First, the dataset provides longitudinal full-year coverage from January through December 2025, enabling analysis of sustained memory maintenance rather than crisis-driven spikes. Second, the bilingual design captures both Arabic (28,031 posts) and English (17,200 posts) discourse, allowing cross-lingual comparison of framing strategies. Third, the intersecting segmentation structure organizes posts simultaneously by language, gender, geography, sentiment, and platform, supporting multidimensional analysis. Fourth, rich metadata including engagement counts, virality indicators, and eighteen rhetorical features enables quantitative analysis of amplification dynamics. Fifth, the ethically gated release follows established frameworks from (Bender and Friedman, 2018) and (Gebru et al., 2021) for responsible data sharing.

The dataset supports multiple NLP tasks including stance classification, virality prediction, cross-lingual transfer learning, code-switching detection, temporal pattern mining, and propaganda detection. We organize our investigation around three

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 33–42

research questions:

RQ1: How do language, gender, geography, platform, and sentiment jointly shape engagement and virality patterns in historical trauma discourse?

RQ2: What distinct rhetorical registers emerge across demographic and linguistic subsets, and how do these reflect different cultural approaches to memory maintenance, advocacy, and identity affirmation?

RQ3: How do cultural and religious calendars structure temporal engagement patterns, and how do counter-narratives transcend linguistic and geographic boundaries?

The paper proceeds as follows: Section 2 reviews related computational work on Israel-Palestine discourse, Arabic NLP resources, and virality research; Section 3 details data collection procedures and corpus composition; Section 4 describes the analysis methodology and processing pipeline; Section 5 presents detailed findings on engagement, rhetorical registers, temporal patterns, and keyword themes; Section 6 discusses theoretical and practical implications; Section 7 addresses limitations and ethical considerations; Section 8 concludes with directions for future work.

2. Related Work

Computational analyses of Israel-Palestine social media discourse have largely focused on acute escalation periods rather than sustained engagement with historical memory. (Shestakov and Zaghouani, 2024) analyzed 370,747 tweets from the 2021 Sheikh Jarrah crisis, highlighting volume asymmetries and weak correlation between hashtag use and virality. Platform-specific studies extend this work: (Liyih et al., 2024) examined YouTube comments during the 2023–2024 Gaza war using neural sentiment models; (Ali, 2025) analyzed Reddit discussions, showing that subreddit norms shape stance expression; (Ng et al., 2024) introduced multimodal analysis through the Love-Hate Dataset; (Antonakaki and Ioannidis, 2025) conducted cross-platform comparisons across Telegram, Reddit, and X; (Nasereddin, 2023) studied differential platform effects on public opinion formation; and (Wang et al., 2025) mapped spatiotemporal sentiment propagation using geotagged tweets. While these studies illuminate platform dynamics and conflict framing, they generally lack longitudinal coverage and engagement metadata centered on Nakba memory discourse.

Recent work has also explored narrative framing and polarization in politically sensitive Arabic discourse through dedicated datasets and shared tasks. In particular, the FIGNEWS shared task (Zaghouani et al., 2024a) introduced a benchmark for analyzing news media narratives, emphasizing the role of framing and bias in shaping public discourse. Complementary efforts in modeling toxic and polarized language, such as multi-label hate speech resources (Zaghouani et al., 2024b), further highlight the interaction between stance, ideology, and linguistic expression in online environments.

Nakba-specific NLP resources have recently expanded through the NakbaNLP initiative (Jarrar et al., 2025). (Ashqar, 2025) demonstrated cultural misalignment in large language model sentiment classification of Nakba oral histories. (AbuHaija et al., 2025) introduced the Nakba Lexicon; (Hamed and Zaidkilani, 2025) developed the NTCC topic classification corpus; (Awad et al., 2025) analyzed narrative cohesion in refugee oral histories; (Nabhani et al., 2025) incorporated argumentation features for propaganda detection; (Garcia-Corral et al., 2025) examined causal attribution in Palestine reporting; (Bilgin Tasdemir and Özateş, 2025) created the NakbaTR NER dataset; and (Mohammed et al., 2025) compared bias detection methods in Israeli-Gaza coverage. These contributions primarily address literary, oral, or news domains and do not provide large-scale social media engagement analysis.

Broader Arabic NLP benchmarks inform our methodological choices. Foundational sentiment datasets include ASTD (Nabil et al., 2015) and ATSAD, while (Chowdhury et al., 2020) and (Mulki et al., 2019) provide offensive language resources. The Curras corpus (Jarrar et al., 2017) supports Palestinian Arabic processing, and (Entman, 1993) framing framework guides rhetorical feature interpretation.

Research on virality and amplification contextualizes engagement dynamics. Emotional arousal predicts sharing behavior (Berger and Milkman, 2012), and algorithmic amplification studies demonstrate systematic visibility concentration in political discourse (Huszár et al., 2022; Ye et al., 2025; Corsi, 2023). (Aboubakr, 2025) conceptualizes digital memory activism as both archival and contestatory, a duality reflected in Nakba discourse online.

No existing resource combines full-year longitudinal coverage, bilingual Arabic-English scope, multi-platform data, demographic segmentation, and engagement metadata for Nakba discourse. The dataset introduced here addresses this gap.

3. Dataset Collection and Composition

3.1. Data Sources and Collection Procedure

Posts were harvested throughout calendar year 2025, from January 1 through December 31, using two commercial social media analytics platforms. Twitter/X data were collected via Meltwater's social media analytics platform, which provides API access to public posts along with demographic annotations derived from profile analysis. Facebook data were collected via CrowdTangle, Meta's research tool for tracking public page content. Both platforms provide engagement metrics including likes, comments, and shares for each post.

Keyword queries were designed to capture posts explicitly engaging with Nakba themes while minimizing false positives from unrelated uses of component terms. English keywords included: "Nakba", "1948 catastrophe", "Right of Return", "Nakba77", "Palestinian displacement 1948", "Nakba Day", and "Nakba commemoration". Arabic keywords included: نكبة (Nakba), نكبة 1948 (Nakba 1948), حق العودة (Right of Return), نكبة فلسطين (Palestine Nakba), ذكرى النكبة (Nakba anniversary), التهجير الفلسطيني (Palestinian displacement), and نكبة 77 (Nakba 77). The complete keyword list with morphological variants is provided in supplementary materials.

3.2. Schema Normalization

Because Meltwater and CrowdTangle use different column naming conventions and data formats, we developed a SchemaAdapter class to normalize exports into a unified schema. The adapter handles three key normalization tasks.

For message content, Meltwater exports use columns named "Opening Text", "Hit Sentence", or "Title" depending on content type, which the adapter maps to a unified "Message" field. CrowdTangle exports use "Message" natively. For timestamps, Meltwater provides separate "Date" and "Time" columns that the adapter combines into a single datetime object, while CrowdTangle provides "Post Created Date" directly. For engagement metrics, Meltwater uses "Replies" and "Reposts" which the adapter maps to "Comments" and "Shares" respectively, while CrowdTangle uses these terms natively.

Total Interactions was computed as the sum of Likes, Comments, and Shares, with Facebook posts additionally including reaction counts (Love, Wow, Haha, Sad, Angry, Care) where available.

The adapter also handles file encoding automatically, detecting UTF-8, UTF-16, and UTF-8 with BOM through byte-order mark inspection, and sniffing delimiters (comma, semicolon, tab, pipe) to accommodate varying export formats.

3.3. Data Cleaning and Deduplication

Language detection combined the langdetect and fastText libraries, requiring greater than 95% confidence for language assignment. Posts failing this threshold were excluded from language-specific sub-corpora but retained in sentiment and platform analyses where language was not a defining criterion.

Deduplication used MinHash locality-sensitive hashing with Levenshtein distance threshold of 5 characters to identify near-duplicate posts while preserving legitimate similar content. This approach removed spam, bot-generated repetitive content, and cross-posted duplicates while retaining organic variations on common themes.

The raw harvest exceeded 142,000 posts. After language filtering, deduplication, and removal of deleted content, 70,312 unique posts remained for analysis.

3.4. Sub-corpora Structure and Overlap

The dataset is organized into twelve intersecting sub-corpora defined by five dimensions: sentiment (Positive, Negative), language (Arabic, English), gender (Male, Female), geography (USA, UK, Israel), and platform (Twitter/X, Facebook). However, not all subsets include segmentation across every dimension. Sentiment labels were provided by Meltwater's automated sentiment classification pipeline. Meltwater's system supports sentence-level sentiment classification across 16 languages, aggregating these into document-level labels of positive, negative, or neutral. The pipeline incorporates deep learning models and accounts for hashtags, emojis, and emoticons as part of its classification input. A continuous feedback loop is maintained to improve model performance over time (**?**). As with any automated pipeline, these labels come with some caveats worth keeping in mind. The deep learning models report sentence-level accuracy of 83% for English, and performance figures for Arabic are not published, meaning classification of Arabic political content that blends grief with resilience or draws on culturally specific affect may not always align with human judgment. These are common constraints of large-scale automated annotation, and findings involving sentiment labels are interpreted accordingly throughout the paper. Geographic labels were inferred from user-declared location fields in account profiles.

Critically, these sub-corpora are not mutually exclusive. A single post may appear in multiple subsets simultaneously. For example, a post by a fe-

male US-based English-speaking author expressing negative sentiment would appear in the Female English, USA, Negative Sentiment, and All English sub-corpora. All statistics should therefore be interpreted as characterizing overlapping populations rather than independent samples, and aggregate totals across sub-corpora must not be summed.

The cross-platform design is intentionally asymmetric: English content derives primarily from Twitter/X, while Facebook content is exclusively Arabic. This asymmetry reflects actual platform usage patterns in our target populations but limits direct bilingual cross-platform comparisons.

4. Methodology

Our analytical framework proceeds in four stages: platform export normalization into a unified schema, engagement metric computation, rhetorical feature extraction at the message level, and bilingual keyword processing. Message-level features include character length, hashtag count, question and exclamation counts, and URL presence.

Platform-specific exports are normalized into a unified schema. Engagement fields (likes, comments, shares, reactions) are harmonized and safely converted to numeric values. Message-level features include character length, hashtag count, question and exclamation counts, and URL presence detected via regex. Message length is additionally grouped into categorical bins (0-100, 100-300, 300-500, 500-1000, 1000+ characters). For Facebook data, engagement rate and basic performance statistics are computed when follower counts are available.

4.1. Engagement and Virality Metrics

The share-to-like ratio captures redistribution behavior relative to passive engagement, calculated as Shares divided by Likes for posts with at least one like, and zero otherwise.

Virality is defined using the 90th percentile of Total Interactions within each sub-corpus. Posts exceeding this threshold are classified as viral:

```
threshold = df['Total
    Interactions'].quantile(0.90)
df['Is_Viral'] = df['Total
    Interactions'] > threshold
```

This relative definition yields approximately 10% viral posts per subset, enabling cross-community comparison while preserving baseline differences in engagement levels. We report both viral percentage and threshold values. A continuous Virality Score is also computed as Total Interactions divided by median Total Interactions.

4.2. Keyword Extraction Pipeline

Keyword extraction applies bilingual preprocessing to Arabic and English text. URLs, email addresses, and @-mentions are removed via regex, and hashtag symbols are stripped while retaining tag text. English text is lowercased. Arabic text is normalized by mapping alef variants (إأآا) to bare alef (ا), converting alef maqsura (ى) to ya (ي), and removing diacritics.

Arabic stemming uses ultra-light stemming that removes only the definite article ال while preserving suffixes. Selected tokens such as الله (Allah) are excluded from stemming. Stopword filtering uses combined Arabic and English lists.

For each retained stem, we compute raw frequency, document frequency, percentage of total tokens, and percentage of posts containing the term. Keywords are ranked by frequency for top-N extraction.

4.3. Temporal Analysis

Temporal analysis aggregates Total Interactions by calendar day using the Post Created Date field (UTC). For each subset, the peak engagement day of week is identified based on total interactions. We note that UTC timestamps may obscure local-time effects in geographically distributed data; timezone-localized analysis is left for future work.

5. Analysis and Findings

5.1. Engagement Concentration and Elite Dynamics

Per-post engagement metrics reveal strong concentration of influence in small subsets. Israel-located accounts achieve 50.62 average likes and 12.90 shares per post despite comprising only 1,413 posts, the smallest geographic subset. UK follows with 49.08 average likes and 17.14 shares per post. These figures exceed USA (25.41 likes), All English (34.20 likes), and All Arabic (15.06 likes).

High-engagement Israeli posts average 489 likes and 136 shares, nearly double UK's figures, indicating dominance by high-influence accounts such as journalists, politicians, academics, and organizations. Engagement concentration is therefore not simply a function of dataset size but of account reach and follower base.

Arabic-language subsets show markedly lower per-post engagement: Arabic Female dataset averages 6.19 likes and 0.94 shares, while Arabic Male averages 18.89 likes and 4.70 shares. Rather than reflecting reduced relevance, this pattern may indicate lower account follower counts or

Table 1: Dataset summary ordered by post volume. Viral percentage uses relative threshold (top 10% within each subset). Facebook Arabic lacks share data due to platform limitations.

Dataset	Posts	Total Likes	Total Shares	Avg Likes	Avg Shares	Viral %
Negative Sentiment	54,424	1,132,196	282,296	20.80	5.19	9.3%
All Arabic (X/FB)	28,031	422,196	86,186	15.06	3.07	9.4%
Arabic Male	17,329	327,310	81,378	18.89	4.70	9.9%
All English (X)	17,200	588,187	165,276	34.20	9.61	9.6%
Male English	13,180	421,984	124,147	32.02	9.42	9.5%
USA	10,906	277,094	63,369	25.41	5.81	9.6%
Facebook Arabic	5,563	63,079	N/A	11.34	N/A	3.7%
Arabic Female	5,139	31,807	4,808	6.19	0.94	9.9%
UK	4,881	239,568	83,680	49.08	17.14	9.4%
Positive Sentiment	4,827	106,390	25,196	22.04	5.22	9.5%
Female English	4,020	166,203	41,129	41.34	10.23	9.7%
Israel	1,413	71,525	18,227	50.62	12.90	10%

platform reach within Arabic-language subsets, resulting in lower absolute engagement figures.

Facebook Arabic operates under a different engagement structure. With zero recorded shares and 11.34 average likes per post, engagement reflects affirmation without redistribution. This likely reflects platform affordances and privacy norms that limit public redistribution on Facebook.

The share-to-like ratio clarifies redistribution dynamics. Mean ratios range from 0.035 for Arabic Female to 0.073 for UK. Critically, the median share-to-like ratio is 0.000 across every Twitter/X dataset, confirming an extreme long-tail structure in which a small minority of posts drives nearly all sharing activity. Amplification is therefore elite-driven across linguistic and geographic communities.

5.2. Virality as Structural Constant

Viral rates cluster tightly between 9.3% and 10.0% across all Twitter/X subsets regardless of language, gender, geography, or sentiment. Despite substantial variation in absolute engagement levels, the proportion of posts exceeding the 90th percentile threshold remains nearly invariant.

This proportional stability suggests that virality operates as a structural property of the platform environment rather than as a community-specific feature. While network size and account influence determine absolute thresholds, the distributional shape of attention appears normalized. Communities differ in magnitude, but not in the statistical structure of amplification.

However, viral thresholds vary significantly. Israel requires 18 interactions to reach viral status, All English and USA require 10, and All Arabic requires 6. Thus proportional virality is constant, but the baseline for success differs across engagement ecologies.

Facebook Arabic is the sole exception, with 3.7% viral posts and a threshold of zero. This confirms that non-share-based architectures fundamentally alter virality mechanics.

Table 2: Temporal patterns and viral thresholds. Peak Int. shows total interactions on the peak day. Threshold indicates interactions needed for viral classification.

Dataset	Peak Day	Peak Int.	Threshold
Facebook Arabic	Thursday	21,264	0
Female English	Thursday	75,446	13
Arabic Female	Thursday	12,255	7
USA	Thursday	94,744	10
All English	Sunday	187,345	10
Male English	Sunday	146,741	9
UK	Sunday	112,231	9
Negative Sent.	Sunday	281,643	8
All Arabic	Monday	102,192	6
Arabic Male	Monday	84,749	8
Positive Sent.	Monday	34,495	6
Israel	Wednesday	40,367	18

5.3. Rhetorical Registers and Communication Strategies

Four rhetorical registers emerge.

Declarative/Witness-bearing. Arabic subsets show 0.00 questions per post, low hashtag use (0.15 to 0.19; 0.02 Facebook), limited URL inclusion (6.3% to 7.9%), and shorter messages (107 to 113 characters). This register asserts rather than interrogates. Positive Sentiment shares this declarative profile with a 0.01 question rate, suggesting that affirmation across languages adopts similar assertive framing.

Advocacy/Evidence-driven. USA and Male English display the highest hashtag rates (0.30 to 0.31), high URL inclusion (13.8% to 14.3%), consistent question use (0.10), and longer messages (126 characters). Posts combine discoverability with documentation, positioning Nakba discourse within global accountability frameworks.

Table 3: Message characteristics across datasets. Hash/Post = hashtags per post; URLs = percentage containing URLs; Quest/Post = questions per post; Excl/Post = exclamations per post; Avg Len = average length in characters.

Dataset	Hash/Post	URLs	Quest/Post	Excl/Post	Avg Len
Positive Sentiment	0.48	12.2%	0.01	0.12	100
USA	0.31	13.8%	0.10	0.06	126
Male English	0.30	14.3%	0.10	0.10	126
All English	0.27	14.1%	0.10	0.09	127
UK	0.23	15.3%	0.09	0.15	131
Arabic Male	0.19	7.9%	0.00	0.09	110
Negative Sentiment	0.17	8.6%	0.03	0.09	116
Female English	0.16	13.3%	0.11	0.08	130
All Arabic	0.15	6.3%	0.00	0.08	110
Arabic Female	0.15	7.3%	0.00	0.09	107
Facebook Arabic	0.02	0%	0.00	0.07	113
Israel	0.02	12%	0.11	0.16	122

Emotive/Dialogic. UK and Israel show highest exclamation rates (0.15 and 0.16), high question rates (0.09 to 0.11), and strong URL inclusion (12% to 15.3%). Israel combines high affect with minimal hashtags (0.02), indicating reliance on follower networks rather than algorithmic tagging.

Networked/Commemorative. Positive Sentiment authors use 0.48 hashtags per post and produce the shortest messages (100 characters). Female English shows strong quote-tweet signal indicated by "qt", reflecting amplification-oriented participation rather than standalone authorship.

5.4. Temporal Patterns

Three peak structures appear. Thursday peaks in Facebook Arabic (21,264), Arabic Female (12,255), Female English (75,446), and USA (94,744). Sunday peaks in All English (187,345), Male English (146,741), UK (112,231), and Negative Sentiment (281,643). Monday peaks for Arabic Male and All Arabic; Israel peaks Wednesday.

These temporal regularities indicate that digital Nakba discourse aligns with broader social rhythms. Thursday concentration among Arabic and female subsets corresponds to the eve of Jumu'ah, when communal and religious engagement intensifies. Sunday concentration among English and male subsets reflects Western media cycles. Israel's Wednesday peak mirrors the Israeli work week. Online amplification therefore follows offline temporal structures rather than operating autonomously.

5.5. Keyword Themes and Counter-Narrative Dynamics

"Nakba" appears in 52% to 66% of posts, highest in Arabic (65% to 66%) and Facebook Arabic (66%), confirming strong thematic cohesion.

"Genocide" appears in 6.2% to 8.5% of English posts, highest in USA (8.5%), indicating legal-accountability framing. "Gaza" appears in Arabic Male (7.25%) and UK (7.68%). "War" peaks in Israel at 13.4%.

"الله" appears in Arabic Female (6.4%), Arabic Male (7.3%), Facebook Arabic (15.5%), Negative Sentiment (5.65%), and Positive Sentiment (8.6%), and is absent from English top keywords, marking a structural cultural-linguistic divide.

"History" appears only in UK (7.2%). "Jews" appears only in Israel (11.6%), alongside "Arabs" and "Arab" (10.2% to 12.7%), reflecting a distinct ethno-demographic framing.

Figure 1 visually reinforces these divergences.

5.6. Gender Patterns

Male-authored datasets are approximately three times larger. Viral rates show parity: Arabic Female and Arabic Male both 9.9% (thresholds 7 and 8); Female English 9.7%; Male English 9.5%. Women achieve proportional virality equal to men despite lower volume.

Arabic Female posts are most declarative (0.00 questions; 7.3% URLs). Male English shows highest hashtag use (0.30). Female English achieves higher shares-per-post (10.23 vs 9.42), indicating strong redistribution despite lower volume. Female subsets peak Thursday; Arabic Male Monday; Male English Sunday.

5.7. Sentiment Distribution

Negative Sentiment contains 54,424 posts; Positive Sentiment 4,827 (11:1). Both show viral rates between 9.3% and 9.5%.

Positive posts are shorter (100 characters) and use 0.48 hashtags, suggesting condensed solidarity signaling. Negative posts are longer (116 char-

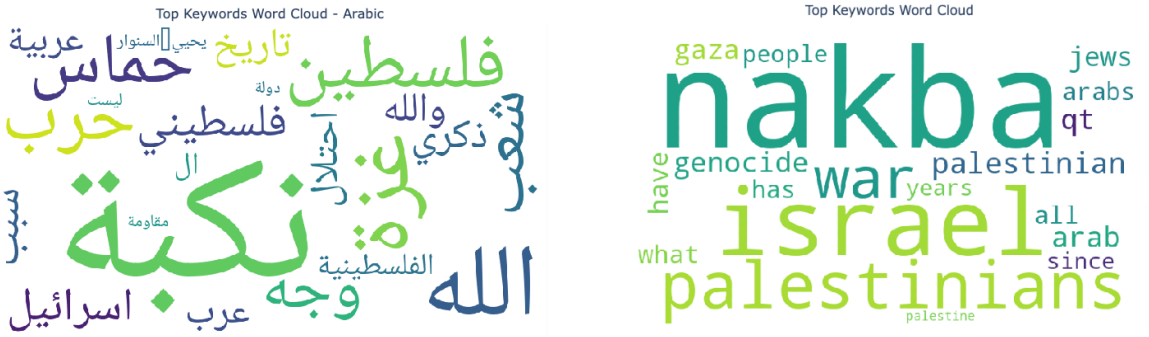

Figure 1: Bilingual keyword word clouds for Arabic (left) and English (right) Nakba discourse. Word size proportional to corpus-level frequency after normalization, light stemming, and stopword removal.

Table 4: Language baseline comparison showing systematic divergence across metrics except viral rate.

Metric	All Arabic	All English
Total Posts	28,031	17,200
Avg Likes/Post	15.06	34.20
Avg Shares/Post	3.07	9.61
Share-to-Like Ratio	0.037	0.052
Viral Rate	9.4%	9.6%
Viral Threshold	6	10
Peak Day	Monday	Sunday
Avg Message Length	110 chars	127 chars
Questions/Post	0.00	0.10
Hashtags/Post	0.15	0.27
URL Inclusion	6.3%	14.1%
Nakba-term %	65.9%	52.6%
Viral Posts (n)	2,647	1,648

acters) and use 0.17 hashtags, reflecting extended articulation of grievances.

5.8. Language Baseline Comparison

English content is longer, more URL-rich, more hashtag-heavy, more questioning, and more shareable. Arabic shows higher Nakba-term penetration and more declarative style. Both achieve near-identical viral rates, indicating shared structural amplification conditions despite divergent rhetorical norms.

"حماس" appears at 3.9% in All Arabic; "qt" appears at 8.1% in All English, reflecting networked repost culture in English-language Twitter.

6. Discussion

6.1. Theoretical Implications

The findings advance understanding of digital political memory in several ways. The near-universal viral rate of approximately 10% across Twitter/X datasets, despite dramatic variation in absolute engagement, community size, and rhetorical style, suggests that platform mechanics impose structural regularities on content spread independent of content characteristics or community identity. This "shape invariance" may reflect algorithmic amplification operating uniformly while absolute reach varies with account size and follower base.

The four-register rhetorical typology maps onto distinct cultural and strategic orientations toward memory work. Arabic discourse performs what (Halbwachs, 1992) termed memory maintenance through declarative witness-bearing that affirms identity without seeking external validation. English discourse pursues global legitimacy through advocacy and evidence, characteristic of diaspora communities seeking international recognition. UK and Israel discourse engages in confrontational dialogue reflecting colonial and national historical positions. Positive and female discourse builds solidarity through networked commemoration that amplifies community voices.

Moreover, Arabic speakers engage with and rebut English-language denial not through translation but through bilingual code-switching that incorporates the denial phrase itself. This finding supports (Aboubakr, 2025) theorization of digital memory activism as simultaneously archival and contestatory.

The cultural calendar findings reveal that digital Nakba commemoration is embedded in broader rhythms of religious and social life. Thursday/Jumu'ah peaks for Arabic and female subsets, Sunday peaks for English and male subsets, and the Israeli Wednesday peak all demonstrate that online discourse follows offline temporal structures rather than operating as a separate digital sphere.

6.2. Practical Implications

For NLP researchers, the dataset enables several benchmark tasks. Stance detection can leverage demographic segmentation to examine how language, gender, and geography predict framing.

Virality prediction can incorporate rhetorical features as predictors alongside engagement history. Cross-lingual transfer learning can exploit the parallel Arabic-English coverage to develop models that generalize across languages.

For platform governance, the finding that elite accounts in small subsets (Israel, UK) drive disproportionate engagement aligns with amplification audit findings (Ye et al., 2025; Corsi, 2023) and suggests that visibility inequality may be particularly acute for politically contested topics. Combined with documented content moderation disparities affecting Arabic and pro-Palestinian content (7amleh, 2025), these patterns warrant attention to equity in algorithmic amplification.

For practitioners in digital humanities and memory studies, the temporal calendar findings suggest that commemoration campaigns should align with community rhythms. Thursday posting may optimize reach for Arabic and female audiences, while Sunday posting may optimize for English and male audiences.

7. Dataset Availability

Access to the dataset requires completion of a request form[1]. Access will be granted to researchers affiliated with recognized academic or research institutions upon submission of a request outlining the intended use of the data. To ensure responsible use, applicants will be required to agree to terms that prohibit redistribution, commercial exploitation, or attempts to re-identify individual users. The dataset will be distributed in a format consistent with platform policies and applicable data protection regulations. Information about access procedures and request forms will be made available in the camera ready version.

8. Conclusion

The Nakba Discourse 2025 dataset provides an unprecedented resource for studying digital political memory at the intersection of NLP, computational social science, and memory studies. Our contributions include the first full-year, bilingual, multi-platform corpus of Nakba discourse with engagement metadata; a four-type rhetorical typology linking linguistic features to distinct memory practices; documentation of structural viral regularities alongside dramatic engagement inequality; evidence for cultural calendar effects structuring temporal patterns; and analysis of pan-linguistic counter-narrative dynamics.

Future work will extend this foundation in several directions: fine-grained emotion labels beyond binary sentiment to capture the complex affective landscape of trauma discourse; multimodal extensions incorporating images and videos; Hebrew-language data enabling trilingual analysis of Israeli discourse; timezone-localized temporal analysis; and longitudinal tracking as the corpus extends into future years.

We invite the research community to build upon this resource for more equitable, culturally-informed multilingual NLP in conflict and memory contexts.

9. Limitations

Several limitations constrain interpretation of our findings. Gender and sentiment classifications were provided by Meltwater's automated pipelines rather than human annotation. Binary gender classification does not capture non-binary identities and may misgender individuals based on name-based inference. Sentiment classification performance on Arabic political content, particularly expressions mixing grief with resilience, has not been independently validated on our corpus.

Geographic labels derive from user-declared location fields, which may reflect ancestral, aspirational, or privacy-protective declarations rather than actual residence. The coverage and accuracy of geographic inference have not been systematically validated. A further limitation is the geographic segmentation of Arabic-language posts, which are not subdivided by Arab country of origin. The current geographic sub-corpora — USA, UK, and Israel — reflect the availability of English-language geographic signals and do not capture the diversity of Arabic-speaking communities across countries such as Jordan, Lebanon, Palestine, Egypt, and the Gulf states. This limits the extent to which regional variation in Arabic Nakba discourse can be examined. Future work should incorporate country-level segmentation within Arabic sub-corpora, enabling comparative analysis of how Nakba memory is constructed and contested across different Arab national contexts, each carrying distinct historical relationships to 1948.

Platform moderation may systematically undersample certain voices. Digital rights organizations have documented bias against Arabic and pro-Palestinian content in content moderation systems. Our findings therefore characterize visible discourse rather than the full population of Nakba-related posting.

The corpus excludes Hebrew, limiting analysis of Israeli counter-narrative dynamics from Hebrew-language sources. The cross-platform design is asymmetric, with English primarily from Twitter/X and Facebook exclusively Arabic, constrain-

[1] https://forms.gle/W7xpLt7io326bR3A6

ing bilingual cross-platform comparisons.

The relative viral threshold (top 10% within each subset) yields approximately 10% viral classification by construction, limiting cross-subset comparison of absolute viral reach. UTC timestamps may introduce artifacts in temporal analysis for a geographically distributed corpus.

10. Ethical Considerations

Our ethical approach follows frameworks established by Bender & Friedman (2018) for data statements and Gebru et al. (2021) for dataset documentation. Usernames are redacted from released data. Geographic information is coarsened to country level. Profile images are excluded.

For Twitter/X content, we release tweet IDs only in compliance with platform terms of service. Researchers may rehydrate full text via API access subject to their own compliance obligations. For Facebook content, we release anonymized feature vectors rather than raw text, with full text available only through gated access requiring researcher verification.

Dataset access is restricted to verified researchers through an application form. Applicants must describe research purpose, institutional affiliation, and data handling protocols. Commercial use is prohibited.

We acknowledge dual-use risks including potential misuse for targeted harassment, surveillance, or propaganda. Access restrictions and anonymization mitigate but do not eliminate these risks. We encourage responsible use and will monitor for misuse.

The study received IRB-equivalent ethics review and exemption determination based on analysis of publicly available data with anonymization protections.

Acknowledgments

This work was made possible by the National Priorities Research Program (NPRP) grant NPRP14C-0916-210015 from the Qatar National Research Fund (QNRF), a member of the Qatar Research, Development and Innovation Council (QRDI)

References

7amleh. (2025). Racism and Incitement Index 2024. 7amleh – The Arab Center for the Advancement of Social Media.

Aboubakr, F. (2025). Archivalism and memory activism: The Nakba (1948) and the Gaza War (2023). *Memory Studies*, 18(2):439–455.

Ali, H. J., Abrar, A., Hossain, S. M. H., and Mridha, M. F. (2025). Social media polarization during conflict: Insights from an ideological stance dataset on Israel-Palestine Reddit comments. *arXiv:2502.00414*.

Antonakaki, D. and Ioannidis, S. (2025). Cross-platform digital discourse analysis of the Israel-Hamas conflict. *arXiv:2601.02367*.

Ashqar, H. I. (2025). Sentiment Analysis of Nakba Oral Histories: A Critical Study of Large Language Models. In *Proceedings of NakbaNLP 2025*, pages 30–36, Abu Dhabi. Association for Computational Linguistics.

AbuHaija, I., Al Mandhari, S., El-Haj, M., Sibony, J., and Rayson, P. (2025). The Nakba Lexicon: Building a Comprehensive Dataset from Palestinian Literature. In *Proceedings of NakbaNLP 2025*, pages 37–47, Abu Dhabi. Association for Computational Linguistics.

Awad, G. A., Rayan, T. N., Dunagan, L., and Gamba, D. (2025). Collective Memory and Narrative Cohesion: A Computational Study of Palestinian Refugee Oral Histories in Lebanon. In *Proceedings of NakbaNLP 2025*, pages 83–102, Abu Dhabi. Association for Computational Linguistics.

Bender, E. M. and Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Berger, J. and Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research*, 49(2):192–205.

Bilgin Tasdemir, E. F. and Özateş, Ş. B. (2025). NakbaTR: A Turkish NER Dataset for Nakba Narratives. In *Proceedings of NakbaNLP 2025*, pages 122–126, Abu Dhabi. Association for Computational Linguistics.

Chowdhury, S. A., et al. (2020). A multi-platform Arabic news comment dataset for offensive language detection. In *Proceedings of LREC 2020*.

Corsi, G. (2023). Visibility and amplification on X after policy changes: An audit study. *Proceedings of ICWSM 2023*.

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.

Garcia-Corral, P., et al. (2025). The missing cause: Causal attribution analysis in Palestine reporting. In *Proceedings of NakbaNLP 2025*.

Gebru, T., et al. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Halbwachs, M. (1992). *On Collective Memory*. University of Chicago Press.

Hamed, O. and Zaidkilani, N. (2025). Arabic Topic Classification Corpus of the Nakba Short Stories. In *Proceedings of NakbaNLP 2025*, pages 48–55, Abu Dhabi. Association for Computational Linguistics.

Huszár, F., et al. (2022). Algorithmic amplification of politics on Twitter. *Proceedings of the National Academy of Sciences*, 119(1).

Jarrar, M., Habash, N., El-Haj, M., Haddad, A. H., Jallad, Z., Mansour, C., Allan, D., Rayson, P., Hammouda, T., and Malaysha, S. (eds.) (2025). *Proceedings of the First International Workshop on Nakba Narratives as Language Resources (NakbaNLP 2025)*, Abu Dhabi. Association for Computational Linguistics.

Jarrar, M., Habash, N., Alrimawi, F., Akra, D., and Zalmout, N. (2017). Curras: An annotated corpus for the Palestinian Arabic dialect. *Language Resources and Evaluation*, 51(3):745–775.

Khalidi, R. (2020). *The Hundred Years' War on Palestine*. Metropolitan Books.

Liyih, A., Anagaw, S., Yibeyin, M., and Tehone, Y. (2024). Sentiment analysis of the Hamas-Israel war on YouTube comments using deep learning. *Scientific Reports*, 14:13647.

Mohammed, M. Y., et al. (2025). Bias detection in media coverage of the Israeli-Gaza conflict: Traditional vs. transformer models. In *Proceedings of NakbaNLP 2025*.

Mulki, H., et al. (2019). L-HSAB: A Levantine Twitter dataset for hate speech and abusive language. In *Proceedings of RANLP 2019*.

Nabhani, S., Borg, C., Micallef, K., and Al-Khatib, K. (2025). Integrating Argumentation Features for Enhanced Propaganda Detection in Arabic Narratives on the Israeli War on Gaza. In *Proceedings of NakbaNLP 2025*, pages 127–149, Abu Dhabi. Association for Computational Linguistics.

Nabil, M., et al. (2015). ASTD: Arabic sentiment tweets dataset. In *Proceedings of EMNLP 2015*.

Nasereddin, S. (2023). Impact of social media platforms on international public opinion during the 2023 Israel-Gaza war. *Journal of International Communication*.

Ng, L. H. X., Lim, A. X. W., and Lee, R. K.-W. (2024). Love-Hate Dataset: A Multi-Modal Multi-Platform Dataset Depicting Emotions in the 2023 Israel-Hamas War. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion)*, pages 1807–1815. ACM.

Pappé, I. (2006). *The Ethnic Cleansing of Palestine*. Oneworld Publications.

Sa'di, A. H. and Abu-Lughod, L. (eds.) (2007). *Nakba: Palestine, 1948, and the Claims of Memory*. Columbia University Press.

Sada Social. (2025). The Digital Rights Index 2024. Sada Social Center for Palestinian Digital Rights.

Shestakov, A. and Zaghouani, W. (2024). Analyzing Conflict Through Data: A Dataset on the Digital Framing of Sheikh Jarrah Evictions. In *Proceedings of the Second Workshop on Natural Language Processing for Political Sciences @ LREC-COLING 2024*, pages 55–67, Torino, Italia. ELRA and ICCL.

Wang, D., et al. (2025). Spatiotemporal sentiment propagation analysis of the Israeli-Palestinian conflict. *Humanities & Social Sciences Communications*.

Ye, J., et al. (2025). Algorithmic amplification audits on X.

Zaghouani, W., Jarrar, M., Habash, N., Bouamor, H., Zitouni, I., Diab, M., El-Beltagy, S., & AbuOdeh, M. (2024). The FIGNEWS shared task on news media narratives.

Zaghouani, W., Mubarak, H., & Biswas, M. R. (2024). So hateful! Building a multi-label hate speech annotated Arabic dataset.