ArabMisinfo: A Decade-Long Arabic Facebook Corpus for Misinformation Discourse and Engagement Analysis

Wajdi Zaghouani[1], Shimaa Ibrahim[1], Mabrouka Bessghaier[1], Md. Rafiul Biswas[2]
[1]Northwestern University in Qatar, Qatar
[2]Hamad Bin Khalifa University, Qatar

{wajdi.zaghouani,mabrouka.bessghaier,shimaa.ibrahim}@northwestern.edu
mbiswas@hbku.edu.qa

Abstract

We present ArabMisinfo, a corpus of 274,837 public Facebook Page posts in Arabic spanning 2014–2024 and focused on news and public affairs. Each record includes page-level metadata, UTC timestamps, link domains, and platform-native engagement signals (likes, comments, shares, and the seven post-2016 reaction types). We accompany the corpus with reproducible diagnostics for engagement and polarization, the Polarization Index (PI), Share-to-Like Anomaly (SLA), Risk Triage Index (RTI), and Engagement Pattern Residuals (EPR). To address the absence of content-level labels, we contribute a discourse role annotation of 10,000 posts (3.6% of the corpus) using Fanar-C-2-27B, classifying each post into one of six roles. The annotation reveals that Arabic Facebook misinformation discourse is dominated by counter-information activity: 30.6% of posts are Rumor Denial or Official Correction, outpacing Unverified Claim posts (12.3%) by a 2.5:1 ratio, while 19.3% represent incidental keyword matches, directly quantifying the sampling noise introduced by keyword-based collection. To comply with platform terms and copyright, we publicly release Post IDs, Page IDs, and full metadata and derived indices; message text is available only under a research data-use agreement (DUA). ArabMisinfo enables engagement-aware analyses of Arabic misinformation ecosystems and, to our knowledge, provides the first discourse-role-annotated Arabic Facebook corpus at this scale.

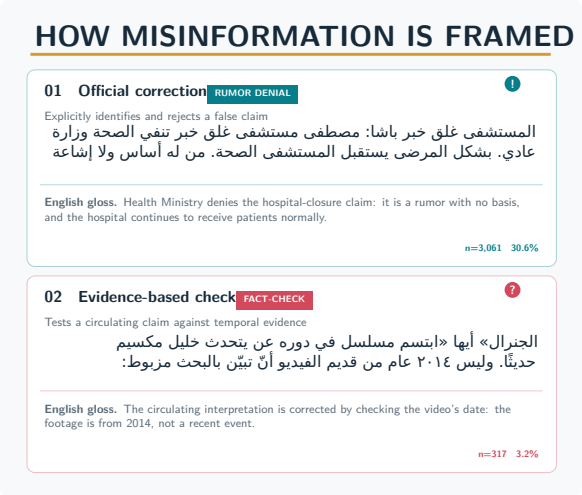


Figure 1: Two representative examples illustrating rumor denial and evidence-based verification in the ArabMisinfo corpus. English glosses are provided for accessibility.

1 Introduction

Online misinformation can distort public talk and public choices, and false claims often spread fast and far on social networks (Dizikes, 2018; Kleinman, 2018). Social media is a main news source across Arabic speaking regions, and Facebook is widely used for public affairs content (Rananga et al., 2024). At the same time, Arabic language brings practical challenges for computation because of rich morphology, dialectal variation,and context specific expressions (Alotaibi and Al-Dossari, 2024).

Most Arabic misinformation resources focus on Twitter. Examples include ArCOV19 Rumors (Haouari et al., 2021b), multi label situational sets (Obeidat et al., 2022), and VERA ARAB (Mostafa and Almogren, 2024). Facebook, despite its central role in Arabic news use (Boulianne and Hoffmann, 2024), remains under resourced. Access also became harder after the retirement of CrowdTangle in 2024, which disrupted common research workflows (Ortutay, 2024; Johns et al., 2024).

This paper introduces **ArabMisinfo**, a Facebook corpus in Arabic with **274,837** public posts from news and public interest Pages collected from **2014** to **2024**. The files keep post text, link context, UTC time stamps, page fields, and detailed engagement for the seven reactions, plus likes, comments, and shares. All descriptive results in this paper are com-

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 3809–3820
November 6-10, 2026 ©2026 Association for Computational Linguistics

puted on the same corpus that we publicly release under a data-use agreement. The corpus provides descriptive and diagnostic analysis of misinformation content on Facebook using various metadata. However, it is not intended for benchmarking for truth labeling or a release of media files.

Our contributions are the following. First, we provide a large Facebook corpus in Arabic with content and page fields in one place. Second, we document a clear schema with coverage notes so that others can understand what is present and what is missing. Third, we present four diagnostic views that rely only on platform signals: a joint view of polarization and share-to-like anomaly, a temporal map of posting and engagement, a rank-based risk triage index, and a robust engagement outlier rule. Fourth, we contribute a discourse role annotation of 10,000 posts using Fanar-C-2-27B (TEAM et al., 2026), classifying each post into one of six roles that characterize how Arabic Facebook pages participate in misinformation discourse, from active rumor denial to fact-checking to incidental keyword use. Together, these views and annotations give transparent baselines for future work on Arabic news and public affairs on Facebook and help target posts that merit closer review without requiring external truth labels.

We review related work (Section 2), detail dataset construction (Section 3), define the diagnostic indices (Section 4), report the core findings (Section 5), discuss implications (Section 6), outline social impact and applications (Section 7), and conclude (Section 8), followed by limitations (Section 9) and the ethics statement (Section 10).

2 Related Work

Most Arabic misinformation datasets are built from Twitter and are text only. Facebook resources in Arabic are rare and small. Table 1 lists representative datasets and shows this pattern. AFND (Khalil et al., 2022) covers news articles, ArCOV19 (Haouari et al., 2021a) and ArCOV19 Rumors (Haouari et al., 2021b) focus on tweets, and the only Facebook item is a small comments set. Work on propaganda and memes adds useful multimodal cases, and recent Arabic deepfake sets exist off platform, but these do not include Facebook post content or page fields. None of the listed resources provide Facebook posts with content, page fields, time stamps, and the full set of reactions. Our dataset fills this gap and enables reaction based diagnostics.

Behavioral diagnostics and platform signals. Our analysis uses platform native signals such as reactions, shares, comments, and time of posting. Prior work shows that items can spread even when endorsement is weak, which motivates the share to like anomaly view (Vosoughi et al., 2018; Tacchini et al., 2017). Reaction categories can be read as coarse affective cues in audience response (Krug, 2016; Lunden, 2020). Heavy tailed popularity calls for robust summaries and outlier rules (Clauset et al., 2007; Hoaglin, 2013). Studies of automated and coordinated activity support the use of temporal heatmaps to spot regular and clustered patterns (Ferrara et al., 2016; Varol et al., 2017; Pacheco et al., 2020). Together these works frame the four descriptive diagnostics we report for Arabic Facebook posts.

3 Dataset Construction

ArabMisinfo contains public posts from Arabic Facebook Pages that publish news and public affairs. The dataset preserves post content and page fields so that work on misinformation and audience behavior can use platform signals directly.

3.1 Source and Coverage

Public posts and engagement counts were collected with the research access tool for public Facebook content from **8 June 2014** to **6 June 2024**. Each record keeps the original timestamp and a UTC copy. We also provide simple calendar fields such as year, month, weekday, and hour to support time based analysis.

3.2 Page Selection and Inclusion

Pages were included when they were public during collection, posted mainly in Arabic, and focused on news or public interest topics such as politics, health, economics, science, and public services. Pages that were non Arabic, purely promotional, or inactive were excluded. The seed list used verified outlets, institutions, public figures, and civil society organizations, then expanded by link tracing to sources that were frequently shared. The final corpus spans news outlets, institutions, non governmental organizations, and public figures.

3.3 Collection and Record Unit

For each included Page we retrieved all available public posts and page fields. Each post is stored

Work / Dataset	Modality	Platform	Size	Focus	Main gap
AFND (Khalil et al., 2022)	Text	News	606K articles	Article credibility labels.	Not social media. No Facebook post fields.
ArCOV19 Rumors (Haouari et al., 2021b)	Text	Twitter	9.4K tweets, 138 claims	Rumor and veracity linked to claims.	No Facebook. Text only.
ArCOV19 (Haouari et al., 2021a)	Text	Twitter	2.7M tweets	Propagation and diffusion.	No Facebook. Text only.
AraCOVID19 MFH (Hadj Ameur and Aliane, 2021)	Text	Twitter	10.8K tweets	Misinformation and hate during COVID19.	No Facebook. Domain specific.
Obeidat et al. (Obeidat et al., 2022)	Text	Twitter	∼6.7K tweets	Fine grained misinformation labels.	No Facebook. Text only.
VERA ARAB (Mostafa and Almogren, 2024)	Text	Twitter	∼20K tweets	Veracity linked to fact checked claims.	No Facebook. Text only.
Laabar and Zaghouani (Laabar and Zaghouani, 2024)	Text	**Facebook**	7,535 comments (5,000 labeled)	Stance, sentiment, emotion, abuse in Tunisian comments.	Not misinformation. Comments only. Narrow topics.
ArMeme (Alam et al., 2024)	Image and text	Various	∼6K memes	Propaganda in memes with OCR.	Not Facebook focused. No page fields.
ArAIEval 2024 (Hasanain et al., 2024)	Text and image	Tweets, news, memes	Public tasks	Propaganda spans and meme labels.	Limited Facebook. No post level fields.
ArEnAV (Kuckreja et al., 2025)	Video and audio	Off platform	387K videos, 765h	Audio and video deepfake detection.	Not Facebook. Off platform.
Arabic Audio Deepfake (DeepFake Audio Rangers, 2024)	Audio	Off platform	≈19.6K clips	Arabic audio deepfake detection.	Not Facebook. Audio only.

Table 1: Key Arabic resources by modality, platform, size, focus, and gap. None provide Facebook post content with page fields, time stamps, and full reactions.

as one record. Reshares and crossposts are kept when they meet time and page criteria and are marked with a type flag. Private profiles and closed groups are not included. Collection followed the platform's research terms in force at the time.

3.4 Schema and Fields

The corpus consists of 274,837 posts along with page metadata. Table 2 defines the primary field groups. The content of the post include both the textual message and any supporting graphic text, if available. Page fields include name, identifier, category, and administrative country. Links include the original link and the final resolved link with a parsed domain. Engagement covers likes, comments, shares, and the seven reaction types such as Like, Love, Wow, Haha, Sad, Angry and Care. Derived fields include hour and weekday, text length, word count, and simple content indicators for photo, video, link, and status.

3.5 Text and Language Handling

We preserve Arabic script as provided, including diacritics and emojis. Control characters are removed and encoding errors corrected. We do not apply language normalization or stemming. Page language fields are kept when available. Content can include Modern Standard Arabic and regional variants. OCR text from images is included when present at the source.We do not redistribute media files; only their metadata and any available source OCR text are included.

3.6 Quality Assurance

Consistency checks verified time range bounds, reaction sums, alignment of page name and category, and link resolution to final domains. Exact duplicate post identifiers were removed. Posts that disappeared during collection are kept with empty text and a status flag. Missing values appear as empty cells; we do not impute them.

3.7 Public Release & Licensing

To comply with platform ToS and copyright, we adopt a hybrid release model. We make publicly available a frozen snapshot containing *Post IDs, Page IDs, page-level metadata, timestamps, link domains, engagement counts (likes, comments, shares, seven reactions), and derived indices (PI, SLA,*

Table 2: Schema overview for ArabMisinfo. Groups show representative columns.

Group	Representative fields (examples)
Post level	`Facebook Id`, `Post Created` (original and UTC), `Type` (Photo, Status, Link, Native Video, YouTube, Live), `Message`, `Image Text`, `Description`
Page level	`Page Name`, `Page Id`, `Page Category`, `Page Admin Top Country`, `Language`
Links and domains	`Link`, `Final Link` (after redirect), `Link Text`, domain parsed from the final link
Engagement	`Likes`, `Comments`, `Shares`; reactions: `Like`, `Love`, `Wow`, `Haha`, `Sad`, `Angry`, `Care`; `Total Interactions`
Derived	time: `post_hour`, `post_weekday`; length: `text_length`, `word_count`; content flags: `is_photo`, `is_video`, `is_link`, `is_status`

RTI, EPR). We do not redistribute media files (images/videos). The *message text* is accessible only under a research data-use agreement (DUA) that prohibits re-publication of verbatim content and mandates takedown on request. Metadata are licensed under an open data license code is released under a permissive license (e.g., MIT). We did not provide the URL during the anonymous review. The URL will be included in the camera ready version.

4 Methods: Platform-Native Diagnostic Indices

Our analysis uses four simple indices that can be computed from native Facebook fields without truth labels. We group reactions into a positive set (Like, Love, Haha, Wow, Care) and a negative set (Sad, Angry), following the public reaction taxonomy introduced in 2016, with Care added in 2020.[1]

4.1 Robustness of Diagnostics

We assess the stability of all indices under alternative assumptions.

Reaction semantics. We re-estimate metrics after (a) treating Haha as neutral, (b) excluding Haha, and (c) collapsing Angry+Sad into a single negative-affect channel; headline conclusions remain unchanged.

Weights and thresholds. For RTI , we test 0.6/0.4 and 0.5/0.5 engagement/volatility weights; for EPR, we vary the outlier cut ($|M| > 2.5, 3.0, 3.5$). Patterns and tail-mass rankings are stable.

Temporal strata. We compute ranks monthly vs. quarterly and observe consistent relative ordering.

Sparsity filters. Excluding pages with fewer than $n \in \{20, 50\}$ posts/year does not materially affect distributional summaries.

[1]Facebook expanded Like into Reactions in February 2016; Care was introduced during April 2020. See Krug (2016); Lunden (2020).

Uncertainty. We report bootstrap confidence intervals for summary stats (e.g., median SLA) using 1,000 resamples by page and year. Seeds and randomization settings are fixed for exact reproducibility.

4.2 Polarization Index (PI)

Let R^+ be the count of positive reactions and R^- be the count of negative reactions for a post. Let $S = R^+ + R^- + \epsilon$ with a small ϵ to avoid division by zero. Define the positive share $p^+ = R^+/S$ and the negative share $p^- = R^-/S$. The Polarization Index is

$$\mathrm{PI} = \left| p^+ - p^- \right| \in [0, 1].$$

PI increases when audience response concentrates on either approval or disapproval and is invariant to the total number of reactions. Grouping reactions into positive and negative sets follows prior use of reaction categories as affective signals on Facebook(Phillips, 2024).

4.3 Share to Like Anomaly (SLA)

Shares spread content beyond the origin network while Likes function as a light endorsement signal. Prior work shows that sharing and endorsement can diverge in the spread of questionable content. False or low-quality items can spread faster and farther even without strong endorsement signals(Vosoughi et al., 2018). Likes and other simple feedback have been used to separate hoaxes from mainstream content in Facebook data(Tacchini et al., 2017). We measure diffusion–endorsement imbalance with

$$\mathrm{SLA} = \log_{10}\left(\frac{\mathrm{Shares} + 1}{\mathrm{Likes} + 1}\right).$$

$\mathrm{SLA} > 0$ indicates the post is shared more than it is liked.

4.4 Discussion Intensity and Risk Triage Index (RTI)

Discussion intensity is defined as comments per one hundred reactions:

$$\mathrm{C100} = 100 \times \frac{\text{Comments}}{\sum \text{Reactions} + \epsilon}.$$

To triage posts that merit review, we combine a rank of negativity and a rank of discussion intensity. Within a fixed temporal stratum (for example, within month), compute percentile ranks r_{neg} for the negative reaction share p^- and r_{C100} for C100. The Risk Triage Index is

$$\mathrm{RTI} = 0.6\, r_{\text{neg}} \; + \; 0.4\, r_{\text{C100}},$$

which yields a unit interval score. Rank-based combination reduces the effect of heavy tails and unit scaling and follows standard practice for robust transformation prior to aggregation(Conover and Iman, 1981).

4.5 Engagement Anomaly Rule (EPR)

Per-post engagement is heavy-tailed in social platforms. We apply a robust outlier rule on the log-transformed interaction count. Let $X = \log_{10}(\text{Total Interactions} + 1)$. Compute the median $\tilde{X}$ and the median absolute deviation $\mathrm{MAD} = \mathrm{median}(|X - \tilde{X}|)$. Define the modified z score

$$M = 0.6745 \times \frac{X - \tilde{X}}{\mathrm{MAD} + \epsilon}.$$

We flag an engagement anomaly when $|M| > 3.5$. This threshold and the use of MAD follow standard robust outlier detection guidance (Hoaglin, 2013) and the NIST Engineering Statistics Handbook for the modified z score with a 3.5 cut point (NIST/SEMATECH, 2012). We use the log transform because engagement follows a heavy-tailed form where log scaling stabilizes variance for summary and visualization(Clauset et al., 2007).

5 Data Analysis

This section focuses on four diagnostic views that describe how content spreads in the corpus. All numbers reported here are taken from the same run that produced the figures.

5.1 Polarization and share to like anomaly

Figure 2 plots the two signals defined in the Methods section. The median values in this corpus are

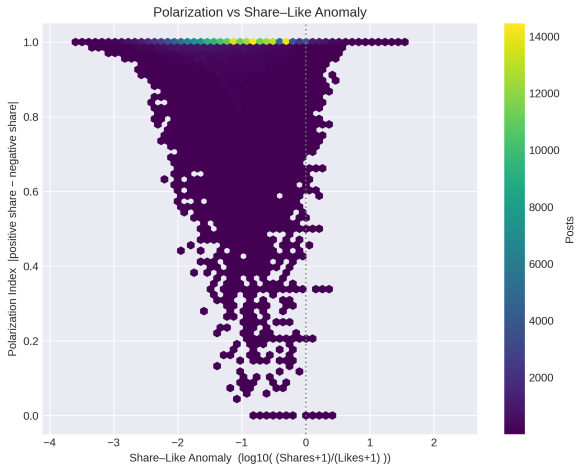


Figure 2: Polarization versus share to like anomaly. The vertical line marks $\mathrm{SLA} = 0$.

PI = 1.00 and **SLA =** -1.00. The first shows that most posts receive unanimous or near unanimous reactions. The second means a typical post has about one share for every ten likes. Points to the right of the reference line have more shares than likes. When that pattern appears together with high polarization (for example $\mathrm{PI} > 0.5$), the item spreads without clear endorsement and should be reviewed with priority. The scatter also forms a cone shape: extreme share to like values occur most often when reactions are unanimous, while mixed reactions pull the anomaly toward the center.

5.2 Temporal Posting and Engagement

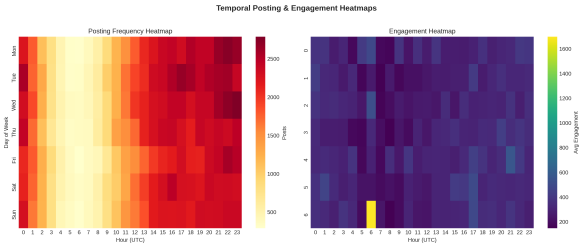


Figure 3: **Temporal patterns.** Left: posting frequency by day of week and hour in UTC. Right: mean engagement on the same grid. Color bars show post counts on the left and average engagement on the right.

Figure 3 summarizes when Pages post and when audiences respond. Posting volume is highest at **22:00**, **23:00**, and **21:00** UTC. The engagement map is not a mirror of the volume map. Cells that are bright on the right but not on the left indicate hours with higher than expected attention and can help flag coordination windows.

5.3 Risk Triage Index

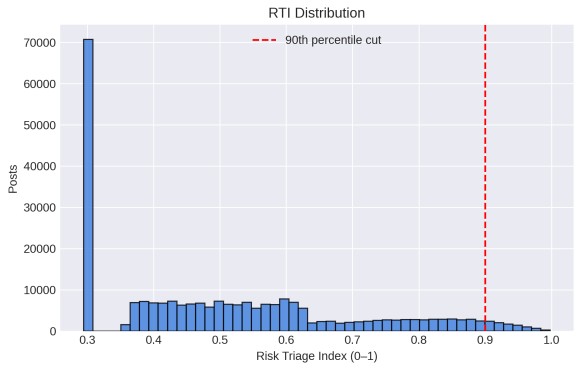


Figure 4: **Risk Triage Index (RTI).** RTI $= 0.6\,\mathrm{rank}(\text{negative reaction share}) + 0.4\,\mathrm{rank}(\text{comments per 100 reactions})$. The dashed line marks the cut at 0.90.

Figure 4 combines two observable signals into one score: the audience share of Sad and Angry and the number of comments per one hundred reactions. The histogram is heavy on the left with a long right tail. Using the fixed cut at RTI ≥ 0.90, **3.445%** of posts are above the line in this run (0.0344495101 as a fraction). This tail is small but distinct and is suitable for focused review.

5.4 Engagement Anomalies

Figure 5 applies a robust outlier rule on the log scaled engagement. Only **0.023%** of posts exceed the anomaly threshold in this run (0.0002292268 as a fraction). These outliers appear in short bursts rather than being spread evenly across time, which is consistent with coordinated amplification and rapid cross page sharing.

Posting is most frequent late in the evening in UTC. Typical posts receive near-unanimous reactions and have many more likes than shares, while a small and informative slice shows high polarization together with more shares than likes. The triage score isolates a compact review set at RTI ≥ 0.90 and the robust anomaly rule focuses attention on very rare but important spikes in engagement.

5.5 Discourse Role Classification

To address the absence of content-level labels and characterize how Arabic Facebook pages participate in misinformation discourse, we annotate a stratified sample of 10,000 posts (3.6% of the full corpus) using Fanar-C-2-27B,[2] an Arabic-optimized large language model. Each post is classified into one of six discourse roles. The annotation text supplied to the model combines the post `Message`, the linked article headline (`Link Text`), and the article description (`Description`), prefixed with Arabic context tags (‘*arabian al-rabat*‘ for headline and ‘*wasf*‘ for description, rendered as <link title:> and <description:> in practice) and truncated to 500 characters, giving the model sufficient field context without exceeding the effective context window.

The six roles are: **Rumor Denial / Official Correction** (post explicitly states that a specific claim is a rumor, incorrect, or officially denied, including authority statements, corrections, and explicit rebuttals); **Unverified Claim / Misinformation Framing** (post presents a specific factual claim without visible verification, correction, or skepticism—the claim is asserted as true without countervailing signal); **Fact-Check / Verification Report** (post investigates a claim using evidence, sourcing, images, or explicit fact-check methodology and reaches a veracity conclusion); **News / Factual Report** (post reports an event, announcement, or development in a standard informational or news format, without obvious misinformation framing or correction language); **Opinion / Political Commentary** (post expresses a subjective viewpoint, editorial position, or political analysis without a specific verifiable factual claim as primary content); and **Incidental / Off-topic** (post uses misinformation-related terms colloquially in non-public-affairs contexts: sports transfer speculation, entertainment gossip, commercial content, religious narratives).

Table 3 reports the distribution. All 10,000 posts received a valid discourse-role label without requiring manual repair of malformed outputs; annotation quality was assessed through post-hoc inspection of label boundaries across 50 randomly sampled posts per role.

Finding 1: Arabic Facebook misinformation discourse is dominated by counter-information. Rumor Denial and Official Correction are the single largest discourse role category at 30.6%, outnumbering Unverified Claim posts (12.3%) by a ratio of 2.5:1. Combined with Fact-Check posts (3.2%), counter-misinformation content accounts for 33.8% of the annotated sample, while misinformation-discourse posts (Denial + Unverified + Fact-Check) together account for 46.1%. This finding reframes the corpus: rather than a collection of false claims,

[2]Fanar API: `https://api.fanar.qa/v1`; model `Fanar-C-2-27B`.

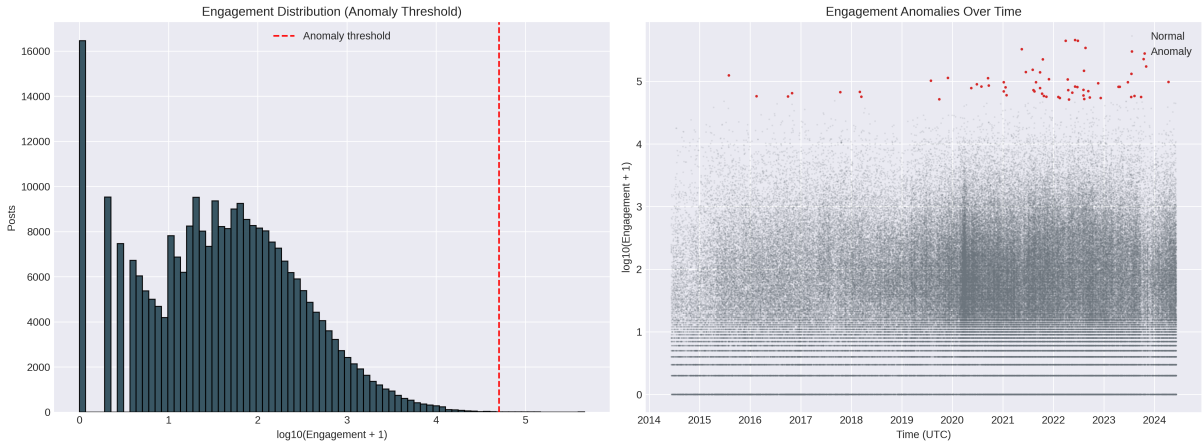


Figure 5: **Engagement anomalies.** Left: distribution of $\log_{10}(\text{Total Interactions} + 1)$ with a robust median and median absolute deviation threshold shown as a red dashed line. Right: timeline of posts with outliers highlighted.

Discourse Role	n	**%**
Rumor Denial / Official Correction	3,061	**30.6**
Incidental / Off-topic	1,931	19.3
Opinion / Political Commentary	1,811	18.1
News / Factual Report	1,645	16.4
Unverified Claim / Misinfo. Framing	1,235	12.3
Fact-Check / Verification Report	317	3.2
Counter-misinfo. total (Denial+FC)	*3,378*	*33.8*
Mis.-discourse total (above + Mis.)	*4,613*	*46.1*

Table 3: Discourse role distribution (n=10,000; 3.6% of corpus). Bold = dominant role. Counter-misinformation posts outnumber Unverified Claim posts by 2.5:1.

Country	**Denial**	**Unverified**	**Denial rate**
Algeria (DZ)	554	152	**78%**
Tunisia (TN)	342	100	77%
Syria (SY)	196	64	75%
Iraq (IQ)	151	60	72%
Morocco (MA)	436	175	71%
Egypt (EG)	393	246	62%

Table 4: Denial rate (Rumor Denial ÷ (Denial + Unverified Claim)) for the six most-represented countries ($n \geq 500$ posts). Palestine (n=193, 81%) and Jordan (n=372, 80%) exceed these rates but from smaller samples.

ArabMisinfo primarily captures the Arabic public-affairs ecosystem's response to misinformation. Engagement signals reflect this: Rumor Denial posts attract the highest median Likes (34), while Fact-Check posts attract the lowest (21), suggesting that simple denial is more socially rewarded than evidence-based verification—a pattern consistent with prior findings on diffusion without endorsement (Vosoughi et al., 2018).

Finding 2: Keyword sampling noise is precisely quantified. The Incidental / Off-topic role (19.3%) captures posts that use *ishā'a* (rumor), *mughalāta* (fallacy), or *taḍlīl* (misinformation) in non-public-affairs contexts—sports transfer speculation, gaming rumors, celebrity gossip, commercial content, and religious narratives. This directly quantifies the sampling bias introduced by keyword-based collection that prior work could only estimate qualitatively. Removing Incidental posts from the corpus leaves an 80.7% on-topic subset for downstream supervised tasks.

Finding 3: Denial rates vary meaningfully across countries. Table 4 reports the denial rate per country for the six most-represented country groups in the annotated subset ($n \geq 500$ posts each). Among these, Algeria (78%) and Tunisia (77%) exhibit the highest denial rates and Egypt (62%) the lowest. Palestine (n=193, 81%) and Jordan (n=372, 80%) show higher denial rates from smaller samples and are reported in the text for completeness. The 2020 collection year shows the highest denial rate of any year (38.9% Rumor Denial vs. 11.7% Unverified Claim), consistent with the COVID-19 infodemic triggering large-scale correction campaigns across Arabic-language social media.

6 Discussion

The four diagnostics in Section 5 describe how Arabic content gains attention on Facebook using platform-native signals. They are not truth detectors but transparent filters that help direct limited review capacity toward posts with unusual diffu-

sion and reaction patterns, consistent with prior findings on selective exposure, diffusion without endorsement, and coordinated activity.

Unanimous reactions and selective exposure Most posts show near-unanimous reactions, pointing to audience homophily rather than open contestation. Reactions should be treated as behavioral signals rather than precise sentiment labels, since their semantics vary across contexts.

Posting rhythms and attention windows Activity clusters around evening hours in many regions, and engagement peaks do not always track posting volume. Social platforms display strong diurnal cycles, and evening usage often correlates with higher spread of questionable content, supporting time windows as an audit dimension.

Share-to-like imbalance as a triage cue Shares are rarer and more effortful than likes, so posts with more sharing than liking can indicate diffusion without visible endorsement. Prior work shows that false or misleading items can travel farther and faster than true ones (Vosoughi et al., 2018), supporting share-to-like imbalance as a scalable screening rule before human review.

Risk triage and engagement anomalies Combining negative reaction share with discussion intensity isolates a tractable subset for review, while a robust anomaly score highlights rare spikes in short bursts. Temporal concentration and synchronized behaviors are characteristic of coordinated campaigns, making bursty anomalies reasonable audit targets even without user-level data.

Discourse roles and corpus character The discourse role classification (Section 5.5) substantially clarifies the corpus character. Rumor Denial and Official Correction account for 30.6% of the annotated subset, outnumbering Unverified Claim posts (12.3%) by 2.5:1, revealing that the corpus captures the Arabic public sphere's *response* to false claims as much as the claims themselves. Training a misinformation classifier on ArabMisinfo without role-aware filtering risks learning correction discourse rather than false claim production. The 19.3% Incidental / Off-topic rate provides a concrete noise estimate for keyword-based collection.

Implications and limits The corpus is best used for descriptive and diagnostic studies, targeted sampling for annotation, and comparative analysis across page types and periods. The discourse role annotation partially addresses the ground-truth gap for the 10,000-post annotated subset; the remaining 264,837 posts retain only platform-signal diagnostics. Reactions are coarse proxies, platform affordances are Facebook-specific, and dialect variation complicates text processing.

Validation paths The Unverified Claim (1,235 posts) and Fact-Check (317 posts) subsets provide a candidate pool for ClaimReview linkage and FEVER-style evaluation. Linking these posts to external Arabic fact-checking repositories would enable supervised precision-recall evaluation of the platform-signal diagnostics without re-annotating the full corpus.

7 Social Impact and Applications

ArabMisinfo delivers a reusable Arabic Facebook corpus with full engagement signals, a clear schema, discourse role annotations, and code that reproduces every figure and metric. The release includes a dataset card, a field dictionary, and a public license and data statement for research use.

Evaluation-ready diagnostics The corpus supplies four simple diagnostics that work at scale using platform fields only. Polarization Index and Share-to-Like Anomaly capture endorsement balance and diffusion imbalance. The Risk Triage Index ranks items for review using negative reaction share and comments per 100 reactions. The Engagement Anomaly Rule flags heavy-tail outliers with a robust score. These views create transparent baselines for Arabic misinformation studies and can be reproduced from the released files.

Discourse-annotated subset for supervised learning The 10,000-post annotated subset with six discourse roles enables supervised classification experiments without requiring full corpus re-annotation. The Unverified Claim (1,235 posts) and Fact-Check (317 posts) subsets provide a candidate pool for future claim verification and ClaimReview linkage, while the Rumor Denial subset (3,061 posts) supports counter-narrative and correction detection. Removing the Incidental / Off-topic posts (19.3%) yields a clean 80.7% on-topic sample for domain adaptation and weak supervision.

Resource for Arabic NLP and measurement Posts include text, page fields, UTC time, link context, and detailed reactions, enabling work on stance detection, claim matching, retrieval, and

domain adaptation. The seven reaction types allow analyses that move beyond like counts and support evaluation of diagnostic models that rely only on platform signals. Country-level discourse role variation (Section 5.5) further supports sociolinguistic and comparative NLP studies across Arabic-speaking regions.

Comparability and reuse Because the indices are defined from platform fields, results are comparable across time and across future Arabic resources. The corpus can anchor cross-platform and cross-language comparisons that test whether diffusion without endorsement is a general pattern or a setting-specific one.

Accountability with safeguards The frozen snapshot covers 2014 to 2024, a period with major public events in Arabic-speaking regions, and supports audits of spread and attention using transparent rules. The release excludes media files and individual users, and ships with a license and data statement to guide responsible use.

8 Conclusion

ArabMisinfo offers a large, well-documented resource of Arabic Facebook posts for work on misinformation in news and public affairs. The corpus covers 274,837 public posts collected between 2014 and 2024, and preserves platform signals including post text, page fields, links, UTC timestamps, and the full set of reactions.

Our four diagnostic views give clear, actionable patterns. Posting is most frequent late in the evening UTC, with the top three hours at 22:00, 23:00, and 21:00. The median post shows a Polarization Index of 1.00, indicating near-unanimous one-sided reactions, and a Share-to-Like Anomaly of -1.00, meaning shares are about one-tenth of likes. A compact high-risk slice is isolated at RTI ≥ 0.90, covering 3.445% of posts. Very rare engagement spikes are also visible: 0.023% of posts exceed a robust anomaly threshold and appear in short bursts over time.

The discourse role annotation of 10,000 posts reveals three findings that go beyond the diagnostic indices. First, Arabic Facebook misinformation discourse is dominated by counter-information: Rumor Denial and Official Correction (30.6%) outpaces Unverified Claim posts (12.3%) by 2.5:1, establishing ArabMisinfo, to our knowledge, as the first large-scale Arabic Facebook corpus to explicitly quantify counter-information dominance through discourse-role annotation. Second, 19.3% of posts are incidental keyword matches, providing the first direct empirical estimate of noise from Arabic keyword-based misinformation collection. Third, Rumor Denial spiked to 38.9% during 2020 (vs. 11.7% Unverified Claim), consistent with COVID-19 infodemic correction campaigns, with meaningful variation across countries (Algeria 78% denial rate; Egypt 62%).

Taken together, these results show how platform-native signals and discourse role classification can surface content that spreads widely without clear endorsement and guide targeted review at scale. Future work should extend the annotation to the full corpus, link Unverified Claim and Fact-Check posts to external ClaimReview records for FEVER-style evaluation, and incorporate account-level covariates and cross-platform data to test the generalizability of these patterns.

9 Limitations

This work acknowledges several important limitations. First, the full 274,837-post corpus lacks human-verified ground-truth labels for misinformation at the claim level. The discourse role annotation (Section 5.5) partially addresses this for the 10,000-post annotated subset, revealing that 30.6% of posts are Rumor Denial / Official Correction, 12.3% are Unverified Claim / Misinformation Framing, and 3.2% are Fact-Check posts. The remaining 264,837 posts still lack content-level labels. For the annotated subset, the weakest label boundary is between Rumor Denial and Unverified Claim—posts that question whether a claim is true ("Is this a rumor or fact?") can plausibly fall into either category; a formal inter-annotator agreement study is in progress.

Second, the keyword-based retrieval strategy overrepresents posts that explicitly mention verification terminology (*ishāʿa*, *mughalāta*, *taḍlīl*) while potentially missing implicit or coded misinformation content. The discourse role annotation quantifies this directly: 19.3% of annotated posts are Incidental / Off-topic—posts using these terms in sports, entertainment, or commercial contexts unrelated to public-affairs misinformation. Removing these posts leaves an 80.7% on-topic subset; future collection should apply topic filters post-retrieval.

Third, the dataset is constrained to Facebook and cannot support cross-platform comparisons beyond

what researchers construct independently. The discontinuation of CrowdTangle in August 2024 limits both temporal extensibility and methodological reproducibility, as future researchers cannot replicate our exact collection procedures or extend the temporal coverage using the same infrastructure.

Fourth, while the dataset includes references to multimodal content such as photos, videos and links, we do not provide the actual media files, only textual metadata and descriptions. This limits applicability for multimodal misinformation detection research, particularly for deepfake or manipulated media analysis. Similarly, our reaction based sentiment proxy, while platform native and scalable, has not been validated against human sentiment annotations and may not capture nuanced emotional responses.

Fifth, the focus on public Pages excludes private discourse, closed groups, and user to user interactions, which may exhibit different misinformation dynamics. The temporal coverage (2014-2024) means the dataset does not capture before 2014 discourse or developments after June 2024. Finally, in this release we do not apply text normalization. If future users normalize the text, it may reduce dialectal and stylistic variation that could be informative for certain research questions.

10 Ethics Statement

This research adheres to ethical principles for computational social science and misinformation studies. All data were collected from public Facebook Pages through the authorized CrowdTangle API in compliance with Meta's terms of service prior to the platform's discontinuation. No private user accounts, personal messages, group content, or individual user profiles were accessed or collected. The broader MARSAD data collection protocol was reviewed and approved by the HBKU Institutional Review Board (IRB # QBRI-IRB-2023-1).

While the dataset comprises publicly posted content, we recognize that large-scale aggregation and systematic analysis may surface patterns or reveal information not anticipated by original posters. To mitigate potential harms, we have: (1) limited the release to post level content and page metadata without individual user identifiers, (2) provided only aggregate engagement metrics rather than individual user interactions, and (3) documented clear intended use cases and restrictions in the accompanying data statement.

We acknowledge significant dual-use risks inherent in misinformation research. While this resource aims to advance detection capabilities and support fact-checking efforts, the dataset could potentially be exploited to: study manipulation strategies for adversarial purposes, target Arabic speaking communities with tailored misinformation, or develop evasion techniques against detection systems. We strongly urge researchers to consider these risks carefully and to conduct their work within appropriate ethical review frameworks.

We note that misinformation disproportionately affects vulnerable populations, including linguistic minorities, politically marginalized groups, and communities with limited media literacy resources. Research using this dataset should be mindful of these power dynamics and should prioritize approaches that empower rather than surveil Arabic speaking communities. We encourage researchers to engage with regional stakeholders, fact checking organizations, and civil society groups when conducting studies with this resource.

Finally, we acknowledge that misinformation judgments are culturally and politically contingent. The discourse role annotation released with this corpus uses an LLM-based pipeline (Fanar-C-2-27B) without full inter-annotator validation. We encourage downstream users to treat the annotations as a starting point, conduct their own validation on subsets relevant to their research questions, and report annotation procedures and inter-annotator agreement transparently.

Acknowledgments

The work was supported by NPRP grant 14C-0916-210015 from the Qatar National Research Fund, part of the Qatar Research Development and Innovation Council (QRDI). The findings reported herein are solely the responsibility of the authors. The author used GPT-5.5 Sol and Claude 4.6 Opus to improve the writing quality through grammatical correction and paraphrasing. The author bears full responsibility for the accuracy, integrity, and content of the manuscript.

Dataset Availability

Researchers are required to agree to the dataset's terms of usage for accessing dataset. The dataset is available for research purposes only and can be requested by completing the access form at https://forms.gle/YUFdA16R6HkSZjp88.

References

Firoj Alam, Abul Hasnat, Fatema Ahmad, Md. Arid Hasan, and Maram Hasanain. 2024. ArMeme: Propagandistic content in Arabic memes. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21071–21090, Miami, Florida, USA. Association for Computational Linguistics.

Taghreed Alotaibi and Hmood Al-Dossari. 2024. A review of fake news detection techniques for arabic language. *International Journal of Advanced Computer Science and Applications*, 15(1).

Shelley Boulianne and Christian P Hoffmann. 2024. Perceptions and concerns about misinformation on facebook in canada, france, the us, and the uk. *International Journal of Public Opinion Research*, 36(4):edae048.

Aaron Clauset, Cosma Rohilla Shalizi, and Mark E. J. Newman. 2007. Power-law distributions in empirical data. *SIAM Rev.*, 51:661–703.

W. J. Conover and Ronald L. Iman. 1981. Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician*, 35(3):124–129.

DeepFake Audio Rangers. 2024. Arabic audio deepfake. `https://huggingface.co/datasets/DeepFake-Audio-Rangers/Arabic_Audio_Deepfake`. Hugging Face Datasets. Version (commit): 0bf925d (2024-09-27). License: ODC-BY. Accessed: 2025-09-28.

Peter Dizikes. 2018. Study: On twitter, false news travels faster than true stories. MIT News, Massachusetts Institute of Technology.

Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Commun. ACM*, 59(7):96–104.

Mohamed Seghir Hadj Ameur and Hassina Aliane. 2021. Aracovid19-mfh: Arabic covid-19 multi-label fake news & hate speech detection dataset. *Procedia Computer Science*, 189:232–241. AI in Computational Linguistics.

Fatima Haouari, Maram Hasanain, Reem Suwaileh, and Tamer Elsayed. 2021a. ArCOV-19: The first Arabic COVID-19 Twitter dataset with propagation networks. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 82–91, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Fatima Haouari, Maram Hasanain, Reem Suwaileh, and Tamer Elsayed. 2021b. ArCOV19-rumors: Arabic COVID-19 Twitter dataset for misinformation detection. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 72–81, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Maram Hasanain, Md. Arid Hasan, Fatema Ahmad, Reem Suwaileh, Md. Rafiul Biswas, Wajdi Zaghouani, and Firoj Alam. 2024. ArAIEval shared task: Propagandistic techniques detection in unimodal and multimodal Arabic content. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 456–466, Bangkok, Thailand. Association for Computational Linguistics.

David C. Hoaglin. 2013. *Volume 16: How to Detect and Handle Outliers*.

Amelia Johns, Francesco Bailo, and Marian-Andrei Rizoiu. 2024. Meta just closed a vital online research tool. it's bad news for the fight against misinformation.

Ashwaq Khalil, Moath Jarrah, Monther Aldwairi, and Manar Jaradat. 2022. Afnd: Arabic fake news dataset for the detection and classification of articles credibility. *Data in Brief*, 42:108141.

Zoe Kleinman. 2018. Fake news 'travels faster', study finds.

Sammi Krug. 2016. Reactions now available globally. Meta Newsroom. Accessed 2025-10-12.

Kartik Kuckreja, Parul Gupta, Injy Hamed, Thamar Solorio, Muhammad Haris Khan, and Abhinav Dhall. 2025. Tell me habibi, is it real or fake? *Preprint*, arXiv:2505.22581.

Sanaa Laabar and Wajdi Zaghouani. 2024. Multidimensional insights: Annotated dataset of stance, sentiment, and emotion in Facebook comments on Tunisia's July 25 measures. In *Proceedings of the Second Workshop on Natural Language Processing for Political Sciences @ LREC-COLING 2024*, pages 22–32, Torino, Italia. ELRA and ICCL.

Ingrid Lunden. 2020. Facebook adds new 'care' emoji reactions on its main app and in messenger. TechCrunch.

Mohamed A. Mostafa and Ahmad Almogren. 2024. VERA-ARAB: unveiling the arabic tweets credibility by constructing balanced news dataset for veracity analysis. *PeerJ Computer Science*, 10:e2432. ECollection 2024.

NIST/SEMATECH. 2012. 1.3.5.17. detection of outliers. NIST Engineering Statistics Handbook. Accessed 2025-10-12.

Rasha Obeidat, Maram Gharaibeh, Malak Abdullah, and Yara Alharahsheh. 2022. Multi-label multi-class covid-19 arabic twitter dataset with fine-grained misinformation and situational information annotations. *PeerJ Computer Science*, 8:e1151. ECollection 2022.

Barbara Ortutay. 2024. Meta kills off misinformation tracking tool crowdtangle despite pleas from researchers, journalists.

Diogo Pacheco, Pik-Mai Hui, Christopher Torres-Lugo, Bao Tran Truong, Alessandro Flammini, and Filippo Menczer. 2020. Uncovering coordinated networks on social media. *ArXiv*, abs/2001.05658.

Justin Bonest Phillips. 2024. Likes vs. loves (and other emoji reactions): Facebook, women, and the gender emoji gap in us election campaigns. *Social Science Computer Review*, 42(4):930–946.

Seani Rananga, Bassey Isong, Abiodun Modupe, and Vukosi Marivate. 2024. Misinformation detection: A review for high and low-resource languages. *Journal of Information Systems and Informatics*, 6(4):2892–2922.

Eugenio Tacchini, Gabriele Ballarin, Marco L. Della Vedova, Stefano Moret, and Luca de Alfaro. 2017. Some like it hoax: Automated fake news detection in social networks. *ArXiv*, abs/1704.07506.

FANAR TEAM, Ummar Abbas, Mohammad Shahmeer Ahmad, Minhaj Ahmad, Abdulaziz Al-Homaid, Anas Al-Nuaimi, Enes Altinisik, Ehsaneddin Asgari, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Asim Ersoy, Masoomali Fatehkia, Mohammed Qusay Hashim, and 18 others. 2026. Fanar 2.0: Arabic generative ai stack. *Preprint*, arXiv:2603.16397.

Onur Varol, Emilio Ferrara, Clayton A. Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online human-bot interactions: Detection, estimation, and characterization.

Soroush Vosoughi, Deb K. Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359:1146 – 1151.