Humanities and Social Sciences Communications

Article in Press

https://doi.org/10.1057/s41599-026-06682-6

Transforming the Northwest frontier: development discourse in Republican China through computational analysis of the historical press

Received: 27 May 2025

Accepted: 2 February 2026

Cite this article as: Ren, T. Transforming the Northwest frontier: development discourse in Republican China through computational analysis of the historical press. *Humanit Soc Sci Commun* (2026). https://doi.org/10.1057/s41599-026-06682-6

Tao Ren

Transforming the Northwest Frontier: Development Discourse in Republican China through Computational Analysis of the Historical Press

Abstract

This study examines the discourse surrounding Northwest China's development during the Republican era (1911–1949). Drawing on 5,461 newspaper and periodical articles from the *Quan Guo Bao Kan Suo Yin* (CNBKSY) and the *Shenbao* database, we developed a multi-stage text extraction workflow leveraging Google Gemini to convert complex historical scans into machine-readable text. After standardising historical character forms, we applied structural topic modelling (STM) to recover 26 coherent themes covering infrastructure, resource extraction, state governance, ethnic relations, and cultural mobilisation. Topic-correlation analysis reveals three higher-order clusters—infrastructure and resources, governance and industry, and cultural-educational development—integrated by overarching concerns for strategic geography and state-led economic planning. Temporal analysis shows dramatic inflections: discourse initially gained momentum through regional development initiatives led by warlords in the 1920s, before the 1931 Manchurian Incident amplified a security-centred rhetoric that fully shifted the focus from exploratory surveys to urgent state-led nation-building after the 1937 Japanese invasion—a focus that gradually ebbed post-1945 as wartime imperatives faded. These patterns illustrate how national crises transformed the Northwest from a peripheral frontier into a strategic heartland in Republican-era Chinese media and policy discourse. By leveraging a large-scale corpus and cutting-edge computational methods, this study moves beyond traditional historiography to provide the most comprehensive analysis yet of how Republican China envisioned, prioritised, and rhetorically shaped its interior frontier during a pivotal period of modern Chinese state-building.

1. Introduction

In the early twentieth century, China's Northwest frontier transformed from a remote periphery into a focal point of national imagination and planning. As mass media expanded rapidly during the Republican era (1911–1949), the notion of 'developing the Northwest' (*Kaifa Xibei* 開發西北) rose to prominence in newspapers and periodicals nationwide. This discourse gained particular urgency after Japan's 1931 invasion of Manchuria, which heightened the interior's perceived strategic importance (Dagongbao, 1932; Wang, 1943, p. 1; Zhang, 1934b, p. 2). Government officials, intellectuals, and journalists increasingly cast the Northwest as both a land of potential riches and a bulwark for national defence, advocating ambitious programs of infrastructural construction, resource exploitation, and administrative integration to incorporate this frontier into the Chinese nation-state.

One contemporary observer noted in 1932 the formation of over thirty new groups–from academic societies to state-sponsored 'reconstruction' teams–alongside a proliferation of periodicals such as *New Northwest*, *Northwest Monthly*, and *Border Reclamation*, all emerging in the wake of the Manchurian Incident (Piao, 1932). A bibliographic survey confirms this trend: whereas only five journals with 'Northwest' in their title existed before 1930, between 1931 and 1945 that number jumped to seventy, before falling off after World War II as wartime urgency waned (Hu, 1985). These publications sought to popularise knowledge about the region's geography, resources, and peoples in order to rally public support for Northwest development (Chuangkanci, 1934; Fakanci, 1936; Yu, 1929). As one contemporary remarked in 1936, after the fall of Northeast China to Japan, 'whether in political or economic terms, the value of the Northwest has greatly increased; thus, issues concerning the Northwest—be it national defence,

transportation, or industry—have drawn the attention of people throughout the country' (You, 1936).

The issue of Northwest frontier (*bianjiang* 邊疆) in China has deep historical roots. Imperial China's discourse referred to territory as *jiangyu* 疆域, *shanhe* 山河, or *jin'ou* 金甌 (Liu, 2011, pp. 120–121; Matten, 2016, pp. 126–128). Unlike modern conceptions, these terms were neither absolute nor static; they shifted with the reach of imperial authority, especially at the margins (Ge, 2011, pp. 25–27; Zhao, 2004, p. 37). This fluidity was particularly pronounced in the vast Northwest frontier, which Mancall (1968, pp. 73–74) characterised as a 'northwestern crescent' forming a distinct ecological and socio-economic zone, fundamentally different from the sedentary agricultural societies of China proper. In essence, frontiers functioned to protect the empire—a role captured by the official term *fan* 藩 (vassal), denoting 'a hedge, a boundary, a frontier; to screen, to protect' (Fairbank, 1968, p. 9; Qinggaozong, 1935, p. 2729; Zeng, 1936, p. 412).

The Qing dynasty (1644–1912) marked a critical turning point in China's relationship with its Northwest frontier. As Elliott (2014, p. 338) observes, the frontier story is 'one of the things that makes the Qing "Qing"'—the geographical framework of modern China was determined through the Qing's expansion, conquest, and consolidation of imperial territory in Inner Asia (Ho, 1967, p. 189). The Manchu rulers, drawing on their non-Han background and Inner Asian connections, developed distinctive institutions and policies for managing frontier regions that differed from those of preceding dynasties (Elliott, 2014, pp. 348–349). Republican-era discourse on the Northwest thus inherited not only the physical territorial legacy of Qing conquest but also the conceptual frameworks through which these frontier spaces had been imagined, administered, and represented. During the transition from the Qing empire to the

Republic, Nationalist elites reframed these spaces from dynastic peripheries into frontiers to be developed and integrated into a modern nation-state.

In Republican-era usage, 'Northwest' (*Xibei*) generally referred to the provinces of Shaanxi, Gansu, Ningxia, Qinghai, and Xinjiang; the historical province of Suiyuan was often included, and at times Mongolia appeared in the ambit as well (Jiang, 2001, p. 42; Ni, 1936, p. 14; Tighe, 2005, pp. 92–93; Xiang, 2018, pp. 265–269; Zhang, 1934a, p. 7; Zhang, 1989, pp. 164–167). In this study we follow this broad convention, noting that shifting administrative boundaries and overlapping jurisdictions sometimes brought adjacent regions (e.g., Chuanbian/Xikang or Tibet) into discussions of the Northwest. In this sense, conceptually, the Northwest was defined in contrast to the eastern heartland of China proper—a periphery within national borders that was nevertheless culturally and ethnically distinct from the eastern core. The region's population included substantial non-Han ethnic and religious communities, including Hui, Uyghur, Mongol, and Tibetan groups. Historically, contemporary writers often depicted the Northwest as a backward frontier inhabited by 'other' peoples (Dikötter, 2015; Lipman, 1997; Newby, 1999). Because frontier was an ideologically charged, inward borderland category, contemporaries cast 'development' as both material improvement and incorporation into the Chinese state.

During Republican China, development and its cognates—construction and management—did not denote a single sector. Press writers bundled material improvement (machines, roads, railways, hydraulic works, mines), territorial and administrative integration (new jurisdictions, fiscal-police reach, settlement and demographic engineering), and civilisational uplift (schooling, hygiene, cultural reform aimed at 'transforming' frontier peoples) into one frontier-building project. This bundle carried forward late-Qing 'self-strengthening'

assumptions that infrastructure and technology generate civilisation (Wu, 2015, p. 27), a vision famously blueprinted in Sun Yat-sen's 'development programmes' of China (Sun, 2021), while also reflecting a high-modernist faith in synoptic planning and social simplification (Scott, 1998) and the extension of the state's 'infrastructural power' across space (Mann, 1984, p. 189). In short, 'developing the Northwest' named both means (technical fixes) and ends (state penetration and civilising mission). We treat this as a discursive definition grounded in press texts: our analysis maps how newspapers assembled these domains into a single project; it does not infer policy implementation or mass consensus from the press itself.

Although the Northwest development movement was broad, scholarship has illuminated it only in parts. Existing studies have typically focused on specific aspects—such as strategic integration of particular frontier provinces, economic and administrative policies, or ethnic frontier governance (Shen, 2007; Zhang, 2021; Zong, 2003)—but no single work offers a corpus-wide view of the press discourse that knit these efforts together. We pursue three linked questions. First, what thematic constellations structured Republican-era discourse on Northwest development? Second, how did those themes co-occur and bridge policy domains? Third, how did their prevalence shift over time across major inflection points? Answering these questions enriches our understanding of Republican China's frontier building and nationalist ideology, yet requires grappling with sources on a vast scale. Thousands of articles and essays about Northwest China appeared in the Republican-era press, far too many for any single researcher to closely read in full. This study meets that challenge by leveraging digital archives and computational text analysis to systematically examine how Republican China discussed the development of its Northwest frontier. We assemble a large corpus of over 5,000 newspaper and periodical pieces from 1911–1949 and employ structural topic modelling (STM) to inductively

identify the major themes in this discourse and to trace their connections and temporal dynamics. By analysing this extensive textual dataset, we reveal patterns and transformations in rhetoric that elude more narrowly focused research, thereby shedding new light on the interplay between media, discourse, and power in modern Chinese history.

The remainder of the paper is organised as follows. Section 2 situates our work within existing scholarship, reviewing the role of print media in shaping discourse in Republican China, studies of Northwest China's development in the Republican era, and computational approaches to historical research. In Section 3, we describe our data and methodological framework, including source collection, text preprocessing, and the implementation of the structural topic model. Section 4 presents our analytical results, highlighting the key topics discovered, their thematic groupings, and the temporal dynamics. Finally, Section 5 synthesises our findings on how press discourse shaped the Northwest as a national project, reflects on the value of digital humanities approaches for historical inquiry, and discusses limitations and directions for future research.

2. Literature Review

2.1 Press as a Discourse-Shaping Arena in Republican China

Newspapers and periodicals have long served as critical forums for public discourse and state-society interaction, a phenomenon particularly evident in Republican China. As Zhou (2006, p. 49) observes, 'it was the appearance of modern newspapers, assisted by the telegraph, that made the formation of large-scale public opinion possible'. Indeed, the late Qing and early Republican decades witnessed an unprecedented proliferation of print media, which fundamentally transformed the landscape of civic discourse. Scholars have documented how new publications

and mass-circulation newspapers revolutionised political communication, catalysing popular engagement in public affairs on a national scale (Mackinnon, 1997; Mittler, 2004; Weston, 2010; Zhao, 2008). The rapid expansion of print culture served as a catalyst for major intellectual and social transformations spanning the late Qing reforms, the ideological ferment of the May Fourth Movement, and the emergence of localised public spheres in interior urban centres. Notably, the influence of this new media environment was especially concentrated among urban, literate constituencies—including intellectuals, officials, and students—who not only formed the primary readership but also played active roles as contributors, shaping and propelling these evolving public discourses through both their reading and writing (Cheek, 2015, pp. 8–9, 94).

Print media in this era not only disseminated information but also forged new collective identities. In this vein, print media functioned as a primary vehicle for cultivating emotional attachments to modern collective identities at an unprecedented scale—effectively generating public sentiment. As Anderson (2006, p. 134) argues, 'print-language is what invents nationalism, not a particular language per se'—by reading the same news and commentary, people who had never met could feel connected by common concerns and ideas. Republican-era newspapers functioned as engines of nationalism, translating abstract concepts of the nation-state into concrete stories, symbols, and debates that resonated with everyday life. Studies of specific newspapers demonstrate this dynamic: Zhao (2025) examines how the *Guangdong Qunbao* in the 1920s employed 'print capitalism' to disseminate socialist propaganda and cultivate cultural strategies for political mobilisation; Wu (2023a) shows how newspapers orchestrated 'public opinion' campaigns during the 1919 Paris Peace Conference and May Fourth Movement, transforming diplomatic disputes into mass political mobilisation; and Veg (2021) demonstrates how the local press in Chengdu catalysed political activism in the 1911 Railroad Protection

Movement and the 1919 May Fourth Movement, solidifying public opinion around both local and national concerns.

Among the pressing national concerns that generated intense media attention was the strategic imperative of developing China's northwestern frontier, which emerged as a subject of sustained debate across the country's burgeoning media landscape. Building on this press-centred perspective, the next subsection reviews how scholarship has treated 'Northwest development' within that media ecosystem and where gaps remain.

2.2 Northwest Development in Republican-Era Scholarship

The idea and practice of developing China's Northwest have attracted substantial scholarly attention. Numerous studies, both in English and Chinese, have examined aspects of this historical phenomenon. In English-language scholarship, Justin Tighe's work stands out: his monograph *Constructing Suiyuan* (Tighe, 2005) and related article (Tighe, 2009) analyse the politics of northwestern territorial development, focusing on how the Chinese state attempted to integrate the Suiyuan region (today part of Inner Mongolia) into the nation-state. Tighe illustrates the processes of state building at the frontier, highlighting the interplay between central policies and local conditions in what was effectively a trial case of frontier governance. Another important contribution is made by Tai (2015), who provides a broad overview of Republican-era Northwest development policies and debates. Tai's research reconstructs how different Nationalist officials and intellectuals conceptualised the Northwest's development, revealing both ambitious economic plans and the ideological currents that framed the 'Northwest question'. From the perspective of ethnicity and frontier peoples, Lin (2011) examines Nationalist strategies toward China's non-Han minorities after the Qing dynasty, with the Northwest development movement featured as a key case. He argues that initiatives in the

Northwest played a pivotal role in shaping the territorial boundaries and ethnic policies of the modern Chinese state, essentially defining China's Central Asian frontier and its security paradigm in the first half of the twentieth century.

Chinese-language scholarship on the Northwest development movement is even more extensive, drawing on rich primary sources to construct detailed historical narratives. Zhang (1989) offered an early, foundational chronology of evolving perspectives on Northwest China's development from the late Qing through the Republican era. Zhang's study synthesises evidence from contemporary newspapers, personal diaries, travel accounts, and government reports to trace how thinking about the Northwest changed over time. This empirical groundwork has been extended by later scholars through focused monographs and edited volumes. For example, Cheng, Wang and Zhang (2007) examines Nationalist government policies and ideological debates on developing the Northwest, while Tian (2007) analyses the cultural and intellectual underpinnings of Republican-era expansion into the Northwest. Wang (2015) investigates the institutional frameworks and economic measures that structured regional development policies during the Nanjing decade, shedding light on how central planning intersected with local execution in the Northwest. A number of studies zoom in on particular facets of the movement: some explore specific sectors such as education (Hu, 2020), transportation infrastructure (Baker, 2024; Shang and Ding, 2023), or the genre of travel writing that promoted Northwest awareness (Nian and Lin, 2019; Shen, 2006). Others examine the activities of individual organisations or key figures involved in frontier programs. For instance, Jia and Hua (2002) analyses how the influential newspaper *Ta Kung Pao* reported on Northwest development in the 1930s, demonstrating the media's role in shaping public perceptions of the region. And Zhang (2002) assesses Nationalist Northwest policy in the decade before the Second Sino-Japanese War,

arguing that prior to 1937, development efforts were driven largely by national security concerns (defence, transportation, water conservancy) tailored to on-the-ground regional conditions.

Taken together, this rich body of scholarship establishes several key insights. First, Northwest development emerged as a crucial component of Republican China's state building project on the frontier, encompassing far more than economic modernisation alone. These studies make clear that development initiatives were deeply entangled with questions of national unity, territorial security, and the construction of modern Chinese identity. Second, the Chinese state's push into the region represented an effort to extend its authority and what it conceived as civilisation to a long-neglected borderland. Finally, scholars increasingly recognise that the Northwest development movement functioned on two levels simultaneously: as a set of concrete policies and institutional programs, and as a discursive project aimed at convincing diverse audiences—from coastal elites to frontier administrators—that this distant region held urgent significance for China's national future.

Yet despite these significant contributions, there are notable gaps in existing studies. Many rely on detailed, close readings of selected archives or print sources, often focusing on a specific province, initiative, or thematic issue (e.g., Tighe, 2005; Wang, 2010; Yang, 2013; Yan and Zhang, 2006), and thereby providing valuable yet limited insights. These approaches typically present deep but focused perspectives, which are crucial but fail to offer a comprehensive view of the diverse and evolving discourse surrounding the development of the Northwest over time. The full complexity of the Northwest's development discourse, encompassing a broad range of themes and voices throughout the Republican period, remains underexplored. How did different aspects of development (e.g., infrastructure, economic policy, cultural integration, military concerns) interrelate in contemporaries' minds? Which ideas

enjoyed broad prominence and which remained marginal? And how did emphases change from the 1910s to the 1940s? These questions are difficult to answer with conventional historiographical methods that necessarily concentrate on manageable subsets of sources.

It is precisely this recognition of the discursive dimension—and the need for a broader empirical foundation—that motivates our study. We shift focus from examining specific policy outcomes (which prior scholarship has documented well) to systematically analysing the discourse itself: the arguments, metaphors, and themes through which Republican-era Chinese intellectuals, officials, and journalists conceived of their Northwest frontier. By adopting a data-driven approach to a large corpus of newspapers and periodicals, we aim to recover the full contours of this discourse—its dominant themes, internal variations, and temporal evolution—at a scale that complements existing case studies. The next subsection explains why computational methods are well suited to address these questions and how prior work has applied them to press discourse.

2.3 Computational Methods in Historical Research

The explosive growth of large-scale, digitised textual corpora is reshaping how scholars investigate the past (Beelen et al. 2025; Rosenzweig, 2003; Zaagsma, 2023). Quantitative analysis of historical data has a long pedigree, dating back to the mid-20th-century ‘cliometrics’ movement (e.g., Fogel and Engerman, 1974), and recent digitisation plus advances in natural language processing have accelerated a broader ‘text-as-data’ turn in humanities and social sciences (Grimmer, Roberts and Stewart, 2021; Wilkerson and Casas, 2017). For historians, computational approaches mitigate the chronic problem of information overload, enabling macro-level views across sprawling archives (Guldi, 2023; Milligan, 2019). This has catalysed a

turn toward mixed methods, where ‘distant reading’ algorithms like topic modelling guide, but do not replace, traditional close reading (Armand and Henriot, 2023; Underwood, 2019).

Recent scholarship illustrates the value of these tools for historical inquiry, especially for analysing print press as discourse-shaping media. One early study is Newman and Block (2006)’s topic modelling of an eighteenth-century American newspaper, which revealed latent themes in colonial-era newsprint. Similarly, DiMaggio, Nag and Blei (2013) applied Latent Dirichlet Allocation to nearly 8,000 newspaper articles on government arts funding across U.S. outlets (1986–1997), demonstrating how the method can identify competing frames within policy debates and track their changing prevalence over time and across publications. More recently, Viola and Verheul (2020) used a ‘discourse-driven’ topic model on immigrant newspapers in the United States, showing how computational analysis can trace the evolution of narratives about identity and assimilation. These precedents underscore why topic modelling is well suited to recover frames, track salience, and map thematic relationships in print press discourse.

Scholarship on Chinese history has likewise leveraged computational methods to analyse large textual corpora. Miller (2013) employs topic modelling on Qing archival records to trace patterns of rebellion, crime, and violence across nearly two centuries (1722–1911), revealing spatial and temporal dynamics that traditional close reading could not easily discern. Moving from historical events to intellectual discourse, Hong and Chen (2024) analyse rhetorical strategies in court remonstrations spanning two millennia of Chinese dynastic history, uncovering a payoff-biased cultural evolution in which later dynasties exhibited higher persuasion success and declining use of less effective rhetorical devices. Their supervised machine-learning approach transforms long-standing interpretive debates about classical rhetoric into testable empirical hypotheses. More recently, Pelzer (2025) analysed over 17,600

biographical entries of Republican-era Chinese engineers to map elite geographic mobility, demonstrating how regional push-pull factors and national political dynamics shaped the movement of technical talent across China's nation-building projects. Extending into the PRC era, Gilkison and Kurzynski (2024) apply text mining and large language model classification to articles from the *People's Liberation Army Daily* (1956–1989), showing how computational analysis of newspaper discourse can expose the ideological strategies through which the state legitimises power. Together, these studies illustrate how computational techniques—whether topic modelling, supervised classification, large-scale prosopography, or text mining of press archives—can reveal macro-level patterns and historical structures that complement close readings of individual sources.

The rise of artificial intelligence has further expanded the toolkit available to scholars. Machine learning and large language models (LLMs) now enable researchers to transcribe degraded manuscripts, decode ancient scripts (Assael et al. 2022; Eberle et al. 2024), and extract structured information—such as named entities, dates, and relations—from unstructured texts (Hou and Huang, 2025; Underwood, 2025; Wencker, Borst-Graetz and Niekler, 2025). Multimodal AI systems such as Google Gemini have demonstrated superior performance over conventional optical character recognition on complex document layouts (Chow, 2024; Gemini Team, 2024), a capability especially valuable for processing historical Chinese newspapers with dense vertical text and mixed typefaces. Recent applications integrate LLMs (e.g., GPT-4o) with historical gazetteers to extract and geocode toponyms from classical Chinese texts (Chen et al. 2025), demonstrating how AI can resolve ambiguities that defeated earlier methods. These advances carry particular significance for Chinese studies, where recent tightening of archival

access (Greitens and Truex, 2020; Mertha, 2024; Northrop, 2022) has made digitised and previously collected sources increasingly critical to sustained scholarship.

Just as important as these technical breakthroughs is the growing accessibility of computational tools. User-friendly software packages and open-source libraries for text cleaning, topic modelling, and natural language processing (Benoit, 2020; Grimmer, Roberts and Stewart, 2022; Short, McKenny and Reid, 2018; Tonidandel et al. 2022) have lowered barriers to entry, enabling scholars without extensive programming training to engage with large-scale textual analysis. In our study, this integrated approach is designed specifically to address the gaps outlined above by reconstructing the contours, emphases, and shifts of the Republican-era Northwest development discourse at scale and in its press context.

3. Data and Method

3.1 Data Collection

This study draws upon two comprehensive digital databases of historical Chinese newspapers and periodicals, which provide primary source materials from late imperial and Republican China. The first source is the *Quan Guo Bao Kan Suo Yin* (CNBKSY),[1] a vast digital repository maintained by the Shanghai Library. CNBKSY encompasses over 50 million indexed entries drawn from more than 50,000 publications, spanning from the late Qing dynasty through the contemporary period. It is one of the most extensive digital archives of Chinese periodical literature available to researchers. The second source is the *Shenbao* Full-Text Database,[2] developed by Get-Hong Communication Company in Taiwan. This database contains digitised

[1] CNBKSY: https://www.cnbksy.com.

[2] *Shenbao* Database: https://tk.cepiec.com.cn/SP.

text of *Shenbao*, one of the most influential and long-running Chinese-language newspapers of the modern period (1872–1949). The *Shenbao* newspaper was a key forum for public discourse in late imperial and Republican China, making its archive an invaluable resource for historical research.

To extract relevant articles on the development of Northwest China during the Republican era (1911–1949), we employed a systematic keyword search strategy across both databases. Table 1 presents our comprehensive search methodology. The search strategy employed two types of queries: combinations of 'Northwest' with six development-related terms, and one standalone mobilisation phrase 'Go to the Northwest'. The selection of these specific terms was based on close reading of the historical literature and preliminary corpus analysis, ensuring comprehensive capture of relevant discourse while maintaining specificity to the Northwest development theme. For each search query, 'Northwest' and the development term must both appear (linked with AND), and these different query combinations were then linked using Boolean OR operators to capture the full range of relevant discourse. All searches were conducted across multiple metadata fields, including titles, subtitles, abstracts, and publication names. For the *Shenbao* Database specifically, we enhanced precision by constraining the proximity of keyword pairs to within 10 Chinese characters, which helped filter out incidental mentions not directly related to regional development. Using Python-based web scraping scripts, we retrieved an initial corpus of 9,677 relevant entries (7,970 from CNBKSY and 1,707 from *Shenbao*).

Table 1: Search Strategy for CNBKSY and *Shenbao* Databases

Search Terms	CNBKSY	*Shenbao*
西北 開發 (Northwest + Exploitation)	No proximity constraint	Within 10 characters
西北 發展 (Northwest + Development)	No proximity constraint	Within 10 characters
西北 建設 (Northwest + Construction)	No proximity constraint	Within 10 characters

西北 經營 (Northwest + Administration)	No proximity constraint	Within 10 characters
西北 考察 (Northwest + Investigation)	No proximity constraint	Within 10 characters
西北 調查 (Northwest + Survey)	No proximity constraint	Within 10 characters
到西北去 (Go to the Northwest)	Exact phrase match	Exact phrase match
Date Range	1911–1949	1911–1949
Total Results	7,970	1,707

These entries encompass a range of document types, including news articles, editorials, academic essays, translated works, government reports, and opinion pieces, reflecting the wide-ranging discussion of Northwest development at the time. We recognise that these databases, centred on CNBKSY periodicals (predominantly from eastern cities) and *Shenbao* (Shanghai-based), reflect coastal and urban elite perspectives rather than the entire spectrum of contemporary viewpoints. Our findings therefore represent media framing of Northwest development from China's political and intellectual centres—what editors and journalists printed—rather than a comprehensive account of local experiences or government enactments.

While the *Shenbao* database provides full-text content (already processed and searchable), the CNBKSY entries are primarily page images (scanned PDFs) without machine-readable text. To obtain usable text from CNBKSY, we implemented a custom OCR (Optical Character Recognition) pipeline using Google Gemini, a state-of-the-art multimodal large language model, for text extraction (Gemini Team, 2024). This model has demonstrated exceptional accuracy in recognising traditional Chinese print, outperforming conventional OCR solutions such as Adobe Acrobat, ABBYY FineReader, and Tesseract in recent evaluations (Chow, 2024; Filimonov, 2025). Additionally, Gemini's API access and relatively low cost made it feasible to apply at the large scale required by our corpus.

To optimise OCR accuracy, particularly for historical documents of varying quality, we implemented a robust three-stage processing pipeline. First, we systematically excluded entries

containing only images or inscriptions, eliminated duplicate articles appearing across multiple publications, and conducted manual reviews to remove irrelevant content. Second, we addressed the challenge of multi-column layouts in historical publications through a semi-automated approach. We developed a script to automatically identify and separate individual columns into discrete PDF pages, followed by manual verification and correction of any segmentation errors prior to OCR processing. Third, for a small subset of documents with severely degraded quality that remained illegible even after image enhancement, we employed a hybrid human-machine approach as a last resort. A research assistant would read these badly faded articles aloud in Mandarin Chinese, creating audio recordings that were then transcribed using OpenAI's Whisper speech-to-text technology (Radford et al. 2022). The resulting transcriptions were manually reviewed and corrected against the original documents to ensure accuracy. This method was applied to only a limited number of documents where standard OCR proved inadequate, ensuring we did not exclude valuable historical content due to poor scan quality. This multi-stage approach combining automated OCR, human-assisted transcription, and manual verification ensures high-quality text extraction while maintaining the integrity of the historical materials.

After completing these initial refinement steps and pre-processing the scanned PDFs, we excluded documents with severely compromised legibility.[3] This left us with 3,270 articles suitable for OCR processing. For these remaining documents, we developed an automated pipeline utilising Python to interface with the Gemini API, employing carefully designed prompts to maximise text recognition accuracy while minimising potential errors. While this approach enabled efficient text extraction, we remained mindful of potential hallucination

[3] For certain articles that are extremely difficult to identify for the human eye, we try to find the full text from other sources, such as Hytung: https://www.neohytung.com, another database of Chinese newspapers and periodicals, as well as published books which compiled articles from the Republican China era, such as Dong et al. (1998).

artefacts inherent to large language models. To maintain data quality, we implemented a rigorous post-processing protocol that included manual verification and correction of apparent errors.

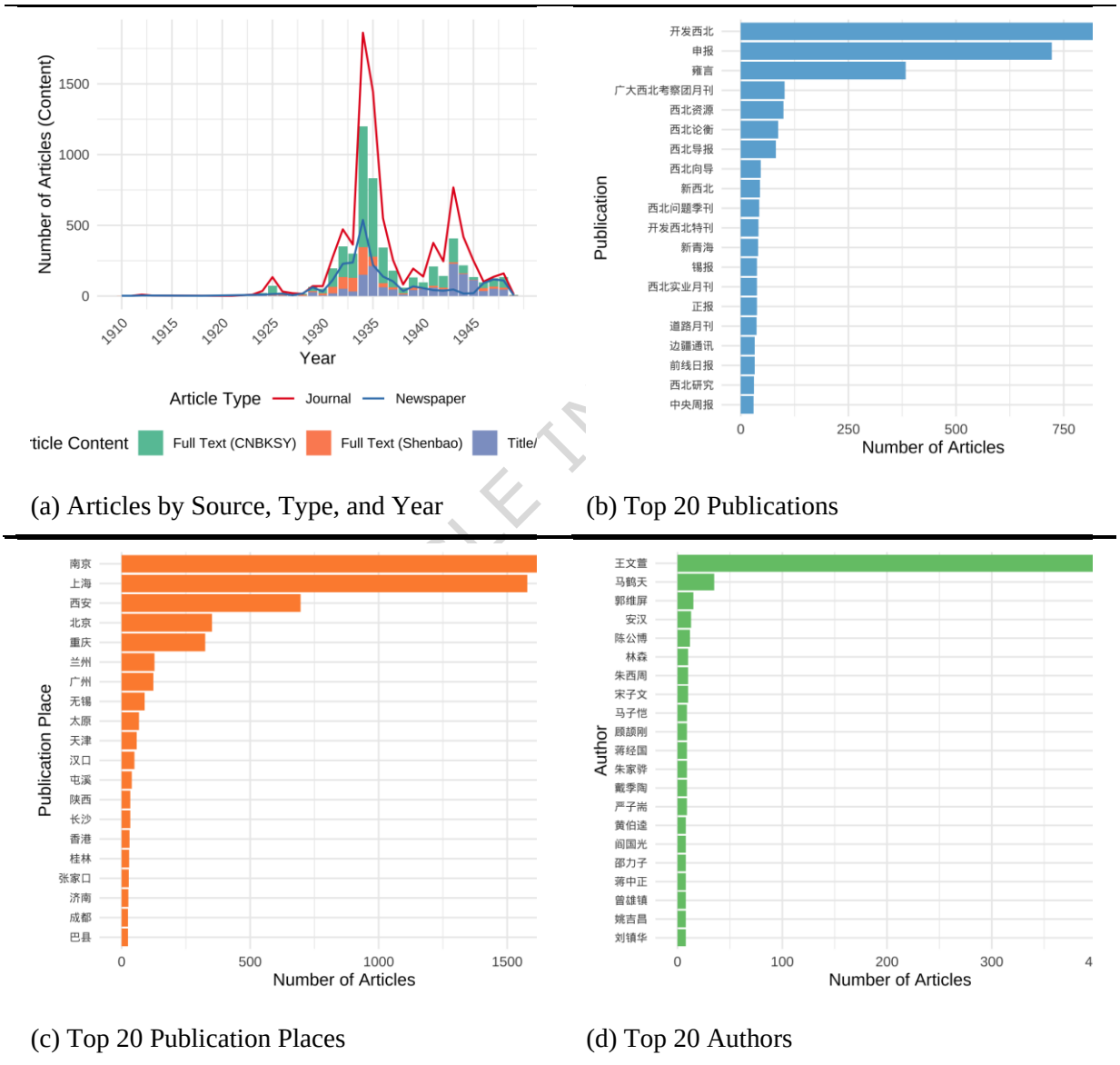
A significant challenge in processing historical Chinese texts lies in the substantial evolution of character forms since the late nineteenth century (Weng, 2023; Zhong, 2019). These historical documents frequently employ variant characters (*yitizi* 異體字) that are rarely encountered in contemporary Chinese texts and may have multiple modern equivalents. Given that contemporary NLP tools are predominantly trained on modern Chinese corpora, we standardised the text by converting variant characters to their modern Simplified Chinese equivalents. We then applied a Traditional-to-Simplified Chinese conversion process using the OpenCC framework[4] to ensure consistency across the corpus and eliminate potential ambiguity in character interpretation. While this standardisation process necessarily involves some loss of historical linguistic nuance, it enables robust computational analysis while maintaining semantic fidelity to the source materials.

The resulting dataset comprises 5,461 articles, with 4,739 sourced from CNBKSY and 722 from *Shenbao*. Of this corpus, 3,270 articles contain complete text from CNBKSY, and 722 articles include full text from *Shenbao*. For articles where CNBKSY scans proved insufficiently legible for full-text extraction, we retained the entries for their significant contextual value and topical relevance,[5] utilising their titles and available abstracts as corpus representations. The dataset demonstrates substantial coverage across publication types, with journal articles comprising 4,258 entries (78% of the total corpus) and newspaper articles accounting for 1,203

[4] Open Chinese Convert (OpenCC) is an open-source library for conversions between Traditional Chinese, Simplified Chinese and Japanese Kanji (Shinjitai): https://github.com/BYVoid/OpenCC.

[5] Structural topic modeling requires that no documents have completely missing text or metadata, which is one reason we chose not to remove CNBKSY entries lacking full text. We included those entries (often with titles or summaries) so that the dataset remained complete, though their influence on the model was minimized.

entries (22%). Notably, 3,992 articles (73.1%) contain complete text from CNBKSY and *Shenbao*. Figure 1 provides a detailed visualisation of the corpus distribution, illustrating temporal patterns across sources and types, while also mapping the distribution across the top 20 publications, publication locations, and authors.



(a) Articles by Source, Type, and Year

(b) Top 20 Publications

(c) Top 20 Publication Places

(d) Top 20 Authors

Figure 1: Summary of Data Distribution, Top Publications, Places, and Authors

Figure 2 illustrates the geographic distribution of articles on Northwest China's development by publication location. The visualisation reveals a distinct spatial pattern with

notable concentrations in eastern urban centres. Shanghai and Nanjing emerge as the primary hubs of discourse, with each generating over 1,500 articles. Other eastern metropolitan areas, including Beijing, Tianjin, and Guangzhou, maintained substantial yet comparatively modest publication outputs. Chongqing gained prominence as a publication centre after 1937, following the Nationalist government's relocation during the Japanese invasion. Within Northwest China itself, Xi'an (Shaanxi) and Lanzhou (Gansu) served as regionally significant publication sites owing to their strategic geographic locations. This spatial distribution reflects the broader political economy of publishing in Republican China, where eastern cities housed the most established press institutions, even as wartime conditions forced shifts in these patterns. The concentration of discourse in cities outside the Northwest region itself demonstrates that development initiatives were conceptualised as a national movement rather than merely a regional concern.

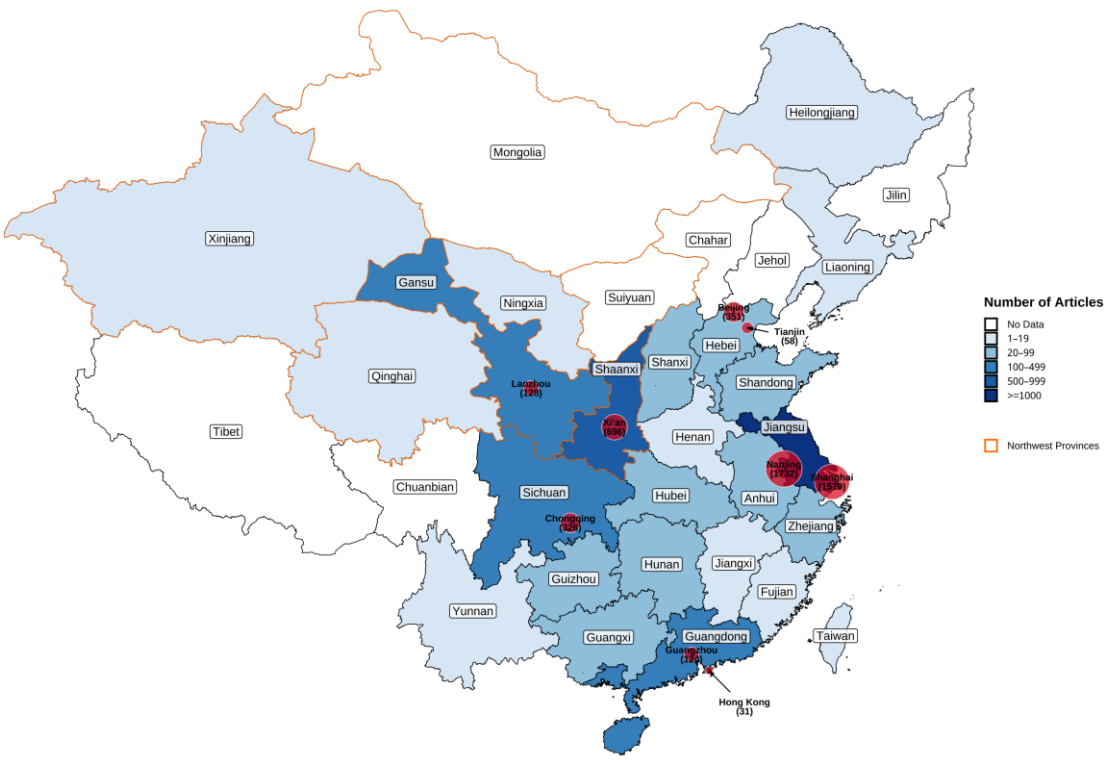


Figure 2: Geographic distribution of articles on Northwest development by publication location. Circle size represents article frequency from each city over the 1911–1949 period. Base map showing 1926 provincial boundaries derived from CHGIS (2012); visualisation by the author.

3.2 Data Cleaning

Text preprocessing for Chinese language presents unique challenges compared to English and other alphabetic languages. While English text processing typically involves straightforward steps like case normalisation and removal of punctuation (Tonidandel et al. 2022, p. 4), Chinese text requires specialised tokenisation techniques due to its lack of explicit word boundaries. This fundamental difference makes Chinese text preprocessing particularly challenging for natural language processing applications.

Word segmentation—the process of dividing Chinese text into meaningful word units—is a critical first step. While several contemporary word segmentation tools exist, such as SnowNLP[6] and THULAC[7], their effectiveness on historical texts can be limited since they are primarily trained on modern Chinese corpora. The ENPChina team's HistText tool (Blouin et al. 2023), specifically developed for 'transitional Chinese' texts from the late 19th century to early 1950s, showed suboptimal performance in our experiments. This limitation is particularly relevant given that, as shown in Figure 1 (a), the majority of our corpus consists of texts from post-1930. Despite being considered 'transitional Chinese', these texts share substantial linguistic similarities with contemporary Chinese. After extensive comparative testing, we found that Jieba[8], when augmented with a custom dictionary tailored for Republican-era Chinese texts, provided optimal segmentation accuracy for our corpus.

Following word segmentation, we addressed the removal of stop words—high-frequency words that contribute minimal semantic value to the analysis. While no universal stop word list exists, we employed a comprehensive approach by combining the established Harbin Institute of Technology (HIT) stop word dictionary with an additional dictionary specifically curated for Republican-era Chinese texts. Furthermore, we enhanced the stop word list by including domain-specific terms that could potentially skew our analysis. These terms, being our initial search keywords, appeared with high frequency in the corpus and their removal was necessary to reduce noise in the subsequent structural topic modelling analysis. After completing the segmentation and stop word removal, our final corpus for STM analysis contained 7,406,390 characters.

[6] Snownlp: https://github.com/isnowfy/snownlp.
[7] THULAC: http://thulac.thunlp.org.
[8] Jieba: https://github.com/fxsjy/jieba.

3.3 Research Method

In this study, we employ a text-as-data methodology (Benoit, 2020; Chiu et al. 2025; Gentzkow, Kelly and Taddy, 2019; Grimmer, Roberts and Stewart, 2022) to analyse the corpus of Republican-era newspaper and periodical writings on Northwest development. Specifically, we utilise structural topic modelling (STM) (Roberts, Stewart and Tingley, 2019), an advanced form of topic modelling techniques, to conduct a comprehensive, data-driven analysis of how contemporary press coverage reflected and shaped evolving narratives about Northwest China's development during this pivotal historical period.

Topic modelling, originally developed for document classification and information retrieval applications (Blei, Ng and Jordan, 2003), has emerged as a powerful analytical tool in the humanities and social sciences for uncovering latent concepts and generating semantic theoretical insights (Ying, Montgomery and Stewart, 2022). These computational methods excel at discovering hidden thematic structures within large-scale textual corpora and generating meaningful topical representations of documents. The fundamental premise of topic modelling is that documents are composed of multiple topics in varying proportions, where each topic represents a probability distribution over words. By analysing patterns of word co-occurrence across thousands of texts, topic models can identify coherent semantic themes without requiring pre-defined categories or manual annotation. The robust capabilities of topic modelling have made these methods indispensable across multiple domains, including text mining, information retrieval, and social science research (Tang et al. 2025). Notably, topic modelling algorithms have demonstrated validity comparable to human coding (Roberts et al. 2014), while complementing traditional analytical approaches (Nelson et al. 2021) and reducing potential analytical biases (Debnath et al. 2020). These advantages make topic modelling particularly

well-suited for examining historical discourse patterns across extensive newspaper and periodical collections.

We implement structural topic modelling (STM) using the stm package (Roberts, Stewart and Tingley, 2019) in R (R Core Team, 2025) for its distinctive methodological advantages aligned with our research objectives. STM provides a probabilistic approach to text analysis by modelling topics as distributions over a fixed vocabulary, where documents are represented as mixtures of these topics (Roberts, Stewart and Airoldi, 2016). This methodology employs unsupervised machine learning algorithms to discover latent semantic structures within texts, offering advantages over both traditional frequency-based methods and conventional topic modelling approaches. Crucially, STM incorporates document-level metadata as covariates affecting topic prevalence (Roberts et al. 2014), allowing us to estimate how topic distributions vary with publication year, location, and source type within a single probabilistic framework.

Compared to alternative approaches, STM is particularly well-suited for our specific analytical objectives. While newer transformer-based methods such as BERTopic (Grootendorst, 2022) offer promising capabilities and convenient visualisations, they rely on pre-trained sentence embeddings that are tuned on contemporary web-era datasets and multilingual parallel corpora rather than Republican-era press. High-quality Chinese and multilingual models exist, but domain mismatch and OCR noise in historical sources can degrade their downstream performance (Todorov and Colavizza, 2022). Moreover, BERTopic typically assesses temporal patterns through post-hoc grouping rather than jointly estimating multi-covariate effects on topic prevalence. For historical Chinese corpora with OCR noise and orthographic variation, STM's bag-of-words representation—paired with our cleaning and standardisation pipeline—provides a robust and interpretable foundation. The robustness and versatility of STM have led to its

widespread adoption across diverse academic disciplines, including political science (Lim, Ito and Zhang, 2025; Xia, 2024), environmental research (Gavriș and Popescu, 2024; Mennig, 2025), urban studies (Cui et al. 2025; Morandell, Wicki and Kaufmann, 2025), psychology (Abraham et al. 2024; Șakar and Tan, 2025), and tourism management (Chen, Anker and Liang, 2025; Ramos-Henriquez and Morini-Marrero, 2025), demonstrating its effectiveness in uncovering meaningful patterns in complex textual data.

4. Findings

4.1 Number and Substance of Topics

After data cleaning and preprocessing, we proceed with the STM analysis to uncover the latent topics in our corpus. Even though the STM model is a powerful tool for analysing text data, it facilitates the exploration of texts in a semiautomated way, which means it requires careful consideration and input from the researcher. The first critical step is determining the optimal number of topics, which is not something that can be defined arbitrarily. One needs to assess what number of topics will match the data. The optimal number of topics (parameter *K*) is dependent on various factors. If *K* is very small, then the corpus will be divided into a few general semantic contexts, whereas if *K* is very large, then the collection will be divided into numerous topics, a few of which may overlap, such that topics may be difficult to interpret. There is no one-size-fits-all or 'right' answer for choosing the number of topics *K* (Grimmer and Stewart, 2013, pp. 285–286; Schmiedel, Müller and vom Brocke, 2019, p. 948); interpretability is an important justification in practice (Blei, 2012, pp. 82–83). A good topic structure should consist of understandable, meaningful, and distinct semantic clusters, with each topic being mutually exclusive.

Following the approach outlined in Roberts, Stewart and Tingley (2019) and Weston et al. (2023), we utilise an iterative procedure to first evaluate models with different numbers of topics ranging from 5 to 50, incrementing by 5 using the `searchK()` function. Next, we inspect both the cross-validation likelihood and semantic coherence to decide how many latent topics are in our corpus. The cross-validation likelihood reflects the fit of each solution in a hold-out sample whereas semantic coherence is a measure of how often words in a topic co-occur. The semantic coherence score measures how often high-probability words within a topic co-occur together, while exclusivity captures how unique the words are to each topic. The held-out likelihood indicates the model's predictive accuracy on unseen data (Roberts, Stewart and Tingley, 2019).

Figure 3 contains a depiction of each metric across the various solutions and, when considering both these criteria, it suggests that the optimal number of topics falls between 20 and 30 topics. The choice of the final models was a trade-off between the statistical results based on a high held-out likelihood and a low residuals dispersion, while downplaying the coherence of topics (a high value for the interpretability of topics is usually easy to obtain). We then refined our topic search to explore all solutions between 20 and 30 topics incrementing by 1 and again visually inspected the cross-validation exclusivity and semantic coherence results of each topic. Finally, we identified 26 topics as the optimal solution based on these multiple criteria to implement the STM model.

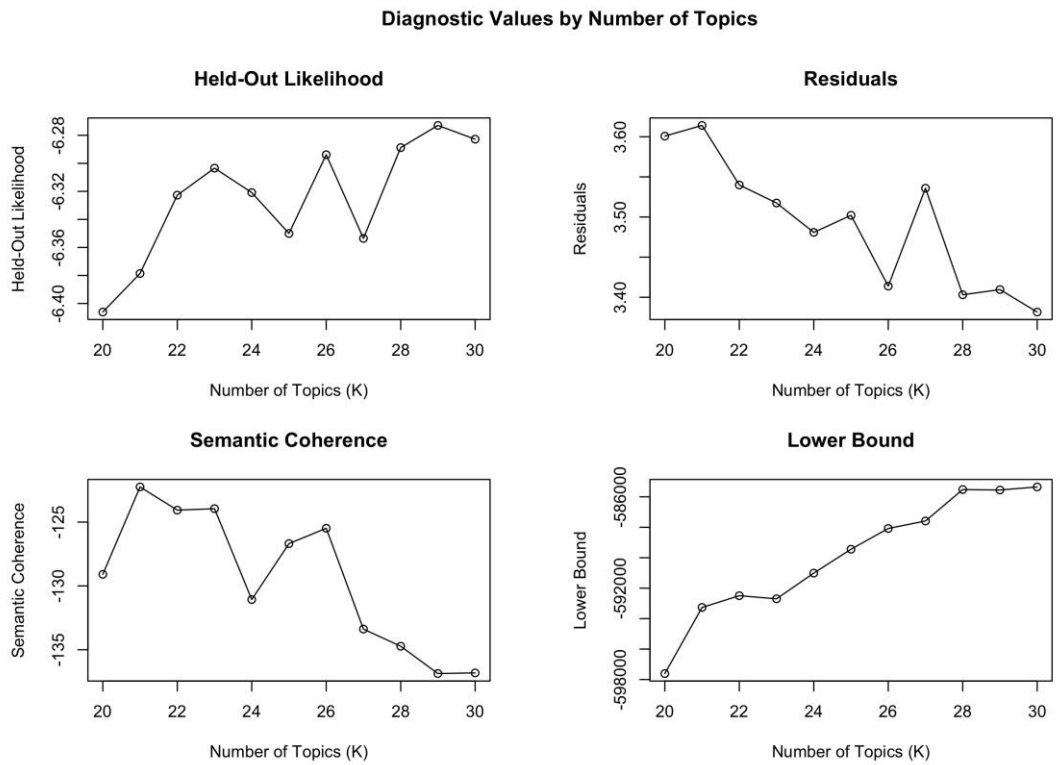


Figure 3: Diagnostic Values by Number of Topics

Due to the stochastic nature of topic modelling algorithms, different random initialisations can produce varying results even with the same number of topics. To ensure the stability and robustness of our findings, we employed the `selectModel()` function to run 20 different models with the 26 topics. This function facilitates comparing models with different random initialisations and selecting the one with optimal semantic coherence and held-out likelihood metrics. As shown in Figure 4, the third model is selected based on its superior performance across these diagnostic metrics, which indicates that this initialisation achieves a good balance between semantic coherence and exclusivity, with relatively consistent performance across multiple runs.

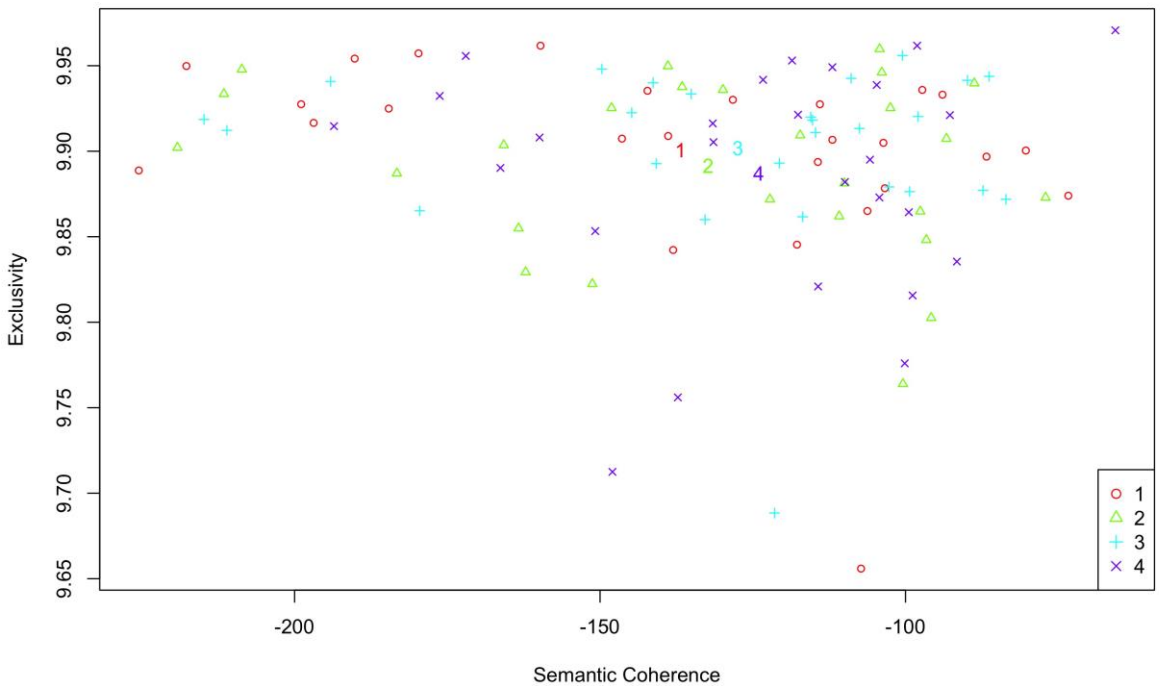


Figure 4: Model Selection Diagnostics

After selecting the optimal model, we examine the semantic content of each topic through their most representative words. For each topic, we present two word sets: highest probability words and FREX (FRequency and EXclusivity) words. The FREX metric combines information about word frequency within a topic and how exclusive words are to that topic, as measured by the ratio of the frequency of the term within a topic to its frequency in the corpus as a whole (Roberts, Stewart and Tingley, 2019, p. 13). The STM output with 26 topics, in which each topic is labelled on the basis of the highest probability and the frequent and exclusive (FREX) words, is presented in the supplementary.

The STM successfully captures the nuanced complexity of the Northwest development discourse through its 26 distinct topics. Each topic represents a significant strand in the intricate

tapestry of public and political conversation during this formative period. Broadly, the discourse encompasses several major thematic areas: infrastructure initiatives, resource exploitation, industrial and agricultural development, educational and cultural projects, national defence and administrative governance, and engagement with ethnic frontier populations. These interconnected themes reflect how the Northwestern frontier was simultaneously conceptualised as an economic opportunity, a strategic military buffer, and a space for cultural integration within the nation-building efforts.

Topic prevalence is the proportion of a document that can be associated with a particular topic (Roberts et al. 2014). STM ascribes a topic prevalence score for every topic to each document. Using topic prevalence, one can examine how frequently various topics appear in the entire corpus. To gain an overview of topical prevalence in the entire corpus, Figure 5 displays the expected proportion of each topic in the corpus. Topics in the figure are ranked in a descending order, showing the proportion of each topic relative to the entire corpus. The figure reveals that topics are not evenly distributed across the corpus. The most prevalent topics are Strategic Surveys and Regional Contention (19), State-led Northwestern Development Programs (22), and Central Administration of Northwestern Frontiers (15). These topics together represent the dominant discourse on administrative governance, strategic surveys, and state-directed development initiatives. In contrast, the least prevalent topics include Geological Surveys and Minerals (8), Anti-Imperialism and National Liberation (7), and Regional Manufacturing Development (21).

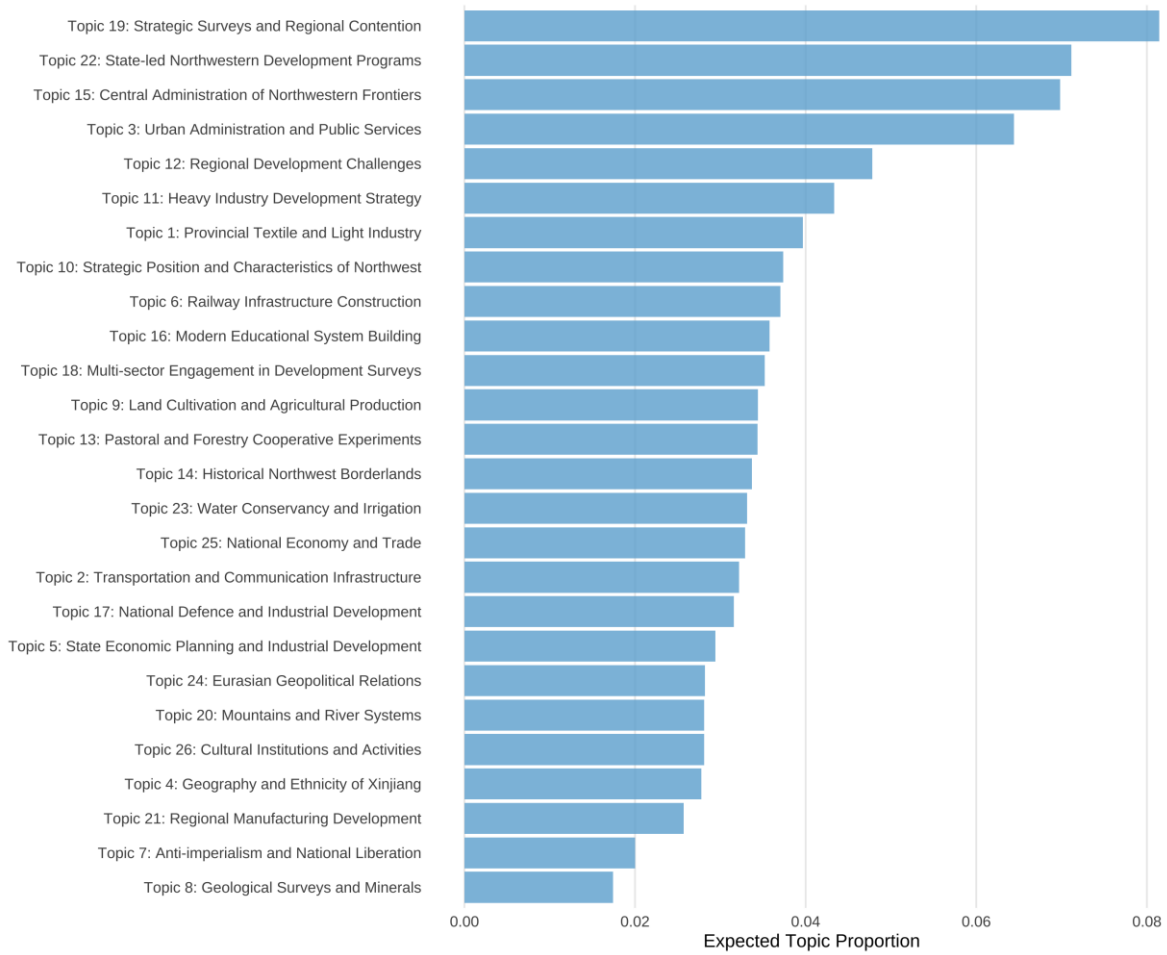


Figure 5: Expected Topic Proportions in Northwest China's Development Discourse

4.2 Correlations and Structure of Topics

Beyond individual topic identification, understanding how topics relate to one another reveals the underlying discursive architecture of Northwest development discourse. STM enables explicit estimation of topic correlations by replacing the Dirichlet distribution in standard LDA with a logistic normal distribution, allowing topics to co-occur within documents (Roberts et al. 2014). This approach generates a network representation where topics function as nodes, connected when they frequently co-occur in the same documents. Such correlation analysis

illuminates the organisational structure of the corpus and identifies overarching themes that transcend individual topics. Figure 6 presents the correlation structure among the 26 topics.

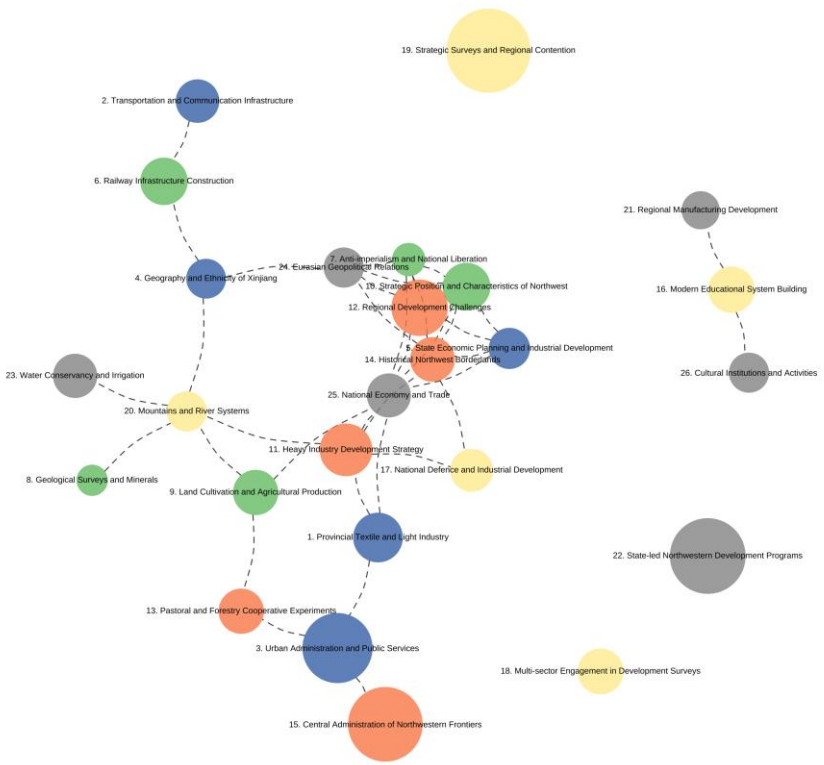


Figure 6: Visualisation of the Topic Correlations

The correlation network resolves into two dominant clusters with a small peripheral satellite. A governance-industry core sits alongside an infrastructure-resources block; link density is highest where planning and strategy concentrate, with security bridging the clusters, while cultural-educational and survey reportage are present but marginal.

State-centric hubs organise the field. The most connected nodes—State Economic Planning and Industrial Development (5), Strategic Position and Characteristics of the Northwest

(10), and Central Administration of Northwestern Frontiers (15)—tie sectoral agendas together, integrating transport (2, 6), water conservancy (23), heavy industry (11), and agriculture (9) into shared conversations about territorial integration and national priorities.

Security bridges governance and production. Anti-imperialism and National Liberation (7) and Eurasian Geopolitical Relations (24) link the planning hubs to defence and heavy industry (17, 11), indicating that security logics frequently trumped strictly economic rationales; contemporaries perceived multifront threats from Japan in the east and the Soviet in the north and west (Forbes, 1986, pp. 116–121).

Infrastructure and resources are framed through planning rather than productivity. Transportation and Communication Infrastructure (2), Railway Infrastructure Construction (6), and Water Conservancy and Irrigation (23) connect more tightly to cross-cutting problems and planning—Regional Development Challenges (12) and State Economic Planning (5)—than to downstream production nodes such as Land Cultivation and Agricultural Production (9) or Regional Manufacturing Development (21). This configuration indicates that technical works were discussed less as narrow productivity fixes and more as instruments for territorial integration and state capacity.

Cultural-educational and survey reporting are prevalent yet structurally peripheral. Modern Educational System Building (16) and Cultural Institutions and Activities (26) sit at the rim of the network with only occasional ties to Regional Manufacturing Development (21), while Strategic Surveys and Regional Contention (19) is frequent in the corpus but near the edge. This pattern suggests that descriptive reportage circulated largely as self-contained material rather than as connective tissue to policy debates, reinforcing the dominance of planning and security as the grammar of integration.

The correlation structure reveals a top-down discursive architecture in which planning and security provide the conceptual framework, infrastructure and resources supply the practical vocabulary, and cultural-educational work serves to link peripheral populations to central state objectives. This hierarchical organisation demonstrates how Northwest development discourse consistently reframed technical projects as instruments of administrative reach and strategic depth, reflecting the Republican state's efforts to extend its authority over frontier territories through both material and discursive means.

4.3 Temporal Dynamics of Topics

A central question for our study is how the Northwest development discourse changed over time. The period 1911–1949 saw tumultuous shifts in China's political and social situation, which we would expect to leave an imprint on public discussions about the frontier. To rigorously analyse temporal trends, we utilised STM's capability to include document dates as a covariate, allowing the model to estimate how the prevalence of each topic varied by year (Roberts, Stewart and Airoldi, 2016). The estimation model is formulated as:

$$Prevalence_{ij} \sim \beta_0 + \beta_1 \times Date_i + \varepsilon_i$$

In this formulation, i denotes the document index and j represents the topic index. $Prevalence_{ij}$ constitutes the matrix of topic prevalence values for each document derived from our topic modelling analysis. $Date_i$ corresponds to the publication date of the ith document in our corpus.

The visualisation in Figure 7 illustrates the chronological evolution of all 26 topics, with vertical dashed lines marking four critical historical junctures: the establishment of the Xibei Jun (1919), the Manchurian Incident (1931), the outbreak of the Second Sino-Japanese War (1937), and Japan's surrender (1945).

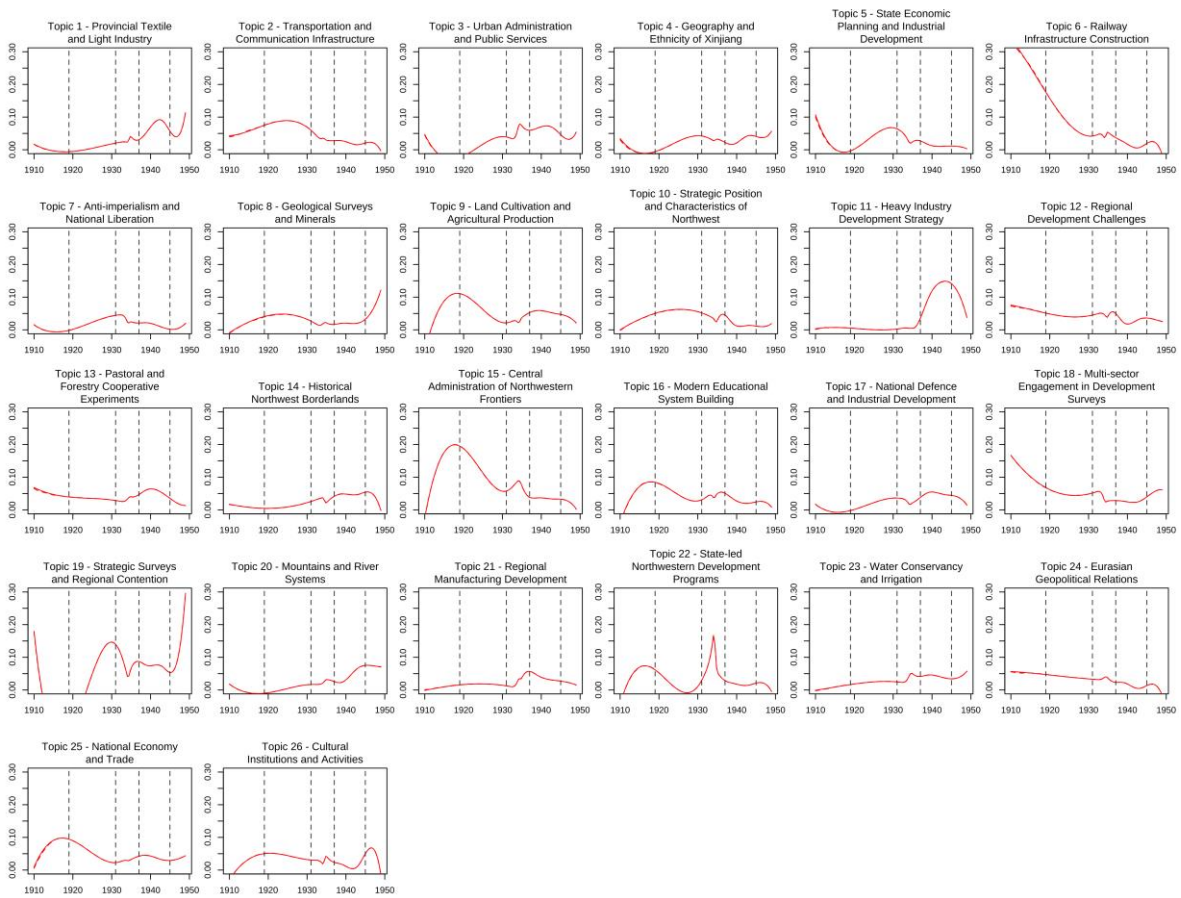


Figure 7: Temporal Dynamics of Topics

The temporal analysis reveals several distinctive patterns. First, discussions of Northwest development were already gaining momentum in the 1920s, largely driven by internal political dynamics. The pronounced peaks in topics such as Land Cultivation and Agricultural Production (9), Central Administration of Northwestern Frontiers (15), and State-led Northwestern Development Programs (22) during the early to mid-1920s correspond to the consolidation of power by the warlord Feng Yuxiang and his Guominjun (Nationalist Army) in the region (Sheridan, 1966, pp. 149–158). Feng, whose name appears as a representative term in Topic 15, initiated numerous state-building and modernisation projects aimed at strengthening his

territorial base. The discourse from this period reflects these top-down efforts at administrative control and economic development, long before the Japanese threat became the primary national concern.

Second, the political landscape of the early 1930s was defined by multiple, overlapping crises that reshaped the discourse. The 1931 Manchurian Incident undoubtedly amplified security-centred rhetoric. However, the period was also marked by the aftermath of the Central Plains War (1930), which led to Feng Yuxiang's defeat and created a power vacuum in the Northwest. The sharp peak in Strategic Surveys and Regional Contention (19) around this time reflects this new reality. The representative terms for this topic include not only academic bodies but also the names of military figures like Sun Dianying and Fu Zuoyi, who moved into the region to establish their own spheres of influence. Thus, the surge in 'surveys' and 'investigations' was not purely academic; it was deeply intertwined with the strategic assessments of new political-military actors vying for control.

Simultaneously, the geopolitical threat from the north remained a potent undercurrent. The Russian Empire and later the Soviet Union's establishment of a proxy state in Outer Mongolia in the 1920s and its direct military intervention in Xinjiang in 1934 heightened anxieties about border security (Cordier, 2016; Paine, 1996). The sustained presence of Eurasian Geopolitical Relations (24), with terms like 'Moscow' and 'Siberia', and Anti-imperialism (7) throughout this period indicates that the 'imperialist' threat was seen as multi-faceted, emanating from both the Soviet Union and Japan.

Third, the outbreak of the full-scale Second Sino-Japanese War in 1937 marked a dramatic inflection point, transforming the language of Northwest development from one of peacetime opportunity to one of wartime necessity. As coastal regions fell to Japan, the discourse

shifted decisively towards the interior. For example, Heavy Industry Development Strategy (11) remains relatively low in prominence through the 1920s and early 1930s but surges after 1937, peaking in the early 1940s. This spike aligns with the wartime effort to relocate industry to the interior. Similarly, National Defense and Industrial Development (17) shows a two-step jump: an increase after 1931 and an even larger one after 1937. The war crystallised the Northwest's role as a strategic rear base, a theme that came to dominate all other developmental concerns.

Fourth, the analysis also reveals interesting temporal relationships between related topics. For instance, while State Economic Planning (5) shows increased prevalence around 1931, Provincial Textile and Light Industry (1) only gains significant attention after 1937, suggesting a shift from central planning to concrete industrial implementation. The Strategic Surveys and Regional Contention (19) shows notable peaks around both 1931 and 1945, indicating how academic attention responded to national crises by intensifying research on the Northwest's strategic potential. Particularly noteworthy is the temporal pattern of topics related to resource development. Geological Surveys and Minerals (8) shows early attention in the 1920s that diminishes after 1931, while Water Conservancy and Irrigation (23) gains increased attention after 1937. This shift suggests an evolution from basic resource surveying to more applied development projects as the Northwest's strategic importance grew.

Fifth, the overarching before-and-after effect of World War II on the discourse. Before the mid-1930s, discussions of the Northwest often framed it as a land of future promise—emphasising surveys, initial infrastructure (like scouting rail routes), migration schemes for Han settlers, and so on. After 1937, the tone shifts to one of immediate exigency: the Northwest must be developed *now* as part of the war effort—build factories, move universities and industries inland, secure grain and minerals for the front. Our topic prevalence graphs quantitatively

capture this comprehensive transformation of priorities. Topics related to surveying and long-term planning either plateau or decline after 1937, while those tied to urgent utilisation and militarisation spike sharply. By war's end in 1945, most development talk recedes significantly. This drop-off likely reflects war weariness and the Nationalist government's waning capacity; indeed, after 1945 China was immediately embroiled in civil war, and Northwest development was deprioritised amid more pressing national struggles. The post-1945 decline in topic prevalence suggests that the idea of the Northwest as a top-tier national concern was somewhat a wartime product—once the external threat lessened, attention shifted away until the PRC would later revive frontier development under a different paradigm.

In sum, viewing the Republican-era Northwest discourse chronologically reveals that it was not a single, steady trajectory but a series of phases shaped by a complex interplay of internal power struggles, long-term geopolitical pressures from the north, and the escalating crisis with Japan. The elites and media of the time maintained a versatile narrative that could pivot as circumstances changed. In the 1920s, the Northwest was an *autonomous frontier* to be consolidated and modernised; by the early 1930s, it was a *contested frontier* in a multi-front crisis; and by the 1940s, it had become a *strategic heartland* to be fortified and mobilised for national survival. These findings demonstrate how discourse analysis can illuminate the reciprocal relationship between historical context and rhetoric: what was said (topics), and how much it was said (topic proportions over time), was deeply intertwined with Republican China's changing fortunes.

5. Conclusion

The development discourse surrounding China's Northwest frontier in the Republican era offers a revealing window into the priorities and predicaments of the modern Chinese state. Our

analysis, combining digital methods and historical interpretation, paints a big picture of a multifaceted national project that evolved significantly from the 1910s to the 1940s. Rather than a monolithic narrative, the Northwest development discourse comprised multiple parallel threads—economic, administrative, cultural, strategic—each rising or falling in prominence with the tides of history. By examining thousands of contemporaneous texts, we showed how government officials, intellectuals, and media outlets conceptualised the northwestern frontier in varying ways: as a storehouse of resources, a target for state building, a home to be integrated, and a strategic redoubt in times of war.

Answering our research questions, in brief: the discourse resolves into two interlinked constellations—a governance-industry core anchored in planning rhetoric and an infrastructure-resources bloc organised around transport, hydraulic, and extraction schemes—with cultural reportage and educational campaigns circulating at the rim. Planning narratives functioned as the hinge that folded specialist agendas into national stories, while security arguments bridged administrative and industrial programmes, showing how anxieties about external threats knit disparate policy domains together. Temporally, attention swings from the consolidation drives of the early-to-mid 1920s, through the survey-heavy scramble that followed the Central Plains War, to the wartime mobilisation after 1937 when defence, heavy industry, and water projects surged before receding once the crisis abated in 1945.

Our evidence pinpoints how press discourse transformed the Northwest: by elevating strategic geography and state planning as hubs that bound sectoral projects into a national storyline, with surges at identifiable moments. These findings reinforce recent scholarship arguing that the frontier was pivotal in the formation of the modern Chinese state (Wu, 2023b). The Northwest was both a laboratory and a proving ground for nationalist statecraft: plans made

there prefigured strategies later applied across other regions. Crucially, the Second Sino-Japanese War acted as a catalyst that forced the state to accelerate and expand its efforts in the interior. What had been a somewhat idealised vision in the 1920s turned into a concrete wartime imperative by the late 1930s—an extreme case of what we might call developmental geopolitics, where international crisis drives internal development agendas. In this sense, the Republican Northwest initiative foreshadowed aspects of the PRC's later approach to frontier development under new geopolitical pressures (such as the Third Front Movement).

Rather than restating that the Northwest 'mattered', our contribution is to specify how discourse made it matter: Planning-oriented networks absorbed sectoral programmes into a governance-industrial core, security rhetoric stitched administrative and productive agendas together, and successive crises from the mid-1920s through the Second Sino-Japanese War intensified those linkages at recognisable inflection points. These patterns arise from corpus-wide co-occurrence and year-by-topic estimates. We therefore characterise our findings as evidence of frequent, historically contingent reframing of technical initiatives within state-building and security narratives in the press discourse.

Methodologically, our study demonstrates the value of digital humanities approaches for historical inquiry. By leveraging a large corpus of texts and employing structural topic modelling, we were able to synthesise a vast array of discourse that no single researcher could read in full. This inductive, data-driven technique allowed us to detect patterns and themes that complement the findings of traditional scholarship. In contrast to existing studies that carefully hand-pick sources and focus on specific aspects of Northwest development, we utilised a method better suited to the scale of digital archives. The combination of an LLM-assisted OCR pipeline and structural topic modelling enabled a scalable and reproducible analysis, showing how AI

technologies can transform access to and analysis of historical materials. At the same time, we integrated domain knowledge throughout the analysis—validating topics by reading original articles and interpreting them in their historical context—illustrating that computational methods and humanistic interpretation work best in tandem.

5.1 Limitations and Future Research

While our approach yields valuable insights, several limitations warrant acknowledgment and suggest directions for future research. Despite using state-of-the-art LLM-based OCR, residual extraction errors and orthographic variation likely introduce noise that can blur low-frequency themes. Data availability also constrained coverage: some CNBKSY entries lacked full text or were too poorly scanned to recover, which may create gaps if the missing documents systematically differed from those we could read. As newer LLM-based OCR models emerge, such as Mistral OCR (Mistral AI Team, 2025) and olmOCR (Poznanski et al. 2025), improved accuracy could further benefit historical text analysis.

The press corpus necessarily underrepresents frontier populations' own voices, including Hui, Uyghur, Mongol, and Tibetan. We therefore read portrayals of minorities as representations by others and flag this asymmetry as a priority for future mixed-source studies. Promising avenues include incorporating minority-language periodicals, local newspapers, gazetteers, and administrative reports, and combining textual analysis with spatial data to add a geospatial dimension. Extending the temporal scope to the post-1949 socialist period would further allow comparison of how these frames persisted or transformed under different institutional regimes.

Beyond these constraints, two extensions are especially promising. Our analysis focused on themes rather than evaluation; augmenting it with sentiment or stance detection tailored to historical Chinese could better gauge attitudes embedded in the discourse (Chen and Mankad,

2025). In parallel, named-entity recognition and network analysis could map connections among key figures and institutions, clarifying which actors bridge policy arenas. Together, these steps would complement the topic-based analysis by addressing questions of evaluation, influence, and implementation.

Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

Informed consent

This article does not contain any studies with human participants performed by any of the authors.

Competing interests

The author declares no competing interests.

Funding statement

This work has received no external funding.

Data availability

Data and analysis are available at: https://github.com/ghuserone/develop-northwest. The repository includes shareable code and the processed corpus used for analysis for peer review and editorial verification; the underlying CNBKSY and *Shenbao* source scans are subject to database licensing and cannot be redistributed.

References

Abraham A et al. (2024) Themes and trends in creativity research between 1894 and 2022: A topic modeling approach. Psychol Aesthet Creat Arts. https://doi.org/10.1037/aca0000677

Anderson B (2006) Imagined communities: Reflections on the origin and spread of nationalism. Revised Edition. Verso, London

Armand C and Henriot C (2023) Beyond digital humanities thinking computationally: A position paper. https://shs.hal.science/halshs-04194570. Accessed 3 Jul 2024

Assael Y et al. (2022) Restoring and attributing ancient texts using deep neural networks. Nature 603(7900):280–283. https://doi.org/10.1038/s41586-022-04448-z

Baker M (2024) Energy, labor, and Soviet aid: China's Northwest Highway, 1937–1941. Mod China 50(3):302–334. https://doi.org/10.1177/00977004231203897

Beelen K et al. (2025) Whose news? Critical methods for assessing bias in large historical datasets. Comput Humanit Res 1:e8. https://doi.org/10.1017/chr.2025.10007

Benoit K (2020) Text as data: An overview. In: L Curini and R Franzese (eds) The SAGE handbook of research methods in political science and international relations. SAGE Publications, London, pp. 461–497. https://doi.org/10.4135/9781526486387.n29

Blei DM (2012) Probabilistic topic models. Commun ACM 55(4):77–84. https://doi.org/10.1145/2133806.2133826

Blei DM, Ng AY and Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3(Jan):993–1022

Blouin B et al. (2023) Unlocking transitional Chinese: Word segmentation in modern historical texts. In: Proceedings of the joint 3rd international conference on natural language processing for digital humanities and 8th international workshop on computational linguistics for uralic languages. Association for Computational Linguistics, Tokyo, pp. 92–101. https://aclanthology.org/2023.nlp4dh-1.11. Accessed 16 Nov 2024

Cheek T (2015) The intellectual in modern Chinese history. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781139108874

Chen B, Anker TB and Liang X (2025) Business continuity management in the sharing economy: Insights from Airbnb online reviews. Tour Manag 107:105067. https://doi.org/10.1016/j.tourman.2024.105067

Chen L and Mankad S (2025) A structural topic and sentiment-discourse model for text analysis. Manage Sci 71(7):5767–5787. https://doi.org/10.1287/mnsc.2022.00261

Chen Y et al. (2025) Geocoding the past world: Unearthing coordinates of early China from texts using generative AI. Int J Geogr Inf Sci:1–27. https://doi.org/10.1080/13658816.2025.2491711

Cheng L, Wang F and Zhang W (2007) Zhongguo jindai kaifa xibu de sixiang yu zhengce yanjiu (中国近代开发西部的思想与政策研究). Shanghai renmin chubanshe, Shanghai

CHGIS (2012) CHGIS V5 shapefiles. Harvard Dataverse. https://doi.org/10.7910/DVN/M7WEFY

Chiu S-H et al. (2025) Studying tech adoption with 'text-as-data': Opportunities, pitfalls, and complementarities in the case of transportation. Environ Plann B: Urban Anal City Sci 52(8):1796–1813. https://doi.org/10.1177/23998083241311039

Chow EHC (2024) An experiment with Gemini Pro LLM for Chinese OCR and metadata extraction. The Digital Orientalist, 5 April. https://digitalorientalist.com/2024/04/05/an-experiment-with-gemini-pro-llm-for-chinese-ocr-and-metadata-extraction. Accessed 21 Oct 2024

Chuangkanci (1934) (創刊詞). Kaifa Xibei 1(1):1–3

Cordier BD (2016) International aid, frontier securitization, and social engineering: Soviet–Xinjiang development cooperation during the governorate of Sheng Shicai (1933–1944). Cent Asian Aff 3(1):49–76. https://doi.org/10.1163/22142290-00301003

Cui N et al. (2025) Using Twitter to understand spatial-temporal changes in urban green space topics based on structural topic modelling. Cities 157:105601. https://doi.org/10.1016/j.cities.2024.105601

Dagongbao (1932) Lun Xibei jianshe (論西北建設), 26 April:2

Debnath R et al. (2020) Grounded reality meets machine learning: A deep-narrative analysis framework for energy policy research. Energy Res Social Sci 69:101704. https://doi.org/10.1016/j.erss.2020.101704

Dikötter F (2015) The discourse of race in modern China. Fully revised and expanded second edition. Oxford University Press, New York

DiMaggio P, Nag M and Blei D (2013) Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. Poetics 41(6):570–606. https://doi.org/10.1016/j.poetic.2013.08.004

Dong Z et al. (eds) (1998) Xibei kaifa shiliao xuanji (1930–1947) (西北开发史料选辑（1930－1947）). Jingji keji chubanshe, Beijing

Eberle O et al. (2024) Historical insights at scale: A corpus-wide machine learning analysis of early modern astronomic tables. Sci Adv 10(43):eadj1719. https://doi.org/10.1126/sciadv.adj1719

Elliott M (2014) Frontier stories: Periphery as center in Qing history. Front Hist China 9(3):336–360. https://doi.org/10.3868/s020-003-014-0025-1

Fairbank JK (1968) A preliminary framework. In: JK Fairbank (ed) The Chinese world order: Traditional China's foreign relations. Harvard University Press, Cambridge, MA, pp. 1–19. https://doi.org/10.4159/harvard.9780674333482.c3

Fakanci (1936) (發刊詞). Bianjiang 1(1):1–2

Filimonov S (2025) Ingesting millions of PDFs and why Gemini 2.0 changes everything, 15 January. https://www.sergey.fyi/articles/gemini-flash-2. Accessed 2 Apr 2025

Fogel RW and Engerman SL (1974) Time on the cross: The economics of American negro slavery. Little Brown, New York

Forbes ADW (1986) Warlords and Muslims in Chinese Central Asia: A political history of Republican Sinkiang 1911–1949. Cambridge University Press, Cambridge

Gavriș A and Popescu C (2024) Encounters of hesitant politics and an unwavering energy transition. Media reflections in Romania. J Cleaner Prod 478:143870. https://doi.org/10.1016/j.jclepro.2024.143870

Ge Z (2011) Zhai zi Zhongguo: Chongjian youguan 'Zhongguo' de lishi lunshu (宅兹中国：重建有关「中国」的历史论述). Zhonghua shuju, Beijing

Gemini Team (2024) Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. https://doi.org/10.48550/arXiv.2403.05530

Gentzkow M, Kelly B and Taddy M (2019) Text as data. J Econ Lit 57(3):535–574. https://doi.org/10.1257/jel.20181020

Gilkison A and Kurzynski M (2024) Vectors of violence: Legitimation and distribution of state power in the *People's Liberation Army Daily* (*Jiefangjun Bao*), 1956–1989. J Cult Anal 9(1). https://doi.org/10.22148/001c.115481

Greitens SC and Truex R (2020) Repressive experiences among China scholars: New evidence from survey data. China Q 242:349–375. https://doi.org/10.1017/S0305741019000365

Grimmer J, Roberts ME and Stewart BM (2021) Machine learning for social science: An agnostic approach. Annu Rev Polit Sci 24:395–419. https://doi.org/10.1146/annurev-polisci-053119-015921

Grimmer J, Roberts ME and Stewart BM (2022) Text as data: A new framework for machine learning and the social sciences. Princeton University Press, Princeton

Grimmer J and Stewart BM (2013) Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Polit Anal 21(3):267–297. https://doi.org/10.1093/pan/mps028

Grootendorst M (2022) BERTopic: Neural topic modeling with a class-based TF-IDF procedure. https://doi.org/10.48550/arXiv.2203.05794

Guldi J (2023) The dangerous art of text mining: A methodology for digital history. Cambridge University Press, Cambridge. https://doi.org/10.1017/9781009263016

Ho P (1967) The significance of the Ch'ing period in Chinese history. J Asian Stud 26(2):189–195. https://doi.org/10.2307/2051924

Hong Z and Chen Y (2024) Persuading the emperors: A quantitative historical analysis of political rhetoric in traditional China. Humanit Soc Sci Commun 11(1):840. https://doi.org/10.1057/s41599-024-03164-5

Hou Y and Huang J (2025) Natural language processing for social science research: A comprehensive review. Chin J Sociol 11(1):121–157. https://doi.org/10.1177/2057150X241306780

Hu J (2020) 20 shiji sanshi niandai Xibei kaifa zhongde gaodeng jiaoyu wenti (20 世纪三四十年代西北开发中的高等教育问题). Zhongguo shehuikexue chubanshe, Beijing

Hu S (1985) Xibeixue chuyi (西北学刍议). Xibei minzu daxue xuebao (zhexue shehuikexue ban) (1):26–34, 25

Jia X and Hua D (2002) *Dagongbao* yu 1930 niandai de Xibei kaifa (《大公报》与 20 世纪 30 年代西北开发). Xibei gongye daxue xuebao (shehuikexue ban) (2):7–13. https://doi.org/10.3969/j.issn.1009-2447.2002.02.003

Jiang J (2001) Weida de Xibei (伟大的西北). Ningxia renmin chubanshe, Yinchuan

Lim J, Ito A and Zhang H (2025) Uncovering Xi Jinping's policy agenda: Text as data approach. Dev Econ 63(1):9–46. https://doi.org/10.1111/deve.12418

Lin H (2011) Modern China's ethnic frontiers: A journey to the west. Routledge, London. https://doi.org/10.4324/9780203844977

Lipman JN (1997) Familiar strangers: A history of Muslims in Northwest China. University of Washington Press, Seattle. https://doi.org/10.6069/9780295800554

Liu X (2011) Bianjiang Zhongguo he 1949 nian (边疆中国和 1949 年). In: G Han (ed) Zhongguo dangdaishi yanjiu (san). Jiuzhou chubanshe, Beijing, pp. 117–136

Mackinnon SR (1997) Toward a history of the Chinese press in the Republican period. Mod China 23(1):3–32. https://doi.org/10.1177/009770049702300101

Mancall M (1968) The Ch'ing tribute system: An interpretive essay. In: JK Fairbank (ed) The Chinese world order: Traditional China's foreign relations. Harvard University Press, Cambridge, MA, pp. 63–89. https://doi.org/10.4159/harvard.9780674333482.c6

Mann M (1984) The autonomous power of the state: Its origins, mechanisms and results. Eur J Sociol 25(2):185–213. https://doi.org/10.1017/S0003975600004239

Matten MA (2016) Imagining a postnational world: Hegemony and space in modern China. Brill, Leiden. https://doi.org/10.1163/9789004327153

Mennig P (2025) Who cares about agriculture? Analyzing German parliamentary debates on agriculture and food with structural topic modeling. Food Policy 130:102788. https://doi.org/10.1016/j.foodpol.2024.102788

Mertha A (ed) (2024) Studying China in the absence of access: Rediscovering a lost art. SAIS China Research Center. https://scgrc.sais.jhu.edu/wp-content/uploads/2024/10/32026_JOHNS_HOPKINS.COVER_SP.pdf. Accessed 11 Dec 2024

Miller IM (2013) Rebellion, crime and violence in Qing China, 1722–1911: A topic modeling approach. Poetics 41(6):626–649. https://doi.org/10.1016/j.poetic.2013.06.005

Milligan I (2019) History in the age of abundance? How the web is transforming historical research. McGill-Queen's University Press, Montreal. https://doi.org/10.1515/9780773558212

Mistral AI Team (2025) Mistral OCR. Mistral AI. https://mistral.ai/news/mistral-ocr. Accessed 18 Mar 2025

Mittler B (2004) A newspaper for China?: Power, identity, and change in Shanghai's news media, 1872–1912. Harvard University Asia Center, Cambridge, MA. https://doi.org/10.1163/9781684173884

Morandell T, Wicki M and Kaufmann D (2025) The planning of urban–rural linkages: An automated content analysis of spatial plans adopted by European intermediate cities. Landscape Urban Plann 255:105258. https://doi.org/10.1016/j.landurbplan.2024.105258

Nelson LK et al. (2021) The future of coding: A comparison of hand-coding and three types of computer-assisted text analysis methods. Sociol Methods Res 50(1):202–237. https://doi.org/10.1177/0049124118769114

Newby LJ (1999) The Chinese literary conquest of Xinjiang. Mod China 25(4):451–474. https://doi.org/10.1177/009770049902500403

Newman DJ and Block S (2006) Probabilistic topic decomposition of an eighteenth-century American newspaper. J Am Soc Inf Sci Technol 57(6):753–767. https://doi.org/10.1002/asi.20342

Ni X (1936) Xijing (西京). Zhonghua shuju, Shanghai

Nian Y and Lin T (2019) 20 shiji 30 niandai Xibei youji zhong de kongjian jiangou yu zhengzhi rentong (20 世纪 30 年代西北游记中的空间建构与政治认同). Hunan shifan daxue shehuikexue xuebao 48(2):96–101. https://doi.org/10.19503/j.cnki.1000-2529.2019.02.012

Northrop K (2022) Open source. The Wire China. https://www.thewirechina.com/2022/01/16/open-source. Accessed 29 Dec 2024

Paine SCM (1996) Imperial rivals: China, Russia, and their disputed frontier. M.E. Sharpe, Armonk

Pelzer T (2025) Engineers on the move: Elite geographic mobility in Republican China. Twent-Century China 50(1):25–55. https://doi.org/10.1353/tcc.2025.a950426

Piao Y (1932) Xibei kaifa yundong de xin zhankai (西北開發運動的新展開). Chulu xunkan 1(3):6–9

Poznanski J et al. (2025) olmOCR: Unlocking trillions of tokens in PDFs with vision language models. https://doi.org/10.48550/arXiv.2502.18443

Qinggaozong (ed) (1935) Qingchao tongdian (清朝通典). Shangwu yinshuguan, Shanghai

R Core Team (2025) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org

Radford A et al. (2022) Robust speech recognition via large-scale weak supervision. https://doi.org/10.48550/arXiv.2212.04356

Ramos-Henriquez JM and Morini-Marrero S (2025) Airbnb customer experience in long-term stays: A structural topic model and ChatGPT-driven analysis of the reviews of remote workers. Int J Contemp Hosp Manag 37(1):161–179. https://doi.org/10.1108/IJCHM-01-2024-0034

Roberts ME et al. (2014) Structural topic models for open-ended survey responses. Am J Polit Sci 58(4):1064–1082. https://doi.org/10.1111/ajps.12103

Roberts ME, Stewart BM and Airoldi EM (2016) A model of text for experimentation in the social sciences. J Am Stat Assoc 111(515):988–1003. https://doi.org/10.1080/01621459.2016.1141684

Roberts ME, Stewart BM and Tingley D (2019) stm: An R package for structural topic models. J Stat Softw 91(2):1–40. https://doi.org/10.18637/jss.v091.i02

Rosenzweig R (2003) Scarcity or abundance? Preserving the past in a digital era. Am Hist Rev 108(3):735–762. https://doi.org/10.1086/ahr/108.3.735

Şakar S and Tan S (2025) Research topics and trends in gifted education: A structural topic model. Gift Child Q 69(1):68–84. https://doi.org/10.1177/00169862241285041

Schmiedel T, Müller O and vom Brocke J (2019) Topic modeling as a strategy of inquiry in organizational research: A tutorial with an application example on organizational culture. Organ Res Methods 22(4):941–968. https://doi.org/10.1177/1094428118773858

Scott JC (1998) Seeing like a state: How certain schemes to improve the human condition have failed. Yale University Press, New Haven. https://doi.org/10.12987/9780300128789

Shang J and Ding X (2023) Xibei guojie tongdao de kapi yu zhonghuaminzu gongtongti yishi de rongzhu: Yi Xinjiang minzhong zuliu yundong wei zhongxin (西北国际通道的开辟与中华民族共同体意识的熔铸——以抗战时期新疆民众筑路运动为中心). Zhongzhou daxue xuebao 40(3):68–76. https://doi.org/10.13783/j.cnki.cn41-1275/g4.2023.03.011

Shen S (2006) Jiangshan ruci duojiao: 1930 niandai de Xibei lüxing shuxie yu guozu xiangxiang (江山如此多嬌——1930 年代的西北旅行書寫與國族想像). Taida lishi xuebao (37):145–216. https://doi.org/10.6253/ntuhistory.2006.37.03

Shen X (2007) Kangri zhanzheng shiqi guominzhengfu de Xibei kaifa (抗日战争时期国民政府的西北开发). Zhejiang daxue xuebao (renwen shehuikexue ban) 37(5):104–113. https://doi.org/10.3785/j.issn.1008-942X.2007.05.015

Sheridan JE (1966) Chinese warlord: The career of Feng Yü-hsiang. Stanford University Press, Stanford

Short JC, McKenny AF and Reid SW (2018) More than words? Computer-aided text analysis in organizational behavior and psychology research. Annu Rev Organ Psychol Organ Behav 5:415–435. https://doi.org/10.1146/annurev-orgpsych-032117-104622

Sun Y (2021) The international development of China: A project to assist the readjustment of post-bellum industries. Springer, Singapore. https://doi.org/10.1007/978-981-16-0961-9

Tai J (2015) The Northwest question: Capitalism in the sands of nationalist China. Twent-Century China 40(3):201–219. https://doi.org/10.1179/1521538515Z.00000000066

Tang Y-K et al. (2025) Bridging insight gaps in topic dependency discovery with a knowledge-inspired topic model. Inf Process Manage 62(1):103911. https://doi.org/10.1016/j.ipm.2024.103911

Tian S (ed) (2007) Xibei kaifa shi yanjiu (西北开发史研究). Zhongguo shehuikexue chubanshe, Beijing

Tighe J (2005) Constructing Suiyuan: The politics of northwestern territory and development in early twentieth-century China. Brill, Leiden. https://doi.org/10.1163/9789047407881

Tighe J (2009) From borderland to heartland: The discourse of the North-West in early Republican China. Twent-Century China 35(1):54–74. https://doi.org/10.1179/tcc.2009.35.1.54

Todorov K and Colavizza G (2022) An assessment of the impact of OCR noise on language models. https://doi.org/10.48550/arXiv.2202.00470

Tonidandel S et al. (2022) Using structural topic modeling to gain insight into challenges faced by leaders. Leadersh Q 33(5):101576. https://doi.org/10.1016/j.leaqua.2021.101576

Underwood T (2019) Distant horizons: Digital evidence and literary change. University of Chicago Press, Chicago. https://doi.org/10.7208/chicago/9780226612973.001.0001

Underwood T (2025) The impact of language models on the humanities and vice versa. Nat Comput Sci:1–3. https://doi.org/10.1038/s43588-025-00819-4

Veg S (2021) Creating public opinion, advancing knowledge, engaging in politics: The local public sphere in Chengdu, 1898–1921. China Q 246:331–353. https://doi.org/10.1017/S0305741021000217

Viola L and Verheul J (2020) Mining ethnicity: Discourse-driven topic modelling of immigrant discourses in the USA, 1898–1920. Digit Scholarsh Humanit 35(4):921–943. https://doi.org/10.1093/llc/fqz068

Wang R (2010) Nanjing guominzhengfu shangceng renshi yu 'Xibei kaifa' (南京国民政府上层人士与‘西北开发’). Xibei nonglin keji daxue xuebao (shehuikexue ban) 10(3):136–140. https://doi.org/10.13968/j.cnki.1009-9107.2010.03.001

Wang R (2015) Weiji xiade zhuanji: Guominzhengfu shiqi de Xibei jingji kaifa yanjiu (危机下的转机：国民政府时期的西北经济开发研究). Zhongguo shehuikexue chubanshe, Beijing

Wang Z (1943) Xibei jianshe lun (西北建設論). Qingnian chubanshe, Chongqing

Wencker T, Borst-Graetz J and Niekler A (2025) Text as data for evaluation: Natural language processing and large language models to generate novel insights from unstructured text data. Evaluation 31(3):369–393. https://doi.org/10.1177/13563890251330911

Weng J (2023) Stop the presses! Publishing Chinese character simplification, 1935–1936. Harv J Asiat Stud 83(2):333–364. https://doi.org/10.1353/jas.2023.a938222

Weston SJ et al. (2023) Selecting the number and labels of topics in topic modeling: A tutorial. Adv Methods Pract Psychol Sci 6(2). https://doi.org/10.1177/25152459231160105

Weston TB (2010) China, professional journalism, and liberal internationalism in the era of the First World War. Pac Aff 83(2):327–347. https://doi.org/10.5509/2010832327

Wilkerson J and Casas A (2017) Large-scale computerized text analysis in political science: Opportunities and challenges. Annu Rev Polit Sci 20:529–544. https://doi.org/10.1146/annurev-polisci-052615-025542

Wu R (2023a) The making of 'public opinion': Media and open diplomacy in China's strategy at Versailles and the May Fourth Movement. Mod Asian Stud 57(4):1355–1386. https://doi.org/10.1017/S0026749X22000609

Wu SX (2015) Empires of coal: Fueling China's entry into the modern world order, 1860–1920. Stanford University Press, Stanford. https://doi.org/10.1515/9780804794732

Wu SX (2023b) Birth of the geopolitical age: Global frontiers and the making of modern China. Stanford University Press, Stanford. https://doi.org/10.1515/9781503636859

Xia S (2024) Fandom culture as a catalyst for propaganda. China Q 259:814–823. https://doi.org/10.1017/S0305741023001650

Xiang H (2018) Rechao, shijian yu kunjing: Kangzhan qian Xibei kaifa de zai shenshi (1928–1937) (热潮、实践与困境：抗战前西北开发的再审视（1928–1937）). Jindai Zhongguo (2):264–292

Yan D and Zhang L (2006) Minguo 'kaifa Xibei' zhong yici weijun de yimin jihua: 1942 nian zhi 1944 nian de Xinjiang yimin (民国'开发西北'中一次未竣的移民计划——1942 年至 1944 年的新疆移民). Minguo dangan (3):105–112

Yang H (2013) Kangzhan shiqi Xibei jingji kaifa sixiang yanjiu (抗战时期西北经济开发思想研究). Zhongguo shehuikexue chubanshe, Beijing

Ying L, Montgomery JM and Stewart BM (2022) Topics, concepts, and measurement: A crowdsourced procedure for validating topics as measures. Polit Anal 30(4):570–589. https://doi.org/10.1017/pan.2021.33

You S (1936) Xibei zhi jiaotong xiankuang jiqi jianshe (西北之交通現況及其建設). Luxiang 3(8):231–236

Yu T (1929) Fakanci (發刊詞). Xin Xibei (1):7

Zaagsma G (2023) Digital history and the politics of digitization. Digit Scholarsh Humanit 38(2):830–851. https://doi.org/10.1093/llc/fqac050

Zeng W (1936) Zhongguo jingying Xiyu shi (中國經營西域史). Shangwu yinshuguan, Shanghai

Zhang H (1934a) Xiyou xiaoji (yi) (西游小記（一）). Lüxing zazhi 8(9):7–10

Zhang L (1989) Jindai guoren de kaifa Xibei guan (近代國人的開發西北觀). Jindaishi yanjiusuo jikan (18):163–188. https://doi.org/10.6353/BIMHAS.198906.0163

Zhang R (1934b) Kaifa Xibei shiye jihua (開發西北實業計劃). Zhuzhe shudian, Beiping

Zhang X (2021) Xibei kaocha yu guozu xiangxiang (西北考察与国族想象). Dissertation, Nanjing daxue. https://doi.org/10.27235/d.cnki.gnjiu.2021.001356

Zhang Y (2002) Kangzhan qianshinian guominzhengfu kaifa Xibei de zhengce quxiang (抗战前十年国民政府开发西北的政策取向). Sichuan daxue xuebao (zhexue shehuikexue ban) (5):121–128. https://doi.org/10.3969/j.issn.1006-0766.2002.05.019

Zhao J (2008) Fenjie yu chonggou: Qingji minchu de baojie tuanti (分解与重构：清季民初的报界团体). Shenghuo, dushu, xinzhi sanlian shudian, Beijing

Zhao S (2004) A nation-state by construction: Dynamics of modern Chinese nationalism. Stanford University Press, Stanford. https://doi.org/10.1515/9781503624498

Zhao X (2025) Running a mainstream revolutionary newspaper: *Guangdong Qunbao* and socialist propaganda in 1920s South China. Labor Hist 66(3):417–429. https://doi.org/10.1080/0023656X.2024.2383968

Zhong Y (2019) Chinese grammatology: Script revolution and literary modernity, 1916–1958. Columbia University Press, New York. https://doi.org/10.7312/zhon19262

Zhou Y (2006) Historicizing online politics: Telegraphy, the internet, and political participation in China. Stanford University Press, Stanford. https://doi.org/10.1515/9780804767583

Zong Y (2003) 20 shiji 30 niandai baokan meijie yu Xibei kaifa (20 世纪 30 年代报刊媒介与西北开发). Shixue yuekan (5):54–58. https://doi.org/10.3969/j.issn.0583-0214.2003.05.008