Mining the Pre-1948 Palestinian Press: Unsupervised Keyphrase Extraction and Temporal Discourse Analysis from Five Historical Arabic Newspapers

Basel Barakat [1], **Nizam Barakat** [2]
[1] School of Computing, Goldsmiths University of London, New Cross, London, UK
[2] Independent Scholar
b.barakat@gold.ac.uk

Abstract

The Palestinian Arabic-language press of the late Ottoman and British Mandate periods constitutes a rich but computationally under-explored archive for studying the evolution of political, cultural, and social discourse in pre-1948 Palestine. This paper presents an end-to-end pipeline for extracting and analyzing thematic content from five historically significant Palestinian newspapers: *Lisān al-ᶜArab* (لسان العرب), *Al-Bushrā* (البشرى), *Al-Karmil* (الكرمل), *Al-Difāᶜ* (الدفاع), and *Filasṭīn* (فلسطين). We describe (i) the construction of a five-source corpus from scanned newspaper images obtained from archival collections, processed using the Google Cloud Vision OCR (GCV-OCR) API, (ii) the adaptation of KeyBERT with an AraBERT backbone for unsupervised keyphrase extraction, and (iii) a purpose-built Python visualization toolkit that produces keyword-frequency heatmaps, longitudinal trend charts, and ranked bar charts with full Arabic script rendering. Experiments across the five subcorpora show that the pipeline yields topically diverse keyphrases reflecting each newspaper's editorial orientation and its distinct representation of Palestinian native perspectives—from pan-Arab nationalism and anti-colonial resistance to religious and communal affairs. Temporal analysis reveals event-responsive patterns that align with major historical developments, including the 1936–1939 Arab Revolt and the intensification of sovereignty discourse toward 1948. The pipeline, data format specifications, and visualization code are provided as supplementary material.

Keywords: Arabic NLP, keyphrase extraction, historical newspapers, Palestinian press, corpus linguistics, KeyBERT, AraBERT

1. Introduction

The Arabic-language press that flourished in Palestine during the late Ottoman period and the British Mandate era (1920–1948) constitutes an invaluable primary source for historians, political scientists, and linguists. These newspapers documented the social, cultural, and political transformations of a formative period—including the rise of national movements, colonial administration, intercommunal tensions, and debates over statehood—in the everyday register of journalistic prose. Yet the vast majority of this material remains locked in physical archives and private collections, inaccessible to computational analysis.

Advances in optical character recognition (OCR) and neural language models now make it feasible to digitize and semantically analyze such corpora at scale; however, Arabic poses particular challenges: its right-to-left script, rich agglutinative morphology, and orthographic variation—especially in early twentieth-century printing—complicate both digitization quality and downstream NLP processing.

This paper addresses these challenges through an integrated four-stage pipeline (Figure 1) spanning document acquisition, OCR-based transcription, transformer-based keyphrase extraction, and temporal visualization. We apply this pipeline to five major Palestinian newspapers published before 1948:

- ***Lisān al-ᶜArab*** (لسان العرب) — a pan-Arab nationalist newspaper reflecting broad Arab political concerns (Wikipedia contributors, 2025b).
- ***Al-Bushrā*** (البشرى) — a publication with social, cultural, and community-oriented coverage.
- ***Al-Karmil*** (الكرمل) — published in Haifa (1908–1944) by Najib Nassar, one of the longest-running and most influential Palestinian newspapers, known for its political commentary and anti-colonial stance (Wikipedia contributors, 2025a).
- ***Al-Difāᶜ*** (الدفاع) — published in Jaffa (1934–1948), a leading daily that covered political developments extensively during the final Mandate years (Wikipedia contributors, 2026a).
- ***Filasṭīn*** (فلسطين) — published in Jaffa (1911–1948), by ᶜĪsā al-ᶜĪsā and Yūsuf al-ᶜĪsā, one of the most prominent and widely circulated Palestinian newspapers, known for its sustained political reporting across both Ottoman and Mandate periods (Wikipedia contributors, 2026b).

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 23–32
11 May 2026.

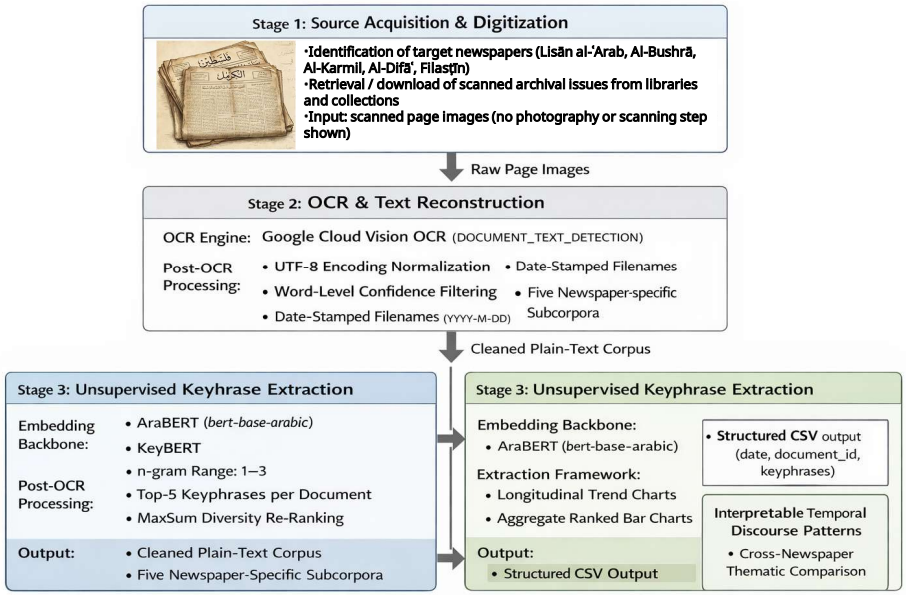

Figure 1: End-to-end pipeline overview. Newspapers from five Palestinian titles are collected and transcribed via GCV-OCR. Post-OCR processing normalizes encoding, filters low-confidence words, and segments pages into date-stamped document units. Plain-text files are processed by the KeyBERT/AraBERT extraction stage, producing a structured CSV per newspaper. The CSVs are utilized by the Arabic-aware visualization toolkit to generate temporal and comparative charts.

Our primary contributions are:

1. **Corpus construction.** We assemble five collections of historical Palestinian newspapers from scanned archival images, process them using the GCV-OCR API, and describe the acquisition workflow, quality filtering criteria, and resulting corpus organization spanning the late Ottoman through British Mandate periods.

2. **Adapted keyphrase extraction.** We apply KeyBERT (Grootendorst, 2020) with AraBERT (Antoun et al., 2020) as the backbone encoder for unsupervised keyphrase extraction from historical Arabic text, with a formal account of the scoring function.

3. **Cross-newspaper comparative analysis.** We compare keyphrase distributions across the five newspapers, revealing distinct editorial orientations and shared patterns of event-driven discourse.

4. **Arabic-aware temporal visualization.** We develop a standalone Python visualization toolkit producing heatmaps, trend lines, and bar charts with correct Arabic script rendering via character reshaping and BiDi reordering.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the corpus and OCR process. Section 4 details the keyphrase extraction methodology. Section 5 presents the visualization framework. Section 6 reports experimental results. Section 7 concludes.

2. Related Work

2.1. Digitization of Historical Arabic Press

The digitization of historical newspapers has received sustained attention in Western digital humanities (Smith et al., 2015), but work on Arabic archival collections remains comparatively sparse. Prior efforts have focused primarily on handwritten manuscript recognition (Al-Khateeb et al., 2011) or on contemporary newswire collections such as the Arabic Gigaword (Graff and Maamouri, 2006). Projects such as the Institute for Palestine Studies digital archive have preserved Palestinian textual heritage, but have primarily produced human-curated reference works rather than machine-readable corpora amenable to NLP processing.

Notable preservation initiatives include the digitization of the historical periodical collection at the Al-Aqsa Mosque Library in East Jerusalem (Matusiak and Abu Harb, 2011), which created archival-quality digital copies of 24 newspaper and magazine titles from the British Mandate period, and the Jrayed digital archive at the National Library of Is-

rael, which provides image-level access to Arabic periodicals from Ottoman and Mandatory Palestine. More recently, El Ganadi et al. (2025) proposed a metadata-driven framework for Arabic library digitization that addresses challenges specific to Arabic-script processing, including calligraphy, diacritics, and ligatures. These projects have made significant strides in preservation and access but have not incorporated NLP-based content analysis pipelines of the kind we present here.

Methodologically, the closest precedent to our work is Grallert (2021), who applied digital history methods—including social network analysis and stylometric authorship attribution—to a 2.65-million-word corpus of four Arabic journals from the late Ottoman Eastern Mediterranean (1906–1918). That study highlighted both the potential and the infrastructure challenges of computational approaches to Arabic periodicals, particularly regarding corpus construction and the socio-technical biases of digitization practices rooted in the Global North. Our work complements this line of research by extending transformer-based keyphrase extraction to Palestinian newspapers and by introducing temporal visualization methods tailored to right-to-left script.

The Palestinian press of the late Ottoman and Mandate periods is of particular scholarly interest because it documents the emergence of national consciousness and debates over British colonial governance in real time. Newspapers such as *Al-Karmil*, founded by Najib Nassar in 1908, *Filasṭīn*, founded by the ʿĪsā brothers in 1911, and *Al-Difāʿ*, which became one of the most widely circulated dailies in the 1930s, are frequently cited in historical scholarship (Ayalon, 1995; Khalidi, 1997) but have rarely been subjected to systematic computational analysis. To our knowledge, the present work is among the first to apply a complete OCR-to-NLP pipeline to archival scanned Palestinian newspaper collections spanning multiple titles and decades.

2.2. Keyphrase Extraction

Keyphrase extraction methods divide broadly into unsupervised statistical approaches—such as TF-IDF and YAKE (Campos et al., 2020)—and embedding-based methods that leverage pre-trained language models. KeyBERT (Grootendorst, 2020) belongs to the latter family, ranking candidate n-grams by their cosine similarity to the document embedding produced by a sentence encoder. While embedding-based methods have shown strong performance on English news corpora, their adaptation to Arabic remains limited. Antoun et al. (2020) introduced AraBERT, a BERT-base model pre-trained exclusively on Arabic text, which provides substantially richer morphological representations than general-purpose multilingual encoders such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020).

Recent work on Arabic keyphrase extraction has explored both supervised and unsupervised settings but has overwhelmingly targeted Modern Standard Arabic from contemporary sources. The application of embedding-based keyphrase extraction to historical Arabic—with its distinct orthographic conventions, vocabulary, and print quality—remains largely unaddressed.

2.3. Temporal Analysis of News Corpora

Temporal keyword analysis has been used to trace the evolution of political discourse in newspaper archives (Lansdall-Welfare et al., 2017). Heatmap visualizations of keyword frequency over time have proven effective for communicating thematic shifts to non-technical audiences in digital humanities contexts (Hamilton et al., 2016). Our visualization tool extends this tradition to right-to-left Arabic script through Unicode reshaping and enables cross-newspaper comparison of discourse patterns.

2.4. NLP for Palestinian Arabic

While substantial NLP resources exist for Modern Standard Arabic (MSA), computational work targeting the specific varieties and registers of Palestinian remains limited. The historical press adds a further temporal dimension: early twentieth-century journalistic Arabic differs from contemporary MSA in vocabulary, sentence structure, and orthographic norms. This creates a domain shift between AraBERT's pre-training distribution and our target corpus, as discussed in the Limitations section.

3. Data Collection and Corpus Construction

Our corpus is derived from five historically significant Palestinian Arabic-language newspapers, all published before 1948 (Jrayed). Together, the five collections span the late Ottoman period through the end of the British Mandate and represent a substantial archive of early twentieth-century Palestinian journalistic prose, encompassing nationalist, cultural, and political registers.

3.1. Digitization via GCV-OCR

Arabic OCR remains a challenging task due to the cursive nature of the script, the presence of diacritics, varying font styles, and—for historical documents—paper degradation and printing irregularities, as shown in Figure 2 (Faizullah et al.,

Figure 2: Filasṭīn Newspaper, issue date 2 November 1932 (Palestine Square, 1932)

2025; Al Ghamdi, 2022). Recent surveys indicate that deep learning architectures, particularly CNN-LSTM-CTC pipelines, have advanced handwritten Arabic recognition (Wagaa and Kallel, 2023), though printed historical Arabic at scale remains underexplored.

Scanned newspaper page images were submitted to the GCV-OCR API (Google Cloud, 2023) using the `DOCUMENT_TEXT_DETECTION` feature, which is optimized for dense, multi-column document layouts and returns a full-text annotation with word-level confidence scores.

Post-OCR processing comprised three steps:

1. **Encoding normalization.** All output was re-encoded in UTF-8 to ensure consistent handling of Arabic diacritics, Hamza variants, and punctuation marks.

2. **Confidence filtering.** Words with confidence below 0.6 were flagged; pages on which more than 30% of words fell below this threshold were excluded to limit noise propagation.

3. **Document segmentation.** Each page is treated as a single document unit, with the publication date encoded in the filename as `YYYY-MM-DD`. Article-level segmentation using layout analysis is left for future work.

3.2. Corpus Statistics

Table 1 summarizes the five subcorpora.

4. Keyphrase Extraction Methodology

We use *keyphrase* throughout this section to denote candidate n-grams extracted by the system (including multi-word expressions), reserving *keyword* for the general discussion of terms and the visualization tool, which operates at token level for frequency counts.

4.1. Extraction Framework

Our system builds on KeyBERT (Grootendorst, 2020), which frames keyphrase extraction as a semantic similarity problem: candidate n-grams are ranked by their cosine similarity to the global document embedding produced by a pre-trained encoder. Formally, given a document D and a set of candidate keyphrases $\{c_n\}$, the top-k keyphrase set is defined as:

$$\text{Keyphrases}(D) = \underset{c_n}{\text{argmax}} \ \ \text{cos-sim}(\mathbf{e}(c_n), \mathbf{e}(D)), \tag{1}$$

where $\mathbf{e}(\cdot)$ denotes the encoder embedding function. Because the formulation requires no labeled training data, it is well-suited to the low-resource setting of historical Arabic news.

4.2. Arabic Language Model Selection

A core design choice is the backbone encoder. General-purpose multilingual models such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) encode Arabic as one of over 100 languages, potentially diluting language-specific morphological patterns. We therefore employ AraBERT (`asafaya/bert-base-arabic`) (Antoun et al., 2020), a BERT-base architecture pre-trained exclusively on Arabic news articles, Wikipedia pages, and web text. AraBERT achieves strong results on a range of Arabic NLP benchmarks—including named entity recognition and sentiment analysis—and its WordPiece vocabulary is trained entirely on Arabic, giving finer subword segmentation than multilingual alternatives. The model is loaded via the Hugging Face Transformers library (Wolf et al., 2020) and wrapped by the KeyBERT interface as a sentence encoder.

4.3. Data Preparation

All input files are stored in UTF-8 encoding. No explicit preprocessing (diacritization removal, stemming, or stopword filtering) is applied at the file

Table 1: Corpus statistics (post-filtering) for the five Palestinian newspaper subcorpora.

Property	*Lisān al-ᶜArab*	*Al-Bushrā*	*Al-Karmil*	*Al-Difāᶜ*	*Filasṭīn*
Publication city	Jerusalem	Haifa	Haifa	Jaffa	Jaffa
Active years	1921–1925	1935–1950	1908–1942	1934–1948	1911–1948
Issues digitized	543	103	1,238	3,874	6,692
Pages digitized	2,022	3,246	8,050	20,505	38,001

level: all tokenization decisions are delegated to the AraBERT tokenizer.

4.4. Keyphrase Extraction Configuration

Extraction hyperparameters are held constant across all documents, as detailed in Table 2.

The Max-Sum algorithm (Carbonell and Goldstein, 1998) greedily selects keyphrases that maximize document-similarity scores while penalizing high pairwise cosine similarity between selected candidates, yielding a topically diverse output well-suited to multi-topic news documents.

4.5. Implementation Details

All experiments are implemented in Python 3 using `keybert`, `transformers` (Wolf et al., 2020), and `pandas`. The pipeline is fully deterministic given fixed model weights, as the Max-Sum algorithm and cosine similarity computations contain no stochastic components.

5. Visualization Framework

Extraction output is stored as a three-column CSV with fields `date`, `filename`, and `keywords`, where the latter contains the top-5 keyphrases for each document. The visualization toolkit ingests this format and produces three complementary chart types described below.

5.1. Data Filtering

During loading, the tool discards rows with unparseable dates and retains only keyphrases containing at least one character in the Arabic Unicode block (U+0600–U+06FF), eliminating OCR artifacts and non-Arabic tokens. The top N keyphrases by overall frequency are selected for plotting (default $N{=}15$; $N{=}8$ for trend lines to preserve readability).

5.2. Chart Types

Three chart types are generated for each subcorpus:

Keyword $\times$ Year Heatmap (Figure 4). A heatmap shows the frequency of the top-N keyphrases (rows) across publication years (columns) using a `YlOrRd` color scale. This view enables rapid identification of temporal surges and declines in specific terminology.

Longitudinal Trend Lines (Figure 5). Line charts plot annual keyphrase frequency per newspaper, one colored line per keyphrase. This view supports fine-grained comparison of relative thematic prominence over time.

Overall Ranked Bar Chart (Figure 3). Horizontal bar charts show the aggregate frequency of the top-N keyphrases across the full time span for each newspaper, providing a domain summary independent of temporal structure.

6. Results and Analysis

6.1. Extracted Keyphrases: Qualitative Overview

The overall keyword rankings (Figure 3) show that each newspaper's extracted vocabulary reflects its distinct editorial character.

Lisān al-ᶜArab yields keyphrases dominated by references to the newspaper itself and to broader Arab political discourse, with the highest-frequency terms including newspaper-name collocations and references to cities such as Jerusalem and Haifa. *Al-Bushrā* surfaces religious and cultural terminology—including Islamic references and community-related terms—consistent with its orientation toward social affairs. *Al-Difāᶜ* produces keyphrases concentrated around Quranic references, personal names associated with political and religious figures, and editorial content, reflecting its status as a politically engaged daily. *Filasṭīn* generates the highest-frequency keyphrases in the corpus, dominated by newspaper-name collocations, references to law and governance, and administrative terminology, consistent with its role as a primary venue for sustained political reporting. *Al-Karmil* yields keyphrases related to education, community organizing, and political institutions, reflecting its editorial focus on political engagement.

The trigram range setting captures compound

Table 2: KeyBERT hyperparameter configuration used in all experiments.

Parameter	Value	Rationale
`keyphrase_ngram_range`	`(1, 3)`	Captures unigrams, bigrams, and trigrams to accommodate multi-word expressions common in Arabic news discourse.
`top_n`	`5`	Extracts the five highest-scoring keyphrases per document.
`stop_words`	`None`	Arabic function words are suppressed implicitly through low cosine similarity to document-level embeddings.
`use_maxsum`	`True`	Applies the Max-Sum algorithm (Carbonell and Goldstein, 1998) to maximize document similarity while minimizing inter-candidate redundancy.
`diversity`	`0.7`	Controls the balance between topical breadth and semantic precision in keyphrase selection.

terminology—multi-word political expressions, institutional references, and editorial collocations—that unigram-only approaches would miss.

6.2. Temporal Patterns

The heatmaps (Figure 4) reveal structured temporal variation across all five newspapers. Several patterns emerge.

Lisān al-ᶜArab, covering the early 1920s, shows concentrated keyphrase activity in 1922–1923, with a sharp decline thereafter, consistent with the newspaper's relatively brief publication window. *Al-Bushrā*, active in the 1940s and early 1950s, displays sparse but focused keyword distributions, with Islamic and cultural terminology appearing in isolated temporal clusters. *Al-Difāᶜ* (1934–1948) exhibits a clear temporal progression: keyphrase frequency intensifies in the mid-1930s and peaks during 1944–1947, coinciding with the post-war political crisis and the final years of the Mandate. *Filasṭīn* shows the most sustained temporal pattern, spanning from the early 1910s through 1941. Keyphrase frequency increases markedly in the 1930s—the period of the Arab Revolt and peak political mobilization—with a visible concentration of high-frequency terms between 1935 and 1940. *Al-Karmil* (1908–1944 in the broader record, though the digitized material here spans the 1920s–1930s) shows temporal variation consistent with the rise of nationalist discourse, with keyphrase activity concentrated in the late 1920s and early 1930s.

The trend lines (Figure 5) provide a complementary view of the temporal dynamics. In *Lisān al-ᶜArab*, the dominant keyphrases show a sharp peak around 1922 followed by rapid decline. *Al-Difāᶜ* exhibits a more complex pattern: certain keyphrases peak in the mid-1930s (around the outbreak of the Arab Revolt in 1936), decline during 1938–1940, and then resurge during 1942–1947. *Filasṭīn* shows the most dramatic temporal variation, with several keyphrases surging sharply in the mid-to-late 1930s—reaching frequencies above 80 per year—before declining in the early 1940s. *Al-Karmil* displays a distinctive double-peak pattern in the early 1920s and late 1920s, with keyphrase activity tapering off in the 1930s as the digitized coverage thins.

6.3. Cross-Newspaper Comparison

A central finding is the emergence of shared keyphrase patterns alongside newspaper-specific vocabulary. Several observations stand out from the comparative analysis of Figure 3.

First, newspaper-name collocations appear as top keyphrases in multiple titles (e.g., "newspaper Lisān al-ᶜArab" in the Arab subcorpus, "newspaper Palestine" in the Filasṭīn subcorpus). This reflects a characteristic feature of early twentieth-century Arabic newspaper typography, in which the newspaper name appeared frequently throughout each issue in mastheads, headers, and self-referential editorial content.

Second, *Filasṭīn* and *Al-Difāᶜ*—the two highest-volume subcorpora and the titles with the most sustained political focus—share a concentration of governance- and conflict-related terminology, though the specific keyphrases differ due to differences in editorial style and temporal coverage.

Third, *Al-Bushrā* stands apart from the other four titles, with its top keyphrases dominated by Islamic religious terminology and references to community figures. This confirms its distinct editorial orientation toward social and cultural affairs rather than overtly political reporting.

Fourth, *Al-Karmil* occupies an intermediate posi-

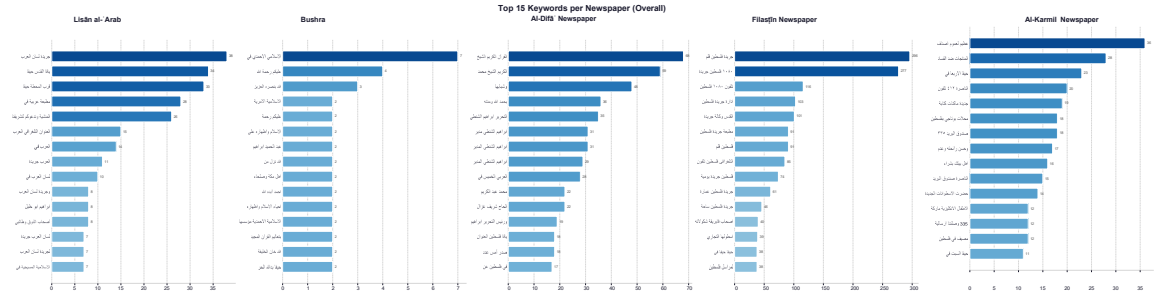

Figure 3: Overall top-15 keyphrases by total frequency for each of the five newspaper subcorpora. Each panel shows the ranked keyphrases extracted by KeyBERT with AraBERT from the respective subcorpus. *Filasṭīn* exhibits the highest absolute frequencies, consistent with its larger digitized volume. All labels are rendered in Arabic script.

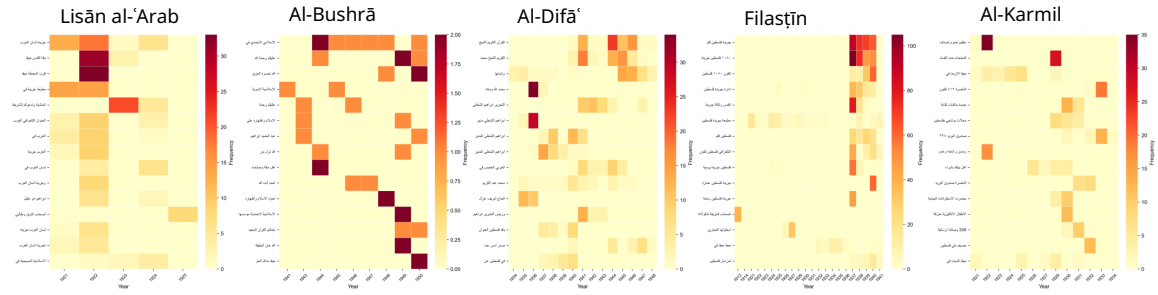

Figure 4: Keyword × Year frequency heatmaps for all five newspaper subcorpora. Rows represent the top-15 keyphrases; columns represent publication years. Color intensity (YlOrRd scale) encodes frequency. Clear temporal clustering is visible in all panels, with *Filasṭīn* and *Al-Difāᶜ* showing the most sustained high-frequency activity.

tion, combining references to political institutions with cultural and educational vocabulary, consistent with its role as a forum for both political commentary and community organizing.

6.4. OCR Noise and Robustness

A qualitative review of extracted keyphrases shows that OCR errors introduce noise primarily as partial words or incorrectly joined characters. The pipeline shows inherent robustness: malformed tokens receive low cosine similarity to the document embedding and are rarely selected among the top-5 candidates. The Arabic character filter applied during visualization provides an additional denoising pass. Nevertheless, pages with borderline OCR confidence remain a latent noise source; image binarization and language-model-based OCR post-correction are priorities for future work.

7. Conclusion

We have presented an end-to-end pipeline for digitizing, extracting, and visualizing thematic content from five major pre-1948 Palestinian Arabic newspapers. By combining GCV-OCR with AraBERT-backed KeyBERT extraction and a purpose-built Arabic-aware visualization toolkit, the system makes historical Palestinian press archives amenable to systematic content analysis without requiring labeled data or language-specific preprocessing beyond UTF-8 normalization.

The five-newspaper design enables both within-title temporal analysis and cross-newspaper comparison, revealing a shared core of political vocabulary alongside distinct editorial orientations. Temporally, all five subcorpora exhibit coherent, event-responsive patterns consistent with known historical developments: the Arab Revolt of 1936–1939, wartime reporting, and the post-war political crisis each leave visible imprints.

Acknowledgements and Ethical Considerations

The authors thank the staff who archived and facilitated access to the newspaper collections. All material consists of early-to-mid twentieth-century archival documents treated for non-commercial academic research. The study aims to preserve Palestinian cultural heritage through computational methods. The terminology and perspectives represented in these historical sources reflect

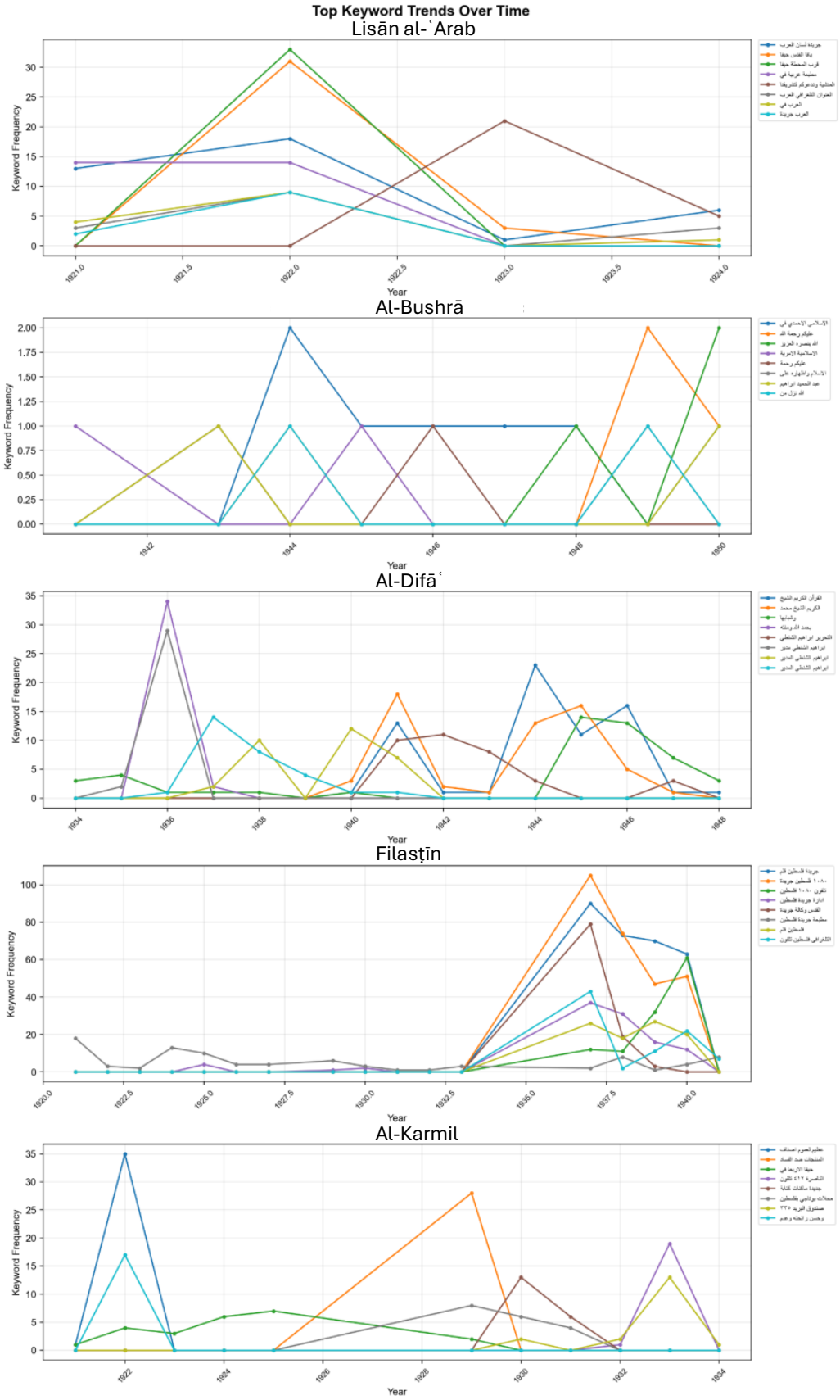

Figure 5: Longitudinal frequency trends for the top-8 keyphrases in each of the five newspaper subcorpora. Each colored line tracks one keyphrase over time. *Filasṭīn* shows the most pronounced temporal spikes, peaking in the mid-to-late 1930s during the Arab Revolt period.

the context of their time and do not necessarily represent the views of the authors.

Limitations

OCR quality varies across newspapers depending on source condition, print quality, and layout complexity; confidence-based filtering mitigates but cannot eliminate noise, especially for pre-1920 material. AraBERT's pre-training on MSA may not fully capture early twentieth-century orthographic conventions, posing a domain-shift challenge common to all transformer-based approaches to historical text. Additionally, newspaper-name collocations appear among top-ranked keyphrases (Figure 3) as an artifact of masthead repetition; post-hoc filtering is left for future work.

Bibliographical References

Mohammed A. Al Ghamdi. 2022. A novel approach to printed Arabic optical character recognition. *Arabian Journal for Science and Engineering*, 47:2219–2237.

Jawad H Al-Khateeb, Jinchang Ren, Jianmin Jiang, and Husni Al-Muhtaseb. 2011. Offline handwritten Arabic cursive text recognition: A survey. *Pattern Recognition*, 44(3):555–567.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools*, pages 9–15.

Ami Ayalon. 1995. *The Press in the Arab Middle East: A History*. Oxford University Press.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. YAKE! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Amina El Ganadi, Luca Gagliardelli, Sania Aftar, and Federico Ruozzi. 2025. Digital Maktaba project: Proposing a metadata-driven framework for Arabic library digitization. In *Proceedings of the 21st Conference on Information and Research Sciences Connecting to Digital and Library Science (IRCDL 2025)*, volume 3937 of *CEUR Workshop Proceedings*.

Safiullah Faizullah, Muhammad Shahid Ayub, Sajid Hussain, and Muhammad Arif Khan. 2025. Advancements and challenges in Arabic optical character recognition: A comprehensive survey. *ACM Computing Surveys*, 58(4):1–37.

Google Cloud. 2023. Cloud Vision API. https://cloud.google.com/vision.

David Graff and Mohamed Maamouri. 2006. Arabic Gigaword, third edition. Technical report, Linguistic Data Consortium.

Till Grallert. 2021. Catch me if you can! Approaching the Arabic press of the late Ottoman Eastern Mediterranean through digital history. *Geschichte und Gesellschaft*, 47(1):58–89.

Maarten Grootendorst. 2020. KeyBERT: Minimal keyword extraction with BERT. https://github.com/MaartenGr/KeyBERT.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1489–1501.

Jrayed. Jrayed: Arabic newspaper archive of ottoman and mandatory palestine. Accessed: 2026-03-30.

Rashid Khalidi. 1997. *Palestinian Identity: The Construction of Modern National Consciousness*. Columbia University Press.

Thomas Lansdall-Welfare, Saatviga Sudhahar, James Thompson, Justin Lewis, and Nello Cristianini. 2017. Content analysis of 150 years of

British periodicals. *Proceedings of the National Academy of Sciences*, 114(4):E457–E465.

Krystyna K. Matusiak and Qasem Abu Harb. 2011. Digitizing the historical periodical collection at the Al-Aqsa Mosque library in East Jerusalem. In Hartmut Walravens, editor, *Newspapers: Legal Deposit and Research in the Digital Era*, pages 271–290. De Gruyter Saur, Berlin.

Palestine Square. 1932. Caricature: Balfour and the woes his ill-fated promise brought to palestine. Accessed: 2026-03-30.

David A Smith, Ryan Cordell, and Abby Mullen. 2015. Computational methods for uncovering reprinted texts in antebellum newspapers. *American Literary History*, 27(3):E1–E15.

Nabil Wagaa and Hassene Kallel. 2023. Analysis of recent deep learning techniques for Arabic handwritten-text OCR and post-OCR correction. *Applied Sciences*, 13(13):7568.

Wikipedia contributors. 2025a. Al-karmil (newspaper) — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Al-Karmil_(newspaper). [Online; accessed 27-February-2026].

Wikipedia contributors. 2025b. Lisan al arab (newspaper) — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Lisan_al-Arab_(newspaper). [Online; accessed 27-February-2026].

Wikipedia contributors. 2026a. Al-difa' — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Al-Difa%27. [Online; accessed 27-February-2026].

Wikipedia contributors. 2026b. Falastin — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Falastin. [Online; accessed 27-February-2026].

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaudhary, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.