International Journal of Computer Science Trends and Technology (IJCST) – Volume 14 Issue 4, Jul-Aug 2026

RESEARCH ARTICLE OPEN ACCESS

Sudanese Arabic as a Low-Resource Dialect: Construction, Audit, and Cross-Period Analysis of a Three-Resource Suite

Mohammed Al. Osman. Abdalsalam[1], Howida Ali Abdulgader [1] and Aldaw Huessin[2]

College of Computer Science and Information Technology, Sudan University of Science and Technology, Khartoum, Sudan [1]

Department of Computer Science, College of Science, Mustansiriyah University, Baghdad, Iraq[2]

Abstract

Sudanese Arabic is poorly served by reusable natural-language-processing resources when compared with Modern Standard Arabic and with better-resourced regional varieties. This paper documents three connected Sudanese Arabic resources: a 6,755,955-sentence social-media corpus, a manually labelled sentiment corpus of 13,293 Google Play reviews, and a hybrid-annotated hate-speech resource of 40,000 annotated sentences with a 38,619-record filtered analytical subset. The contribution is a data and resource study rather than a new classifier. We report provenance, construction, annotation, quality control, and intended uses, and compare the three resources along a common set of structural dimensions. The large corpus then supports a cross-period comparison between a 2015–2020 partition of 6,295,975 sentences and a 2023–2025 conflict-era partition of 459,980 Telegram sentences, combining normalized lexical frequencies, conservative count thresholds, equal-size vocabulary sampling, domain-lexicon aggregation, and windowed collocation analysis. Verified conflict vocabulary—militias, armed actors, fighting, displacement, and affected locations—rises steeply in normalized terms, and the immediate context of the word for war moves from international and historical reference toward local actors and places. The unfiltered ranking, by contrast, is dominated by Telegram channel templates, which is why source-aware validation must precede any reading of frequency change as linguistic change. Every reported statistic is tied to a hashed source file and a reproducible script.

Keywords—Sudanese Arabic, low-resource dialect, corpus construction, diachronic analysis, sentiment dataset, hate speech, social media, Arabic NLP.

I. INTRODUCTION

Arabic NLP operates across a continuum that runs from Classical Arabic through Modern Standard Arabic (MSA) to the regional varieties used in everyday communication. Dialectal Arabic departs from MSA, and from other dialects, in vocabulary, morphology, spelling, code-switching, and pragmatic convention [1], [2]. Social-media writing exposes these differences most sharply: no orthography is enforced, and users mix dialect, MSA, names, emojis, hashtags, and foreign-language material in a single message. Normalization choices made early in a pipeline therefore determine how much of that evidence survives to the modelling stage [3].

Resource availability is uneven across the varieties. MADAR, NADI, and QADI have made controlled dialect-identification research possible [4]–[6], and morphologically annotated collections have widened coverage for selected dialects. Multi-dialect benchmarks do not, however, remove the need for dialect-specific resources. Country-level or city-level labels conceal internal variation, and a corpus built for dialect identification is rarely suitable for sentiment work, harmful-language analysis, or temporal discourse research. Recent studies continue to describe Algerian, Hassaniya/Mauritanian, and Libyan Arabic as low-resource settings that require targeted collection and adaptation [7]–[9].

Sudanese Arabic sits in exactly this position. It carries variation shaped by Sudan's geography, multilingual setting, and cross-border exchange, yet it remains computationally underrepresented. Existing Sudanese resources—sentiment datasets [10], a Sudanese-oriented pretrained encoder [11], and Sudanese portions of multidialect collections—demonstrate feasibility but do not amount to a single documented suite spanning general social-media language, customer-experience sentiment, and harmful discourse. Nor do they supply the temporal baseline needed to ask how vocabulary and recurring topics differ between the period surrounding the Sudanese revolution and the conflict that began in April 2023.

This paper addresses that gap by documenting three datasets built by the authors' research team: a social-media corpus of 6,755,955 sentences drawn from Twitter and Telegram; a labelled set of 13,293 Google Play reviews from eight Sudanese service applications; and a hybrid-annotated hate-speech corpus derived from the broader collection. Earlier publications report modelling experiments on parts of this material. Those classification results are cited but not reproduced here, because the unit of contribution in the present work is the resource itself: what was collected, how it was organized, which labels were created, which quality checks were applied, and which questions the combination now supports.

The paper makes four contributions. It documents a three-resource Sudanese Arabic suite with explicit provenance, purpose, composition, annotation, and quality control. It offers a reproducible structural comparison showing how general-domain, sentiment, and harmful-language resources serve distinct but related needs. It presents a cross-period comparison of the 2015–2020 and 2023–2025 partitions using normalized lexical change, equal-size vocabulary sampling, domain-

lexicon aggregation, topic-proportion comparison, and collocation evidence. And it records the quality-control decisions and source limitations behind those numbers, including early non-Sudanese and MSA-news contamination and the platform mismatch between the two partitions.

The argument throughout is that low-resource dialect development takes more than raising sentence counts. A usable resource ecosystem has to join broad coverage with task-specific labels, transparent quality control, and analyses that separate linguistic patterns from collection artefacts. Four objectives follow: (1) construct and document a large-scale Sudanese Arabic corpus that enriches the computational resources available for the dialect; (2) compare the three constituent datasets in scale, source, temporal coverage, annotation, and intended use; (3) identify the lexical items disproportionately represented in the 2023–2025 partition once unequal Arabic-token totals are controlled; and (4) determine whether the immediate contexts of recurrent words, especially الحرب (war), differ between the two periods, or whether apparent differences are better explained by platform composition, channel templates, or corpus-construction decisions.

Section II reviews related resources and states the gap. Section III gives an overview of the suite. Sections IV–VI describe the large temporal corpus, the sentiment corpus, and the hate-speech corpus. Section VII sets out the shared methodology and Section VIII reports the cross-period results. Section IX discusses implications for low-resource dialect NLP, Section X covers data availability and reproducibility, and Section XI states limitations and ethical considerations. Section XII provides resource cards and recommended uses, and Section XIII concludes.

II. RELATED RESOURCES AND RESEARCH GAP

A. Arabic Dialect Resources

Arabic dialect processing is shaped by geographic variation, non-standard spelling, and domain dependence. MADAR provides fine-grained dialect material, NADI supports country-level identification, and QADI targets naturally occurring dialect identification [4]–[6]. These benchmarks improved comparability, but they were built for dialect identification rather than for longitudinal analysis or Sudanese-specific content moderation.

The wider Arabic NLP literature returns repeatedly to dataset availability and representativeness as the binding constraints [1], [2]. Recent work builds new dialect resources rather than assuming that MSA or pan-Arabic data transfer uniformly. A 2025 Algerian study frames dialect identification explicitly as a low-resource transfer problem [7]; a 2025 Hassaniya study describes the Mauritanian variety as low-resource and constructs domain-specific Facebook data for sentiment analysis [8]; and work on Libyan Tripolitanian speech reports the same need for a dedicated local corpus [9]. A dialect, in short, can be widely spoken and still computationally underrepresented.

B. Sudanese Arabic Resources

SudaBERT introduced a Sudanese-oriented pretrained encoder [11]. Mhamed et al. developed SudSenti2 and SudSenti3 and evaluated a CNN architecture on them [10]. Both sentiment datasets are third-party resources and are not claimed here; the Google Play corpus documented in Section V is a separate dataset built by the our team.

C. Available Sudanese Arabic Datasets and Their Limitations

Published Sudanese Arabic text resources remain concentrated in sentiment analysis and other narrowly defined tasks. One of the earliest resources is the telecommunications sentiment dataset introduced by Ismail et al. [12]. It contains 4,712 Twitter posts discussing telecommunications services in Sudan and was manually labelled by three native Arabic speakers. The corpus provides an early benchmark for Sudanese Arabic sentiment classification, but its relatively small size, single-platform origin, and exclusive focus on one commercial sector limit its representativeness and transferability to other domains. It does not cover harmful language, general social-media discourse, or conflict-era communication, and its size is insufficient for language-model pretraining.

SSA-SDA is another early Sudanese Arabic sentiment resource. It contains 5,456 political tweets and was developed for subjectivity and three-class sentiment analysis, covering Positive, Negative, and Neutral labels [13]. The study compared manual annotation with automatically generated labels and also introduced a Sudanese polarity lexicon. Its political focus makes it useful for studying opinionated discourse, but the corpus is limited to Twitter, contains only a few thousand records, and combines manual and automatic annotation procedures. Consequently, annotation-source effects must be considered when using it for supervised evaluation, and its political-domain vocabulary may not transfer reliably to customer reviews, everyday conversations, or conflict-related harmful-language analysis.

Abuuznien et al. introduced a further Sudanese Arabic sentiment dataset containing 2,116 tweets related to the Sudanese ride-hailing services Tirhal and Halan [14]. The dataset comprises 768 Positive, 841 Negative, and 507 Neutral tweets. Although it supports three-class evaluation and reflects a locally relevant service domain, it is the smallest of the principal published Sudanese sentiment datasets. Its class distribution is moderately unequal, and its restriction to two ride-hailing services limits lexical, topical, and platform diversity. Models trained on this corpus may therefore learn service-specific expressions rather than sentiment patterns that generalize across Sudanese Arabic domains. SudSenti2 and SudSenti3 provide broader sentiment resources

developed from different platforms [10]. SudSenti2 contains 4,000 Facebook and YouTube posts, including 2,027 Positive and 1,973 Negative instances. Neutral posts were removed during construction, making the dataset suitable only for binary sentiment classification. SudSenti3 contains 7,109 Twitter posts, divided into 2,523 Positive, 2,639 Negative, and 1,947 Neutral instances. Three native Sudanese Arabic speakers independently labelled the source posts, and majority decisions were used to establish the final labels. These datasets provide stronger annotation documentation than several earlier resources, but they remain relatively small, platform-specific, and sentiment-oriented. SudSenti2 cannot support three-class evaluation, while the smaller Neutral class in SudSenti3 requires the use of macro-F1 and class-specific recall in addition to overall accuracy. Neither resource provides broad unlabeled coverage, harmful-language labels, or matched temporal partitions.

Beyond sentiment analysis, Lisan supplies morphologically annotated data for Yemeni, Iraqi, Libyan, and Sudanese Arabic [15]. The complete collection contains approximately 1.2 million tokens, but its distribution across dialects is highly unequal. The Yemeni component contains approximately 1.05 million tokens, whereas the Sudanese, Iraqi, and Libyan components contain approximately 50,000 tokens each. The Sudanese material was manually collected from Facebook and YouTube posts and comments and annotated with segmentation, part-of-speech information, lemmas, and English glosses. This detailed annotation makes the Sudanese component valuable for morphology, tokenization, and linguistic analysis. However, approximately 50,000 tokens provide limited lexical and topical coverage for continued pretraining or large-scale statistical analysis. The corpus also does not include sentiment, hate-speech, source-period, or conflict-related labels.

Sudanese-Flores extends the FLORES+ machine-translation benchmark to Sudanese Arabic through 2,009 translated sentences from the DEV and DEVTEST partitions [16]. The resource enables controlled evaluation of translation into and from Sudanese Arabic and addresses an important benchmark gap. Nevertheless, its size and design make it primarily an evaluation dataset rather than a training corpus. Because its sentences are translations of an existing benchmark, it does not replace naturally occurring Sudanese conversations, social-media posts, customer reviews, or conflict discourse. It also provides no sentiment or harmful-language annotations. Tarab provides a substantially larger multidialect collection of Arabic lyrics and poetry [17]. The complete corpus contains 2.56 million verses and more than 13.5 million tokens covering Classical Arabic, Modern Standard Arabic, and six regional varieties, including Sudanese Arabic. Its linguistic, geographic, and historical metadata support comparative and cultural analysis. However, the published corpus description does not separately quantify the number of Sudanese verses or tokens. Tarab is also strongly concentrated on lyrics and poetry, where figurative language, repetition, rhyme, and artistic conventions differ from everyday social-media communication. It therefore expands Sudanese representation within creative-text resources but cannot be treated as a general Sudanese Arabic corpus or as a substitute for task-specific labelled datasets.

Taken together, the published Sudanese Arabic resources remain fragmented by task, size, platform, period, genre, and annotation type. The telecommunications, SSA-SDA, ride-hailing, SudSenti2, and SudSenti3 datasets provide useful sentiment labels, but each contains only a few thousand records and is restricted to political or commercial social-media domains. Lisan provides detailed morphological annotation but only about 50,000 Sudanese tokens. Sudanese-Flores is a small translated evaluation benchmark, while Tarab is a large multidialect creative-text corpus whose Sudanese portion is not separately quantified.

None of these resources alone combines million-scale naturally occurring Sudanese Arabic text, separate sentiment and harmful-language annotations, cross-period coverage, and source-aware quality auditing. The present suite addresses this fragmentation by connecting a broad social-media collection with task-specific sentiment and hate-speech resources, while explicitly documenting the limitations arising from platform composition, domain coverage, annotation procedures, and representativeness.

D. Resource Descriptors and Diachronic Analysis

Resource-descriptor work emphasises provenance, labelling design, component structure, and supported use cases [18]. Diachronic corpus research recommends pairing macro-level statistical patterns with collocation and concordance evidence rather than interpreting topic-model output alone [19], and temporal word-representation research shows the value of measuring how lexical usage shifts between time segments [20]. The present study follows that direction, combining normalized frequencies, vocabulary-growth curves, lexicon categories, topic proportions, collocations, and qualitative validation.

Outside Arabic, recent conflict- and hate-focused resources reflect the same priorities: the CEHA conflict-event dataset for the Horn of Africa [21], the MetaHate meta-collection [22], and the explanation-enhanced HateCOT dataset [23] each pair explicit provenance with task-ready labels for low-resource evaluation. This suite extends that practice to Sudanese Arabic conflict discourse.

III. RESOURCE SUITE OVERVIEW

TABLE I. VERIFIED STRUCTURAL COMPARISON OF THE THREE RESOURCES

Property	Large temporal corpus	Sentiment corpus	Hate-speech corpus
Primary purpose	General corpus and temporal analysis	Customer-experience sentiment	Harmful-language research
Verified / intended size	6,755,955 sentences	13,293 reviews	40,000 annotated; 38,619 filtered analytical records
Sources	Twitter and Telegram	Google Play	Sampled Twitter and Telegram material
Period	2015–2020 and 2023–2025	Dates stored per review	2015–2020 and 2023–2025
Annotation	Unlabelled corpus	Manual team labelling with reconciliation	Weak supervision, two LLMs, native-speaker verification
Principal labels	Time and source metadata	Positive, Negative, Neutral	HATE, OFFENSIVE, NEUTRAL; binary mapping also supplied
Quality control	Filtering, script checks, deduplication, source audit	Reconciliation and label validation	Tiered resolution and stratified human review
License	Academic-research use; identifier assigned at public release	Academic-research use; identifier assigned at public release	Academic-research use; identifier assigned at public release

The three resources were created for different purposes but share a Sudanese Arabic focus. Table I summarises their verified properties, and Fig. 1 presents the suite as a progression from broad language coverage to increasingly task-specific annotation.

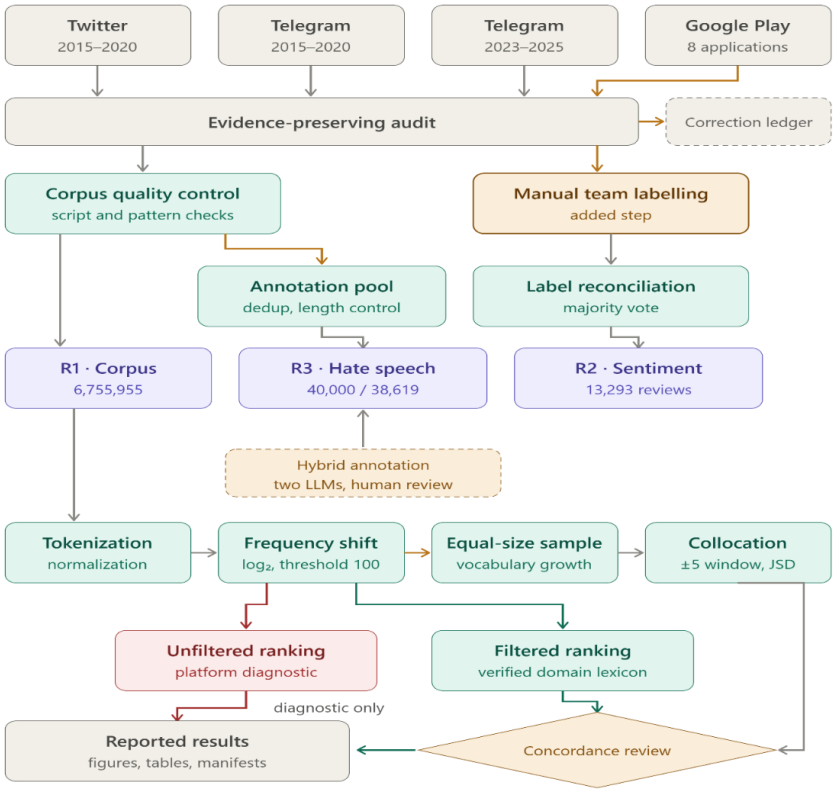

Fig. 1. Connected functions of the three-resource Sudanese Arabic suite.

A. Common Documentation Model

Each resource is documented under the same six headings: identity, provenance, structure, annotation, quality control, and availability. Identity defines the resource and keeps third-party comparison datasets from being counted as contributions. Provenance records platforms, periods, and file locations. Structure covers the sentence or review unit, fields, labels, and derived variants. Annotation states whether a resource is unlabelled, manually labelled, or hybrid annotated. Quality control records filtering, reconciliation, agreement, and review evidence. Availability separates a verified repository link from an intended but unconfirmed public release.

The shared model is what makes comparison meaningful across a two-order-of-magnitude difference in scale. A 6.76-million-sentence pretraining corpus should not be judged by the criteria applied to a manually labelled review dataset; equally, the presence of labels does not compensate for narrow domain coverage. Scale, representativeness, annotation depth, and ethical sensitivity are therefore treated as separate dimensions throughout.

B. Dataset Schemas

In the large corpus, one physical line is the sentence record and the file name supplies source grouping and temporal partition. The files do not carry a verified per-record author, channel, or exact timestamp for every earlier sentence, so any claim that would require those fields is excluded from this paper.

The sentiment CSV schema contains id, text, source_app, sector, score, sentiment, date, thumbs_up, char_count, and token_count. The sentiment field is the research label; score is the application-store rating. The two must not be conflated—a review labelled Neutral may still carry a high or low star score when its text is mixed or not directly evaluative.

The hate-speech TSVs contain text and label. The three-class file uses HATE, OFFENSIVE, and NEUTRAL; the binary file maps the two harmful categories to HARMFUL and retains NEUTRAL. Annotation-source outputs and agreement files are intermediate evidence, not additional independent corpora.

IV. DATASET 1: LARGE SUDANESE ARABIC TEMPORAL CORPUS

A. Sources and Temporal Partitions

The corpus holds exactly 6,755,955 sentence records. An earlier partition of 6,295,975 sentences covers 2015–2020 and comprises three Telegram files (2,271,187, 1,748,262, and 829,830 sentences) alongside one Twitter file of 1,446,696 sentences. A later partition of 459,980 sentences was collected during 2023–2025 across 14 Telegram channels.

Normalizing keyword counts per 1,000 tweets across four phases of the revolutionary period—Before (2018), Revolution (Dec–Apr), Post-Bashir (Apr–Aug), and Transitional (Aug onward)—yields the phase-dependent pattern in Fig. 2. During mobilization, anti-regime terminology dominates: the protest chant *Tasqut* ("Fall") reaches 478.4 per 1,000 tweets, with references to Al-Bashir at 25.7 and general revolution discourse at 30.9. After Omar al-Bashir's removal, attention moves from mobilization to the terms of transfer, and mentions of the Military (*Al-Askar*) rise to 40.9 while calls for Civilian (*Al-Madaniyya*) rule peak at 6.5. In the transitional phase the focus shifts again to institutional leadership: Prime Minister Hamdok reaches 42.6 per 1,000 tweets while *Tasqut* falls back to 24.2. The sequence tracks the documented political chronology, which is the first indication that the collection captures period-specific discourse rather than undifferentiated noise.

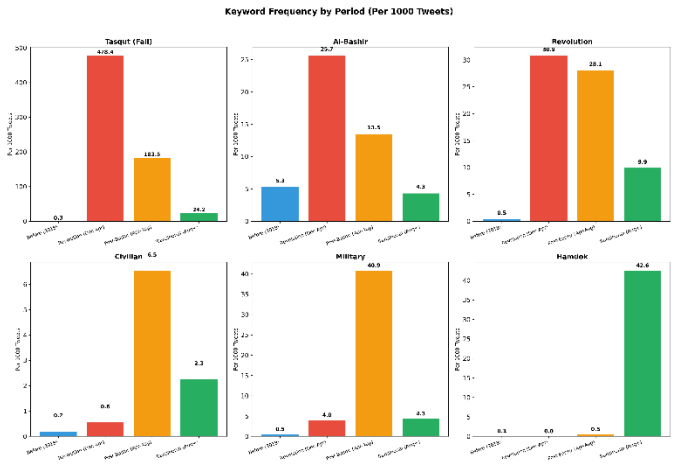

Fig. 2. Keyword frequency per 1,000 tweets across four phases of the 2018–2019 period.

B. Construction and Quality Control

Construction kept raw collection, general corpus storage, and stricter task-specific filtering apart. Early filtering passes admitted non-Sudanese material and MSA-style news content, so quality control added source review, Arabic-script checks, removal of identified non-Sudanese location patterns, and suppression of news-template patterns when the annotation pool was built. Duplicate removal and sentence-length controls were applied to the hate-speech sampling pool only.

The corpus layers must be kept distinct. The 459,980-record conflict-era file is the author-confirmed partition; a stricter Arabic-containing subset of 306,132 sentences matches the annotation-pool documentation. The audit identified 153,848 later records with no Arabic-script character under the Unicode script check. Those records are retained for provenance and excluded from Arabic lexical statistics; they may encode links, channel metadata, or Latin-script material, but their precise composition would require a separate content taxonomy and is not inferred here. The complete earlier corpus is likewise not identical to the fully deduplicated pool used for hate-speech sampling.

C. Intended Uses

The corpus supports continued pretraining, vocabulary analysis, source comparison, and temporal discourse research. It is not a balanced sociolinguistic sample of the Sudanese population. Public social-media channels privilege digitally active users and public discourse, and regional, demographic, and platform coverage is incomplete.

V. DATASET 2: MANUALLY LABELLED SENTIMENT CORPUS

A. Scope and Composition

The sentiment resource contains 13,293 Google Play reviews spanning eight applications in banking, telecommunications, and transport. Fig. 3 gives the application breakdown: Bank of Khartoum (4,630), MySudani (3,276), Tirhal (2,500), MyCashi (1,558), FaisalBank (1,159), SahApp (116), MTN_Sudan (52), and Raseedo_Sudani (2). After labelling and consensus reconciliation the corpus holds 5,954 Positive, 4,265 Negative, and 3,074 Neutral reviews (Fig. 4).

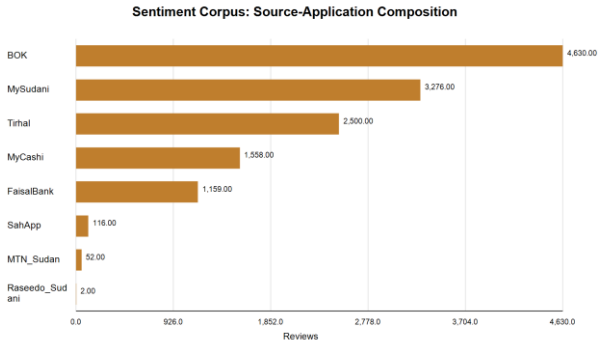

Fig. 3. Application composition of the labelled sentiment corpus.

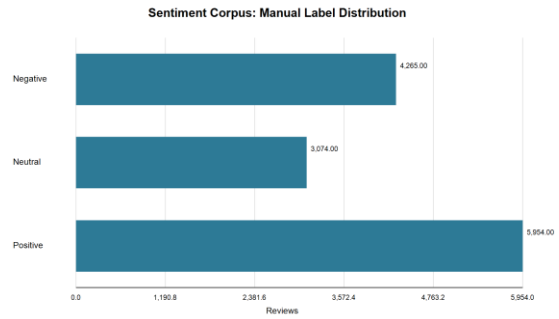

Fig. 4. Sentiment-label distribution across the 13,293-review corpus.

Broken down by sector (Fig. 5), banking contributes the largest volume with relatively balanced engagement (2,841 Positive, 2,447 Negative, 2,059 Neutral); telecommunications is more evenly split (1,483 Positive, 1,302 Negative, 661 Neutral); and transport leans positive (1,630 Positive against 516 Negative and 354 Neutral). Because the eight applications contribute unequally, aggregate percentages partly reflect source composition rather than sentiment alone.

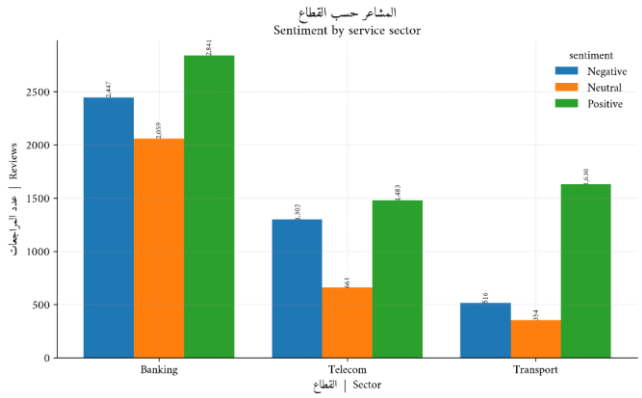

Fig. 5. Sentiment distribution by service sector across banking, telecommunications, and transport applications.

B. Validation and Reconciliation

Direct parsing surfaced minor structural anomalies, missing identifiers, and typographic label errors in the raw dataset. Reconciliation was deliberately conservative: unambiguous spelling errors were normalized, and incomplete records were restored only where independent predecessor files agreed unanimously. All valid original sentiment decisions were preserved and the raw source files were left untouched.

The resulting analytical copy passes the stated identifier, schema, and label-constraint checks. Every modification—spelling normalization, structural restoration, identifier recovery—is listed in the correction ledger. Restricting predecessor data to missing or incomplete records is what prevents older annotation versions from overwriting valid labels in the target dataset.

C. Annotation Documentation

Labels were assigned by the research team and audited before release. The audit found 210 exact duplicate records (about 1.58%) and 20 texts carrying conflicting labels (about 0.15%). Fifteen of the twenty conflicts (75.0%) were Positive-versus-Negative disagreements, concentrated on the boundary between mild criticism and constructive feedback. Conflicts were resolved by majority vote, with ties adjudicated by a lead researcher. The reconciled distribution is 44.79% Positive, 32.08% Negative, and 23.12% Neutral.

VI. DATASET 3: HYBRID-ANNOTATED HATE-SPEECH CORPUS

A. Sampling and Resource Layers

The hate-speech project drew 40,000 sentences from the cleaned social-media pool using period- and source-aware sampling: 19,000 sentences from the 2023–2025 Telegram material, 18,500 from earlier Twitter material, and 2,500 from earlier Telegram material.

The three-class TSV is a filtered analytical subset of 38,619 parsed records—7,168 HATE, 8,474 OFFENSIVE, and 22,977 NEUTRAL (Fig. 6)—produced after a minimum-length filter. Under the binary mapping the same subset yields 15,642 HARMFUL records against 22,977 NEUTRAL. The 40,000-sentence annotation resource and the 38,619-record filtered file are reported separately rather than as interchangeable totals.

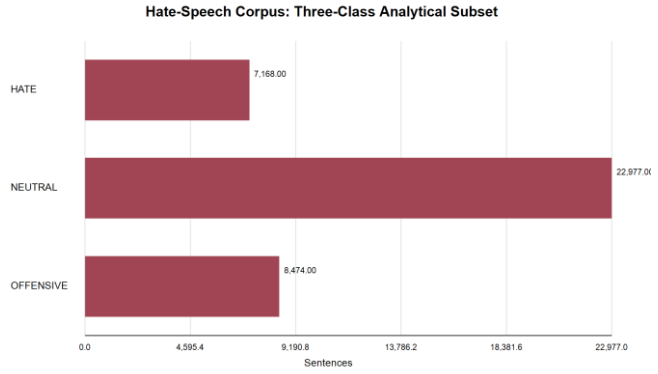

Fig. 6. Three-class distribution of the 38,619-record hate-speech analytical subset.

Fig. 7 shows how message length behaves across the three classes. The distributions overlap heavily, but their modes separate: OFFENSIVE peaks near eleven words, NEUTRAL near fourteen, and HATE near eighteen, with the HATE curve carrying the heaviest tail beyond thirty words. Short insults and longer group-directed passages therefore sit at opposite ends of the same range, which is a practical argument for contextual annotation: length alone separates none of the three classes, and a purely lexical rule applied to short messages would systematically misread the longer ones.

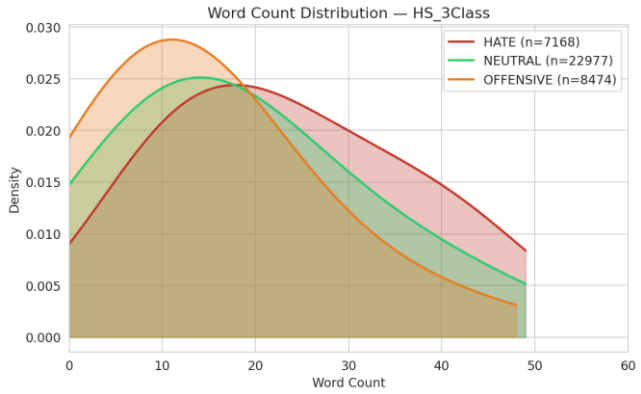

Fig. 7. Word-count distribution by class in the three-class hate-speech subset.

B. Hybrid Annotation

The annotation pipeline combined 42 Sudanese-oriented labelling functions designed according to weak-supervision and data-programming principles [24] with sentence-level annotations generated by GPT-4o mini [25] and Llama-3.1-70B-Instruct [26].

These consisted of 17 HATE functions, 8 OFFENSIVE functions, 11 NEUTRAL functions, and 6 conflict-resolution functions. The functions encoded lexical and contextual evidence related to group targeting, profanity, violence, political references, humanitarian reporting, and neutral-news contexts. The present procedure followed the general labelling-function principle introduced by Ratner et al. [24], but it did not use Snorkel's generative label model. Instead, the outputs of the 42 functions were combined through deterministic priority and conflict-resolution rules to produce one auxiliary weak-supervision label for each sentence.

This aggregated label was then treated as one annotation source alongside GPT-4o mini [25] and Llama-3.1-70B-Instruct [26]. GPT-4o mini processed each complete sentence using a five-shot Arabic prompt with Sudanese contextual framing. The prompt provided explicit definitions of HATE, OFFENSIVE, and NEUTRAL, together with five labelled examples selected to represent contextually ambiguous cases. These included the distinction between a neutral news report mentioning a political or armed group and a sentence using the same group reference derogatorily. Processing was conducted in 2,000 batches of 20 sentences, covering the complete 40,000-sentence collection. The annotation process required approximately 2.85 million input tokens and 593,000 output tokens and was completed in 157.7 minutes without reported processing errors.

Llama-3.1-70B-Instruct was deployed locally using 4-bit NormalFloat quantization [27]. It received the same Arabic five-shot prompt, label definitions, examples, and output schema used for GPT-4o mini. Processing the 40,000 sentences required approximately 22.8 hours at an average rate of 0.49 sentences per second.

The two LLM annotations and the aggregated weak-supervision label were combined through a three-tier resolution procedure. All three sources agreed on 15,983 sentences, for which the unanimous label was retained. Two of the three sources agreed on 21,330 sentences, for which the majority label was retained. For the remaining 2,687 sentences, the three sources assigned three different labels. GPT-4o mini was used as a heuristic tie-breaker because it had the highest participation rate in the Tier 2 majority decisions. This rule was an aggregation procedure rather than evidence that GPT-4o mini was the most accurate annotator; Tier 3 labels should therefore be treated as less certain than unanimous or majority-supported labels. After automated resolution, three native speakers of Sudanese Arabic reviewed a stratified sample of 10,000 sentences, representing 25% of the complete corpus. The sample covered the three final classes and the two source platforms. The reviewers changed 176 labels, corresponding to 1.76% of the reviewed sample and 0.44% of the complete corpus.

VII. COMPARATIVE METHODOLOGY

A. Immutable Sources and Reproducible Outputs

All analyses run against immutable source files. For each file the audit records absolute path, byte size, modification time, physical record count, and SHA-256 digest. Schema validation then flags malformed rows, missing identifiers, duplicate records, and invalid label values. Any accepted correction is written only to a derived analytical copy and documented in a separate ledger. Algorithm 1 formalizes this evidence-preserving procedure.

B. Arabic Tokenization and Normalization

The temporal analysis treats Unicode letter and mark sequences as tokens. Alef variants are mapped to bare alef, final alif maqsura to ya, waw and ya hamza forms to standalone hamza, and tatweel is removed. Raw and normalized files are kept separate. The scheme is deliberately conservative, since aggressive stemming would erase the dialect-specific evidence the corpus exists to preserve [3].

C. Normalized Frequency Shift

Raw counts are not comparable across partitions of unequal size. For period p, token w, observed count c_p(w), and Arabic-token total N_p, the frequency per million tokens is

$$F_p(w) = 10^6 \cdot c_p(w) / N_p \quad (1)$$

and the direction and magnitude of change are summarized by

$$L(w) = \log_2 [(F_{2023-25}(w) + \varepsilon) / (F_{2015-20}(w) + \varepsilon)] \quad (2)$$

where ε = 0.05 prevents division by zero. Publication figures additionally require at least 100 occurrences in each period, a

conservative threshold that keeps rare spelling errors from dominating the ranking.

D. Equal-Size Vocabulary Growth

To separate corpus-size effects from lexical accumulation, 459,980 earlier sentences were drawn by deterministic reservoir sampling and compared with all 459,980 later sentences. After deterministic shuffling, cumulative vocabulary was measured at 25,000-sentence checkpoints:

$$V_p(n) = | \cup_{i=1..n} T(d_i) | \quad (3)$$

where $T(d_i)$ is the set of normalized Arabic tokens in sentence d_i.

E. Lexicon, Topic, and Collocation Analysis

The domain lexicon groups Sudanese and conflict-related terms into categories covering war parties, violence reporting, humanitarian language, conflict locations, revolution terminology, political hate, tribal hate, threats, and profanity. Category frequencies are read descriptively, since the categories differ in lexicon size and may overlap. Topic proportions were estimated separately for the two partitions using a 10-topic Latent Dirichlet Allocation model [28] and compared by Jensen–Shannon divergence.

For seed word w, collocate u, period p, and a symmetric five-token window, the normalized collocation rate is

$$C_p(u \mid w) = 1000 \cdot c_p(u, w; \pm 5) / c_p(w) \quad (4)$$

Function words and verified channel boilerplate are excluded from presentation, while the full raw collocation output remains available for audit.

F. Validation and Sensitivity Rules

Four safeguards were fixed before any output was interpreted. A term must occur at least 100 times in both partitions to enter the main shift ranking. The unfiltered ranking is retained so that collection artefacts stay visible. Substantive figures draw on the supplied domain lexicon rather than on words selected because they support a preferred reading. And collocation profiles are normalized by seed frequency.

The vocabulary-growth comparison uses a fixed reservoir-sampling seed recorded in the analysis source. That makes the curve reproducible but does not eliminate sampling variance, so the curve is read descriptively; repeated seeds with confidence bands are left to a publication extension. For the same reason, lexicon-category totals are not compared as though the categories had equal coverage, because their term inventories differ in size. Algorithm 2 summarizes the source-aware cross-period workflow used to calculate the diagnostic and verified lexical-shift rankings, the sentence-count-matched vocabulary-growth curves, and the normalized collocation profiles.

VIII. CROSS-PERIOD RESULTS

A. Corpus and Token Totals

The four earlier files contain 65,520,282 Arabic tokens under the stated token definition. The later file contains 9,775,522 tokens in total, of which 5,271,512 are Arabic, and 306,132 of its sentences contain Arabic script—consistent with the stricter annotation-pool count and confirming that the complete later file and the filtered Arabic subset serve different purposes.

The unfiltered ranking is retained as a diagnostic of source and platform artefacts, whereas substantive interpretation is restricted to terms that satisfy the minimum-count threshold, occur in both partitions, belong to the verified domain lexicon, and have undergone concordance review.

```
Algorithm 1: Evidence-Preserving Dataset Audit
Input:
    Source file set S = {s₁, s₂, ..., sₘ}
    Expected schema definitions E
Output:
    Audit manifest M
    Validated tables V
    Correction ledger C
 1  M ← ∅; V ← ∅; C ← ∅
 2  for each source file s ∈ S do
 3      assert IsReadOnly(s) = True
 4      R_s ← Records(s)
 5      h_s ← SHA256(s)
        b_s ← ByteSize(s)
        t_s ← ModificationTime(s)
        n_s ← |R_s|
 6      Q_s ← ValidateSchema(R_s, E)
        A_s ← ExtractAnomalies(s, Q_s)
 7      M ← M ∪ {(s, h_s, b_s, t_s, n_s)} ∪ A_s
 8      V_s ← CopyValidRecords(R_s, Q_s)
        V ← V ∪ V_s
 9  end for
10  for each proposed repair
        r = (id, v_old, v_new, reason_r, evidence_r) do
11      if IsDeterministic(r) = True
            or UnanimousAgreement(r) = True then
12          C ← C ∪ {
                (reason_r, id, v_old, v_new, evidence_r)
            }
13          V ← ApplyRepair(
                V, id, v_old, v_new
            )
14      else
15          M ← M ∪ {
```

```
                (id, "unverified_repair", r)
            }
16      end if

17  end for
18  assert HasExpectedKeys(V, E) = True
19  assert ∀k ∈ Keys(E):
        IsUnique(V_k) = True
20  assert SatisfiesConstraints(V, E) = True
21  return M, V, C
```

B. Vocabulary Growth

At the common endpoint of 459,980 sentences, the deterministic earlier sample contains 4,659,816 Arabic tokens and 381,137 cumulative normalized types, while the later partition contains 5,169,217 Arabic tokens and 267,245 types. Fig. 8 shows the full curves.

```
Algorithm 2: Source-Aware Cross-Period Analysis
Input:
    Earlier document set D_0
    Later document set D_1
    Verified domain lexicon K
    Minimum count threshold θ
    Smoothing constant ε
    Reservoir-sampling seed s
Output:
    Unfiltered shift ranking R_raw
    Verified lexical-shift map ΔF
    Vocabulary-growth curve V_cum
    Seed-word collocation matrix C_seed

 1  assert RecordCount(D_0) = DeclaredSize(D_0) and
RecordCount(D_1) = DeclaredSize(D_1)
 2  for each period p ∈ {0, 1} do
 3      T_p ← { ArabicTokens(d) for each document d ∈ D_p }
 4      D̃_p ← { NormalizeTokens(ArabicTokens(d)) for each
document d ∈ D_p }
 5      c_p(w) ← Count(w, D̃_p)
        N_p ← TotalTokenCount(D̃_p)
 6      F_p(w) ← 10^6 × c_p(w) / N_p for each w ∈
Vocabulary(D̃_p)
 7  end for
 8  W ← Vocabulary(D̃_0) ∪ Vocabulary(D̃_1)
 9  for each term w ∈ W do
10      L(w) ← log2 [(F_1(w) + ε) / (F_0(w) + ε)]
11  end for
12  R_raw ← SortDescending({ (w, L(w)) : w ∈ W })
13  W_valid ← { w ∈ Vocabulary(D̃_0) ∩ Vocabulary(D̃_1) : c_0(w)
≥ θ and c_1(w) ≥ θ }
14  ΔF ← SortDescending({ (w, L(w)) : w ∈ W_valid and w ∈
K })
15  D̃'_0 ← ReservoirSample(D̃_0, size = |D̃_1|, seed = s)
16  V_cum ← ComputeVocabularyGrowth(D̃'_0, D̃_1, checkpoints =
25,000)
17  C_seed ← ComputeCollocations(seed_terms = K,
earlier_documents = D̃_0, later_documents = D̃_1, window = ±5,
normalize_per = 1,000)
18  assert ConcordanceReviewed(ΔF) = True
19  return R_raw, ΔF, V_cum, C_seed
```

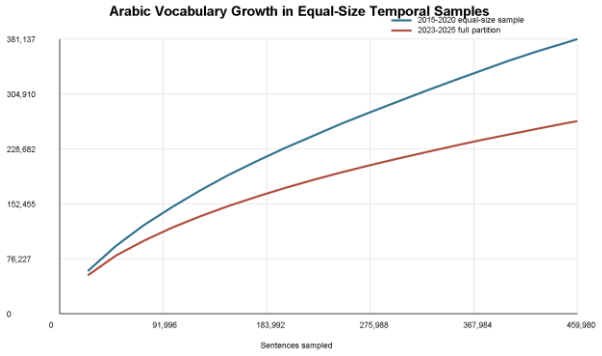

Fig. 8. Vocabulary growth for equal-size temporal samples.

The gap should not be read as evidence that Sudanese Arabic became intrinsically less diverse. The earlier sample mixes Twitter with three Telegram sources, whereas the later partition is drawn from Telegram channels whose reporting conventions repeat. The curve is evidence about the collected corpora and their communicative composition, not about the dialect.

C. Period-Exclusive Vocabulary

Fig. 9 isolates the types that occur in one partition only. The two lists reveal different sources of period-exclusive vocabulary. The earlier-only list is dominated by hamza-bearing spellings of common MSA function words—أن, أو, إلى, أنا, إلا, إن, أي, أنت—which points to a systematic orthographic difference between the partitions rather than to vocabulary that fell out of use. The later-only list mixes platform furniture (تلغرام, تابعونا, قناتنا, نافذتك, العاجلة) with conflict-specific reference (المليشيا, دقلو, المدرعات, الحربي, مقبرة). Reading the second list without the first would overstate lexical innovation; reading either without source awareness would misattribute a channel convention to the dialect.

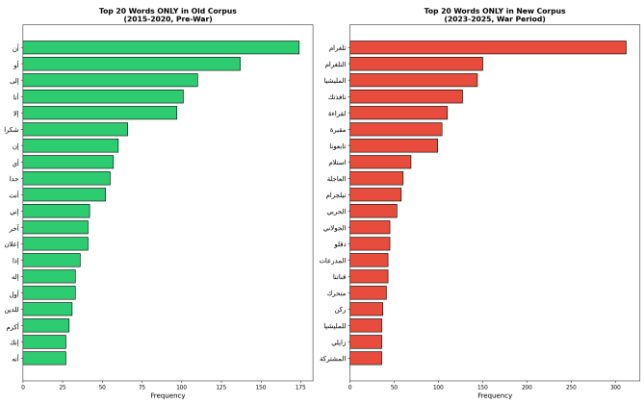

Fig. 9. Types occurring in only one partition: earlier-only (left) and later-only (right). The diagnosis of raw tokens prior to normalization was established

before normalization and is not used as evidence of standardized lexical change.

D. Raw Frequency Shifts and Platform Diagnostics

The unfiltered ranking is led by Telegram and channel-template expressions equivalent to "follow us," "Telegram," and "news." This diagnostic shows that an unfiltered ranking is strongly affected by platform-specific expressions. The diagnostic ranking is retained in the supplementary outputs but is not used as a substantive figure. Fig. 10 shows the unfiltered terms most elevated in 2023–2025 and Fig. 11 those most strongly associated with 2015–2020; the latter shows which forms are overrepresented in the heterogeneous earlier Twitter/Telegram mixture before the domain lexicon is applied.

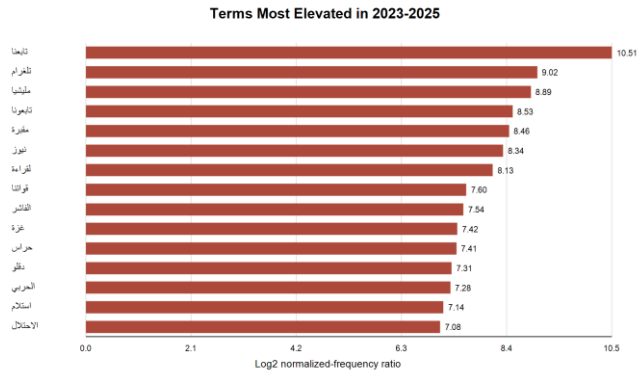

Fig. 10. Unfiltered terms elevated in 2023–2025.

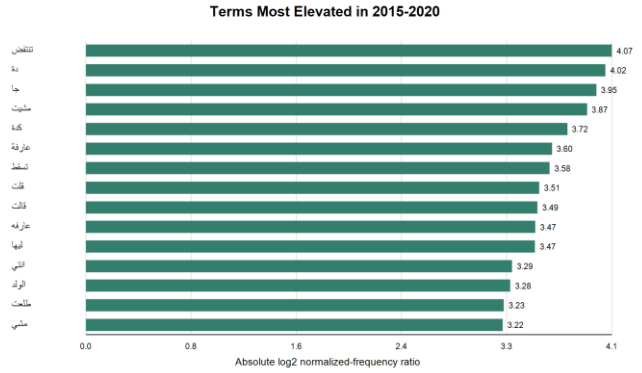

Fig. 11. Unfiltered terms elevated in 2015–2020.

Fig. 12 places the same comparison on a per-10,000-word basis and separates rises from falls. The terms that rise are news-register and platform items (عاجل, الرابط, تابعنا, نيوز, قناة) together with conflict actors (الجيش, السريع); the terms that fall are largely conversational dialect markers (ما, دي, دا, بس, ده, عشان, ياخ) alongside period-specific entities such as حمدوك and الصحة. The later partition is not simply a wartime version of the earlier one—it is a more news-like register with fewer conversational turns, which is a property of the channels sampled as much as of the period.

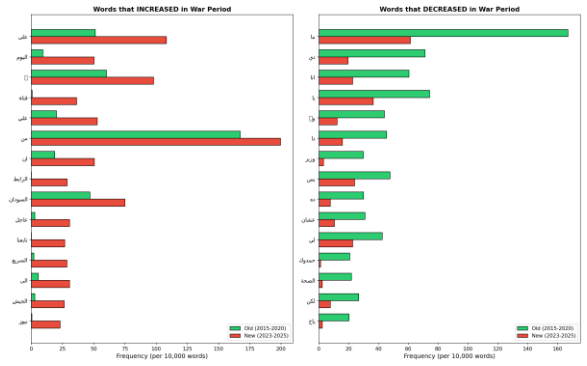

Fig. 12. Words that increased (left) and decreased (right) in the conflict-era partition, per 10,000 words.

E. Conflict-Related Lexical Shifts

Once the conservative count threshold is applied and interpretation is restricted to verified conflict and humanitarian lexicon entries, the later period shows large normalized increases for terms referring to militia, Al-Fashir, fighting, destruction, clashes, siege, Kordofan, Nyala, armed leaders, and conflict locations. Fig. 13 visualizes the strongest verified content shifts.

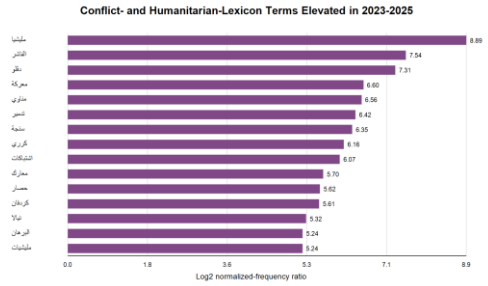

Fig. 13. Conflict-related terms elevated in the 2023–2025 corpus.

Specific normalized examples include مليشيا (militia), which rises from 1.786 to 871.856 occurrences per million Arabic tokens, and الجنجويد (Janjaweed), from 24.160 to 1,916.907. اشتباكات (clashes) rises from 3.281 to 223.276 per million and النازحين (displaced people) from 2.091 to 69.430. These are corpus-frequency findings, not measurements of attitudes toward the groups referenced.

At category level, war-party terminology rises from 397.953 to 6,500.412 occurrences per million among matched unigram entries, violence-reporting terms from 451.830 to 4,542.719, and humanitarian terms from 366.986 to 894.999. Neutral-news markers rise from 288.521 to 4,303.509, which is the clearest single reason to keep subject matter and channel genre analytically apart.

F. Topic-Level Comparison

Topic proportions tell a more restrained story than the lexical rankings. Across ten estimated topics, the Jensen–Shannon divergence between the two partitions is 0.1016 (Fig. 14). Four topics gain share in the later material—0.099 to 0.133, 0.093 to 0.127, 0.107 to 0.127, and 0.064 to 0.091—while three lose it, from 0.129 to 0.085, 0.131 to 0.091, and 0.127 to 0.089. No topic disappears and none approaches dominance.

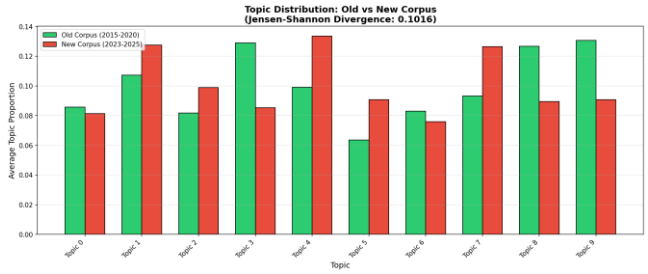

Fig. 14. Topic proportions in the earlier and later partitions (Jensen–Shannon divergence 0.1016).

The observed divergence is modest compared with the magnitude of the lexical-frequency shifts in Section VIII.E are so large. The topics are estimated over unlemmatized dialect text, and their highest-weight terms are largely function words, so the model distributes documents across broadly similar distributions in both periods. Topic proportions are therefore reported here as a coarse stability check rather than as the primary evidence, which is consistent with the recommendation to validate topic output against collocation and concordance material rather than relying on it alone [19].

G. The Changing Context of "War"

The term الحرب occurs 2,464 times in the earlier corpus and 3,278 times in the later corpus—a difference far smaller than its change in company. In the earlier material its frequent collocates correspond to "world/global," "peace," "civil," "third," and "cold," the vocabulary of international and historical reference. In 2023–2025 the normalized collocates are Sudan, beginning/since, Rapid Support, support, Khartoum, and army. Figs. 15 and 16 give the period-specific profiles.

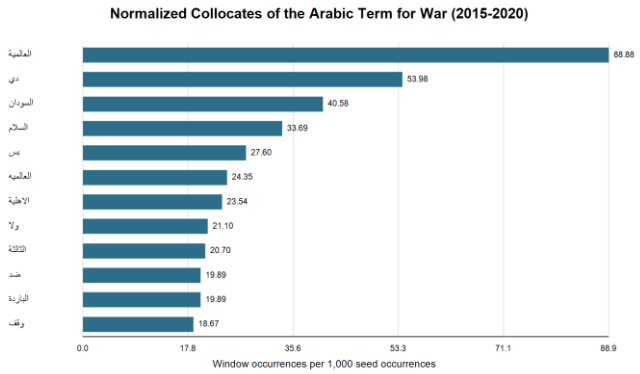

Fig. 15. Normalized collocates of the Arabic term for war, 2015–2020.

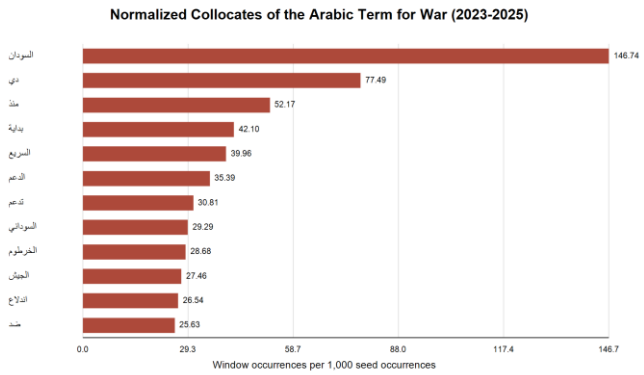

Fig. 16. Normalized collocates of the Arabic term for war, 2023–2025.

Contextual change of this kind is stronger evidence than frequency alone, because the word moves from a heterogeneous international frame toward recurring local actors, places, and temporal framing while its raw count barely moves. The finding is still corpus-specific and is not a causal estimate of population-level language behaviour.

IX. DISCUSSION

A. One Dialect, Three Resource Functions

The three resources answer different layers of the same problem. The large corpus supplies scale and temporal breadth without task labels. The sentiment corpus supplies manually assigned polarity in a service-review domain that is comparatively narrow. The hate-speech corpus reaches politically and socially sensitive language through a scalable hybrid procedure whose labels carry the qualifications that partial human review implies.

Their connection is methodological rather than thematic. Broad corpus evidence identifies frequent forms and changing topics; task-specific datasets establish how those forms behave in evaluative or harmful contexts. Together they support pretraining, domain adaptation, corpus linguistics, sentiment analysis, and content-moderation research without any one dataset having to stand for all Sudanese Arabic.

B. Implications for Low-Resource Dialects

The study supports a resource-first view of low-resource dialect NLP, consistent with recent Mauritanian, Libyan, and Algerian work built on targeted local datasets [7]–[9]. The transferable lesson is not that Sudanese labels can be reused for neighbouring varieties. It is that provenance, orthographic variability, task-specific annotation, and community validation have to be documented at the level of the dialect and the domain.

C. Source-Aware Analysis

The strongest raw temporal shifts in this corpus are channel boilerplate. Without the unfiltered diagnostic in Section VIII.D, those forms could have been presented as emerging conflict vocabulary; without the period-exclusive lists in Section VIII.C, an orthographic difference between partitions could have been presented as vocabulary loss. Treating platform artefacts as evidence about the collection process, while reserving verified semantic categories, normalized rates, and collocations for substantive claims, supports a distinction between source-related artefacts and substantive lexical patterns [19].

D. Interpreting Hate-Related Vocabulary

The domain lexicon contains group names, political labels, profanity, threats, violence descriptions, humanitarian terms, and neutral-news indicators. The presence of such a term is not

itself a hate-speech annotation: a report condemning violence may contain the same actor or group word as an abusive post. Sentence-level labels come from the hybrid pipeline; the temporal lexicon analysis measures only the prevalence of lexical categories.

That distinction is what connects the large corpus to the hate-speech resource without collapsing them. The corpus shows where potentially sensitive vocabulary occurs and how it moves; the labelled corpus supplies contextual class judgments. Future work can combine them to ask whether the proportion of HATE, OFFENSIVE, and NEUTRAL uses of the same term changes by time or source, but that question needs labels across matched temporal samples and is not answered here.

E. Sentiment as a Complementary Domain

Customer reviews extend the suite beyond political social media into praise, dissatisfaction, requests, mixed evaluations, and service-specific terminology. Because the reviews come from eight applications across three sectors, they support domain analysis the Twitter/Telegram corpus cannot. The resource is a task-specific component, not evidence of population-wide emotional change: it may support sentiment modelling, cross-application transfer, and the study of Sudanese evaluative expression, but it cannot estimate national sentiment about the conflict without a separately justified sampling design.

X. DATA AVAILABILITY AND REPRODUCIBILITY

All three resources—the large social-media corpus, the sentiment dataset, and the verified hate-speech corpus—will be released as open-source assets on Masader once documentation and copyright-compliance procedures are complete. Standard license identifiers and repository links will be finalized at public release.

The accompanying workspace separates analysis, derived data, figures, tables, manuscript, and logs. A source manifest records every file behind a numerical result. Corpus scanning is streaming-based, so the 6.76-million-sentence collection is never loaded into memory in full, and derived frequency files retain the observed counts for both periods so that every displayed ratio can be recomputed.

The sentiment correction ledger records the affected identifier, the action taken, the final label, and three evidence files. Hate-speech manifests record physical line counts, parsed records, hashes, and class totals. Figure scripts consume the derived CSV files rather than hard-coded values, which is what keeps prose, tables, and figures from drifting onto different dataset versions.

XI. LIMITATIONS AND ETHICAL CONSIDERATIONS

The corpora come from public digital sources and do not represent every Sudanese region, age group, community, or register. The partitions differ in platform composition—the earlier period combines Twitter and Telegram, the later is Telegram-only—so the analysis reports corpus change and makes no claim that war caused a quantified change in the speech of Sudanese Arabic users generally.

The sentiment dataset is domain-specific and unevenly distributed across applications. The hate-speech dataset is hybrid annotated, with human review covering a stratified quarter of the intended corpus rather than every record, and hate-related terms occur in reporting, counterspeech, quotation, and group self-reference, so lexicon counts are not label counts.

The temporal design has a gap at 2021–2022 and provides no continuous annual series. It supports a comparison between two supplied periods, not a year-by-year change-point model; references to revolution and war describe the historical positioning of the partitions, and the measured contrast is 2015–2020 against 2023–2025.

Orthographic normalization reduces surface variation but does not fully resolve Arabic clitics, spelling variants, or multiword expressions, so vocabulary counts describe types under the stated rules rather than dictionary lemmas. The evidence in Section VIII.C indicates that hamza-bearing spellings are distributed unevenly across the two partitions, which is a further reason to treat type counts as partition-specific. The deterministic equal-size sample controls sentence count but not platform, author, topic, or Arabic-token count. These boundaries do not invalidate the analysis; they define the level at which it is reproducible.

The data contain sensitive political, ethnic, and conflict-related expression. Published figures aggregate counts and expose no user identities. Any public release should document content warnings, permitted uses, privacy safeguards, and a procedure for reporting harmful or incorrectly included material.

XII. RESOURCE CARDS AND RECOMMENDED USES

A. Large Temporal Corpus Card

Identity. A Sudanese Arabic social-media collection organized into an earlier 2015–2020 partition and a conflict-era 2023–2025 partition, with the physical text line as the record unit.

Motivating need. Sudanese Arabic lacks a general corpus at the scale required for lexical analysis, continued pretraining, and temporal comparison. Small labelled datasets matter for evaluation but cannot supply the breadth of forms needed to study spelling, vocabulary, source variation, or changing public topics.

Composition. 6,755,955 sentences in total; 6,295,975 in the earlier partition across three Telegram files and one Twitter file, and 459,980 Telegram records from 14 channels in the later partition. The source-composition figures make this imbalance visible rather than hiding it behind a single total.

Recommended uses. Continued language-model pretraining under appropriate data governance; tokenizer and vocabulary

studies; source-aware corpus linguistics; discovery of candidate Sudanese expressions; construction of carefully sampled annotation pools; comparison of normalized lexical or collocational patterns.

Uses requiring additional evidence. Annual trend estimation, individual-level behaviour analysis, demographic inference, geographic prevalence estimation, and causal claims about the revolution or the conflict. Each would require timestamps, user controls, matched sources, or demographic metadata beyond what is verified here.

Quality notes. Sentence totals are exactly reproducible. The later file contains 306,132 Arabic-script records plus non-Arabic and metadata-like records, and the complete collection must remain distinguishable from the stricter Arabic analytical subset.

B. Sentiment Corpus Card

Identity. A labelled Sudanese Arabic Google Play review dataset built by the present research team, distinct from SudSenti2, SudSenti3, and the earlier telecom dataset used in related work.

Motivating need. Public social-media corpora provide breadth but rarely reliable sentiment labels. A domain-specific review corpus supplies explicit evaluative text together with application metadata, sector, star score, date, and an assigned polarity.

Composition. 13,293 unique review IDs from eight applications: 7,347 banking, 3,446 telecommunications, 2,500 transport. Positive is the largest class, then Negative, then Neutral. The application distribution is highly unequal and should inform evaluation splits.

Recommended uses. Supervised three-class sentiment analysis; cross-application and cross-sector evaluation; analysis of Sudanese evaluative expression; examination of disagreement between textual labels and star ratings; domain adaptation; held-out-application robustness tests.

Uses requiring additional evidence. Population sentiment estimation, conflict sentiment tracking, author demographic profiling, and claims of general coverage across Sudanese Arabic. The corpus measures language in selected application reviews.

C. Hate-Speech Corpus Card

Identity. A Sudanese Arabic harmful-language dataset in binary and three-class formulations. The annotation collection contains 40,000 sentences; the supplied minimum-length analytical TSV contains 38,619 records.

Motivating need. Generic Arabic moderation datasets miss Sudanese political labels, coded group references, local profanity, and conflict-specific usage. Fully manual labelling of a large sensitive corpus is expensive, and purely lexical labelling cannot resolve reporting, quotation, counterspeech, or implicit targeting.

Composition. 7,168 HATE, 8,474 OFFENSIVE, and 22,977 NEUTRAL records in the three-class subset; 15,642 HARMFUL against 22,977 NEUTRAL under the binary mapping. The selected material covers earlier Twitter and Telegram sources and later Telegram material.

Recommended uses. Harmful-language classification; comparison of binary and three-class formulations; study of hybrid annotation; error analysis separating group-targeted from general offensive language; human-in-the-loop moderation research; evaluation of Sudanese-aware encoders.

Uses requiring additional evidence. Automated punitive moderation, prevalence estimates for hate speech across Sudan, identification of individual users, or any claim that a lexicon match is equivalent to hate. High-impact deployment requires an independently adjudicated evaluation set and stakeholder review.

Quality notes. Two LLM annotations, an auxiliary weak-supervision output, tiered resolution, and stratified native-speaker review provide multiple evidence layers. Because human review covers 10,000 sentences, the corpus is described as hybrid annotated with sample-based human quality control.

D. Cross-Resource Experimental Designs

The suite supports experiments no single component allows. A model may be continued-pretrained on the large corpus and then evaluated separately on sentiment and hate speech, which tests whether broad Sudanese exposure improves task adaptation without treating task labels as pretraining text. Tokenizer fragmentation can be compared across the two partitions, and highly fragmented words checked against errors in the labelled datasets.

A second design concerns domain transfer. Sentiment models can be evaluated on held-out applications or sectors, and harmful-language systems across earlier and later sources. Reporting those settings separately keeps in-domain accuracy from being mistaken for robustness.

A third design joins corpus linguistics to annotation. Shifted terms identified in the large corpus can be sampled for native-speaker concordance review, and their occurrences in the hate-speech resource separated into HATE, OFFENSIVE, and NEUTRAL contexts. That would test whether a word's rising public frequency corresponds to rising harmful usage or simply to more reporting. This paper establishes the resources and descriptive evidence that extension needs.

XIII. CONCLUSION

This paper documents a connected Sudanese Arabic resource suite: a 6,755,955-sentence temporal corpus, a 13,293-review labelled sentiment dataset, and a 40,000-sentence hybrid hate-speech resource with a 38,619-record filtered analytical subset. The contribution lies in their construction, documentation, reconciliation, structural comparison, and cross-period analysis.

he results provide both substantive and methodological findings. Substantively, the conflict-era material carries marked normalized increases in verified terms for armed actors,

military engagement, displacement, humanitarian conditions, and affected locations, and the collocational environment of the word for war shifts from international reference to local actors and places while its raw frequency barely moves. Methodologically, the raw rankings are led by Telegram channel templates and the period-exclusive vocabulary lists are shaped by an orthographic difference between partitions. Normalized counts, conservative thresholds, source awareness, domain-lexicon validation, and collocation evidence are what separate the substantive findings from the artefacts.

Three limitations bound these results. The partitions differ in platform composition, with the earlier period drawing on Twitter and Telegram and the later on Telegram alone. Domains such as medical, legal, and formal educational discourse remain outside the suite. And automated preprocessing and weak supervision, though audited, leave residual noise and platform-specific bias.

Work in progress extends the suite toward Sudanese varieties spoken outside the major urban centres, adds fine-grained annotation for named-entity recognition and dialectal machine translation, and evaluates current large language models on these benchmarks to test cross-lingual transfer and sensitivity to dialectal nuance in conflict-era discourse. The suite gives Sudanese Arabic NLP a documented foundation, and it illustrates a wider point about low-resource dialect research: progress depends on complementary resources, transparent provenance, task-appropriate annotation, and claims that stay proportional to the evidence.

REFERENCES

[1] K. Darwish et al., “A panoramic survey of natural language processing in the Arab world,” *Communications of the ACM*, vol. 64, no. 4, pp. 72–81, 2021, doi: 10.1145/3447735.

[2] Y. Matrane, F. Benabbou, and N. Sael, "A systematic literature review of Arabic dialect sentiment analysis," Journal of King Saud University—Computer and Information Sciences, vol. 35, no. 6, Art. no. 101570, Jun. 2023, doi: 10.1016/j.jksuci.2023.101570.

[3] J. Khan, K. Ahmad, S. K. Jagatheesaperumal, and K.-A. Sohn, “Textual variations in social media text processing applications: Challenges, solutions, and trends,” *Artificial Intelligence Review*, vol. 58, no. 3, art. 89, 2025, doi: 10.1007/s10462-024-11071-z.

[4] H. Bouamor et al., “The MADAR Arabic dialect corpus and lexicon,” in *Proc. LREC*, 2018, pp. 3387–3396.

[5] M. Abdul-Mageed et al., “NADI 2023: The fourth nuanced Arabic dialect identification shared task,” in *Proc. ArabicNLP*, 2023, pp. 600–613, doi: 10.18653/v1/2023.arabicnlp-1.62.

[6] A. Abdelali, H. Mubarak, Y. Samih, S. Hassan, and K. Darwish, “QADI: Arabic dialect identification in the wild,” in *Proc. WANLP*, 2021, pp. 1–10.

[7] M. Chabane, F. Harrag, and K. Shaalan, "Advancing low-resource dialect identification: A hybrid cross-lingual model leveraging CAMeLBERT and FastText for Algerian Arabic," Expert Systems with Applications, vol. 284, Art. no. 127816, Jul. 2025, doi: 10.1016/j.eswa.2025.127816.

[8] M. E. M. Chrif, E. B. M. Mahmoud, O. El Beqqali, and M. F. Nanne, "Improving e-government through sentiment analysis in the context of developing countries: A Mauritania case study," Journal of Theoretical and Applied Information Technology, vol. 104, no. 1, pp. 68–81, Jan. 2026.

[9] M. O. AboSarafa and M. A. Arteimi, "Development of smart voice agent with case study: Libyan voice assistant," Academy Journal for Basic and Applied Sciences, vol. 7, no. 1, Jun. 2025, doi: 10.5281/zenodo.15505226.

[10] M. Mhamed et al., "A deep CNN architecture with novel pooling layer applied to two Sudanese Arabic sentiment data sets," Journal of Information Science, vol. 52, no. 1, pp. 285–306, 2026, doi: 10.1177/01655515231188341.

[11] M. Elgezouli, K. N. Elmadani, and M. Saeed, "SudaBERT: A pre-trained encoder representation for Sudanese Arabic dialect," in Proc. 2021 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE), 2021, pp. 1–4, doi: 10.1109/ICCCEEE49695.2021.9429651.

[12] R. Ismail, M. Omer, M. Tabir, N. Mahadi, and I. Amin, “Sentiment analysis for Arabic dialect using supervised learning,” in Proc. 2018 Int. Conf. Computer, Control, Electrical, and Electronics Engineering (ICCCEEE), Khartoum, Sudan, 2018, pp. 1–6, doi: 10.1109/ICCCEEE.2018.8515862.

[13] M. E. M. Abo, N. A. K. Shah, V. Balakrishnan, M. Kamal, A. Abdelaziz, and K. Haruna, “SSA-SDA: Subjectivity and sentiment analysis of Sudanese Dialect Arabic,” in Proc. 2019 Int. Conf. Computer and Information Sciences (ICCIS), Sakaka, Saudi Arabia, 2019, pp. 1–5, doi: 10.1109/ICCISci.2019.8716466.

[14] S. Abuuznien, Z. Abdelmohsin, E. Abdu, and I. Amin, “Sentiment analysis for Sudanese Arabic dialect using comparative supervised learning approach,” in Proc. 2020 Int. Conf. Computer, Control, Electrical, and Electronics Engineering (ICCCEEE), Khartoum, Sudan, 2021, pp. 1–6, doi: 10.1109/ICCCEEE49695.2021.9429560.

[15] M. Jarrar, F. A. Zaraket, T. Hammouda, D. M. Alavi, and M. Wählisch, "Lisan: Yemeni, Iraqi, Libyan, and Sudanese Arabic dialect corpora with morphological annotations," in Proc. 20th ACS/IEEE Int. Conf. Computer Systems and Applications (AICCSA), 2023, pp. 1–7, doi: 10.1109/AICCSA59173.2023.10479250.

[16] H. M. A. Samil and D. I. Adelani, "Sudanese-Flores: Extending FLORES+ to Sudanese Arabic dialect," in Proc. 7th Workshop on African Natural Language Processing (AfricaNLP 2026), Rabat, Morocco, Mar. 2026, pp. 243–247, doi: 10.18653/v1/2026.africanlp-main.25.

[17] M. El-Haj, "Tarab: A multi-dialect corpus of Arabic lyrics and poetry," in Proc. 2nd Workshop on NLP for Languages Using Arabic Script, Rabat, Morocco, Mar. 2026, pp. 37–46, doi: 10.18653/v1/2026.abjadnlp-1.5.

[18] C. Pereira et al., "UstanceBR: A social media language resource for stance prediction," *Language Resources and Evaluation*, vol. 60, art. 14, 2026, doi: 10.1007/s10579-025-09896-3.

[19] J. Zou and X. Tang, "Discerning diachronic sinsign topic shifts: A case study of UK HIV news," *Applied Corpus Linguistics*, vol. 5, no. 3, art. 100163, 2025, doi: 10.1016/j.acorp.2025.100163.

[20] M. Couto, A. Perez, J. Parapar, and D. E. Losada, "Temporal word embeddings for early detection of psychological disorders on social media," Journal of Healthcare Informatics Research, 2025, doi: 10.1007/s41666-025-00186-9.

[21] R. Bai, D. Lu, S. Ran, E. M. Olson, H. Lamba, A. Cahill, J. Tetreault, and A. Jaimes, "CEHA: A dataset of conflict events in the Horn of Africa," in Proc. 31st Int. Conf. Computational Linguistics (COLING), Abu Dhabi, UAE, Jan. 2025, pp. 1475–1495.

[22] P. Piot, P. Martín-Rodilla, and J. Parapar, "MetaHate: A dataset for unifying efforts on hate speech detection," Proceedings of the International AAAI Conference on Web and Social Media, vol. 18, no. 1, pp. 2025–2039, 2024, doi: 10.1609/icwsm.v18i1.31445.

[23] H. Nghiem and H. Daumé III, "HateCOT: An explanation-enhanced dataset for generalizable offensive speech detection via large language models," in Findings of the Association for Computational Linguistics: EMNLP 2024, pp. 5938–5956, 2024, doi: 10.18653/v1/2024.findings-emnlp.343.

[24] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré, "Snorkel: Rapid training data creation with weak supervision," The VLDB Journal, vol. 29, nos. 2–3, pp. 709–730, 2020, doi: 10.1007/s00778-019-00552-1.

[25] OpenAI, "GPT-4o mini: Advancing cost-efficient intelligence," Jul. 18, 2024. [Online]. Available: https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/ [Accessed: Jul. 29, 2026].

[26] A. Grattafiori et al., "The Llama 3 Herd of Models," arXiv preprint arXiv:2407.21783, 2024.

[27] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs," in Advances in Neural Information Processing Systems, vol. 36, pp. 10088–10115, 2023, doi: 10.52202/075280-0441.

[28] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," Journal of Machine Learning Research, vol. 3, pp. 993–1022, 2003.