A Critical Review of Gujarati Language in the Digital Age: Emerging Practices and Concerns

Amit Bag
Kolkata, India.
Email: amit18.sdi@gmail.com

Abstract

Gujarati, an Indo-Aryan language spoken across western India and a substantial global diaspora, has undergone marked transformation as speakers have moved communication, literacy, and cultural expression onto digital platforms. This review synthesises scholarship on the linguistic, technological, and social dimensions of this shift, drawing on work in sociolinguistics, computer-mediated communication, natural language processing, and digital inequality research. The paper traces how script rendering, romanisation, and transliteration have shaped everyday Gujarati typing practices; how code-mixing and code-switching with English and Hindi have become normalised in messaging and social media; and how uneven access along caste, gender, and regional lines continues to condition who participates in Gujarati's digital life. It further considers the diaspora's role in sustaining Gujarati online, the state of language-technology resources such as morphological analysers and machine translation systems, the vulnerability of regional-language audiences to misinformation, and policy efforts, including India's National Education Policy 2020, to support mother-tongue digital literacy. The review finds that Gujarati's digital presence is expanding rapidly in volume while remaining thin in infrastructural depth, producing a paradox of visibility without full technological or educational support. Persistent gaps include the scarcity of annotated Gujarati corpora, limited fact-checking capacity in the language, and unequal digital access among women and economically disadvantaged groups. The paper closes by identifying directions for future research and practice, and by outlining the limitations of the present narrative synthesis.

Keywords: *Gujarati language, digital communication, computer-mediated communication, code-switching, language technology, digital divide, South Asia, script and identity*

1. Introduction

1.1 Gujarati in its Sociolinguistic Setting

Gujarati belongs to the Indo-Aryan branch of the Indo-European language family and is written in a script historically derived from Devanagari, from which it differs chiefly in lacking the continuous horizontal head-stroke that joins letters in the parent script. Its dialectology, including variation retained among long-settled diaspora communities, has been documented in detail for the Gujarati spoken in Cape Town, South Africa, where speakers show a distinctive prominence of retroflex over dental consonants relative to homeland varieties (Mesthrie & Chavda, 2020). The language is the primary medium of public and domestic life in the Indian state of Gujarat and carries official status alongside Hindi and English in several neighbouring union territories. Beyond the subcontinent, Gujarati has long travelled with a mercantile and professional diaspora to East Africa, the United Kingdom, North America, and Southern Africa, producing exactly the kind of dialectal divergence documented in the Cape Town case (Mesthrie & Chavda, 2020). This combination of a large home population and a geographically dispersed diaspora means that Gujarati's digital life is never confined to a single national

internet; it is instead distributed across servers, applications, and social networks that span continents and regulatory regimes.

The language's contact history is equally relevant to its digital present. Centuries of exchange with Persian, Arabic, Portuguese, and English left lexical and stylistic traces that made Gujarati speakers, particularly in trading and professional classes, comparatively receptive to bilingual and multilingual repertoires (Mesthrie & Chavda, 2020). This receptivity has carried into digital communication, where English supplies much of the technical, professional, and youth-culture vocabulary that circulates alongside Gujarati script and speech.

1.2 The Digital Turn in South Asian Language Ecologies

The expansion of low-cost smartphones, cheap mobile data, and Unicode-compliant Indic-script rendering has reshaped how Indian-language speakers, including Gujarati speakers, encounter and produce written and spoken text online. Broader South Asian scholarship on computer-mediated communication (CMC) has documented how sociolinguistic method, once developed for face-to-face interaction, has been extended to online settings, revealing that language choice, script choice, and code-switching in digital spaces are neither random nor purely technologically determined but are socially patterned practices (Androutsopoulos, 2006a; Herring, 2007). This extension of method was not without contention: subsequent commentary on the field cautioned against treating "media influence" on language as a simple, one-directional force, arguing instead that digital media reshape sociolinguistic practice through complex, mediatised processes rather than through direct causal transmission (Androutsopoulos, 2014). Early studies of the "multilingual internet" showed that speakers of non-Latin-script languages frequently improvised romanised or transliterated forms of writing when native-script input was unavailable or inconvenient, a pattern well attested for Gujarati, Hindi, Punjabi, and other Indic languages (Danet & Herring, 2003; Kirov et al., 2024).

For regional Indian languages generally, and for Gujarati specifically, this digital turn has proceeded unevenly. Regional-language natural language processing (NLP) resources, including annotated corpora, morphological analysers, and machine translation systems, remain comparatively underdeveloped relative to Hindi, Bengali, or Tamil, a disparity attributed in the literature to Gujarati's status as a lower-resource language within India's own multilingual technology landscape (Baxi & Bhatt, 2025; Singh et al., 2021). At the same time, social platforms report substantial and growing consumption of Gujarati-language video, text, and voice content, driven by expanding internet access outside metropolitan centres (Mohan & Punathambekar, 2019). The result is an apparent paradox: rising volumes of everyday Gujarati digital communication coexisting with thin technological and institutional infrastructure to support, standardise, or safeguard that communication. Recent journal evidence on code-mixed Indian-language processing further shows that transliterated, multilingual, and noisy social-media text requires specialised approaches to language identification and correction, supporting the need for Gujarati-specific resources capable of handling authentic digital usage (Jain et al., 2024; Chanda et al., 2025). In parallel, contemporary research on digital inequality indicates that participation should be assessed not only through connectivity but also through the distribution of capabilities and benefits across socio-economic groups, providing a wider analytical context for the uneven digital expansion described in this review (Heeks, 2022; Das & Chatterjee, 2023).

1.3 Scope, Research Gap, and Objectives

This review addresses the following gap: while individual strands of scholarship exist on Indic script technology, code-mixed natural language processing, digital-divide economics, and South Asian diaspora sociolinguistics, no synthesis has drawn these strands together specifically around Gujarati as a case for understanding how a mid-sized, diasporic, historically contact-rich Indian language behaves in digital environments. The review does not attempt an exhaustive census of every technical system or platform relevant to Gujarati; rather, it draws together peer-reviewed evidence across sociolinguistics, human-computer interaction, natural language processing, and digital-inequality research to characterise the emerging practices through which Gujarati speakers write, speak, and are represented online, and the concerns these practices raise for language maintenance, digital equity, and information integrity. The specific objectives are threefold: first, to describe how script, transliteration, and code-mixing shape everyday digital Gujarati; second, to examine the structural conditions, including caste, gender, and regional disparities, that determine who can participate fully

in this digital ecology; and third, to assess the state of language-technology infrastructure and policy response, and to identify priorities for future research.

2. Methods for Literature Selection

2.1 Databases and Indexing Sources

Literature was identified through searches of Web of Science, Scopus, Google Scholar, and PubMed, supplemented by six field-specific indexing sources selected for their relevance to the review's interdisciplinary scope: the ACM Digital Library (for computational linguistics and human-computer interaction work), the Linguistics and Language Behavior Abstracts database, JSTOR, SSRN, the Directory of Open Access Journals, and IEEE Xplore. This combination was chosen to capture sociolinguistic, computational, and social-science literature, since no single database indexes all three fields comprehensively.

2.2 Search Strings and Date Range

Search strings combined the term "Gujarati" with descriptors including "digital communication," "computer-mediated communication," "code-switching," "code-mixing," "script," "transliteration," "natural language processing," "digital divide," "social media," and "diaspora," joined using Boolean operators (for example, "Gujarati AND (digital OR online OR internet)"). Where Gujarati-specific literature proved sparse, searches were broadened to comparable Indic-language and South Asian terms (Hindi, Punjabi, Santali, Marathi, Indian languages) to situate Gujarati within the wider regional pattern, consistent with the review's comparative framing. The search covered literature published between 2003, corresponding to the emergence of dedicated scholarship on multilingual computer-mediated communication, and 8 April 2026. The full strategy is summarised in Table 1.

2.3 Inclusion and Exclusion Criteria

Included sources were peer-reviewed journal articles reporting empirical, theoretical, or technical work bearing directly on Gujarati or closely comparable South Asian languages in digital contexts. Authoritative institutional data on language demographics and digital access were also consulted. Excluded sources were conference proceedings without subsequent journal publication, unrefereed preprints, trade or industry reports lacking peer review, books, and single-language technical manuals with no bearing on sociolinguistic or policy questions. Non-English sources were excluded unless an English-language peer-reviewed translation or summary was available, a restriction that inevitably favours internationally indexed scholarship over Gujarati-medium academic writing.

2.4 Rationale for a Narrative Approach

A narrative rather than systematic review design was adopted because the subject matter spans disparate methodological traditions, including corpus linguistics, ethnographic fieldwork, survey research, and system-building computational papers, for which a single set of systematic-review eligibility criteria would be poorly suited. Narrative synthesis is recognised as appropriate when the evidence base is heterogeneous in method and when the review's purpose is to integrate findings across disciplines rather than to pool comparable effect sizes (Green et al., 2006). This approach permits the identification of cross-cutting themes, such as the recurring tension between digital visibility and infrastructural neglect, that would be difficult to surface through a narrowly bounded systematic protocol.

2.5 Screening, Duplication, and Selection of Influential Studies

Records identified through database searching were screened first by title and abstract, then by full text, against the inclusion criteria above. Duplicate records retrieved across multiple databases were identified by matching digital object identifiers and author-year combinations and were counted once. Because Gujarati-specific empirical work is limited in volume, particular attention was given to identifying studies that were frequently cited within their subfields, that introduced widely adopted datasets or frameworks, or that represented the only available peer-reviewed treatment of a given subtopic; such studies were retained even where sample sizes or

scope were modest, on the grounds that their absence would leave the review unable to address that subtopic at all. Studies addressing India more broadly, without disaggregating by language, were included only where their findings were plausibly generalisable to Gujarati-speaking populations or where they supplied essential structural context, such as national digital-divide patterns.

Table 1. Summary of Search Strategy

Element	Detail
Core databases	Web of Science, Scopus, Google Scholar, PubMed
Field-specific databases	ACM Digital Library, Linguistics and Language Behavior Abstracts, JSTOR, SSRN, DOAJ, IEEE Xplore
Core search terms	Gujarati; digital communication; code-switching; code-mixing; script; transliteration; natural language processing; digital divide; social media; diaspora
Comparator terms	Hindi, Punjabi, Santali, Marathi, Indian languages (used where Gujarati-specific evidence was sparse)
Date range	2003 to 8 April 2026
Review type	Narrative, with structured search reporting
Source types included	Peer-reviewed journal articles; authoritative institutional data
Source types excluded	Conference-only proceedings, preprints, books, trade reports, non-peer-reviewed grey literature

3. Sociolinguistic Profile and Digital Ecology of Gujarati

3.1 A contact Language Moving Online

Gujarati's long history of contact with English, Persian, Arabic, and Portuguese, accumulated through trade, colonial administration, and migration, produced a speaker population accustomed to code alternation well before the advent of digital media (Mesthrie & Chavda, 2020). This historical readiness for multilingual repertoire-switching helps explain why Gujarati speakers, when confronted with English-dominated software interfaces, keyboards, and platform defaults, tended to adapt through improvisation rather than resistance, blending scripts and languages within single messages rather than insisting on monolingual Gujarati-only communication. Broader sociolinguistic theorising on computer-mediated communication supports this reading: language choice online is shaped by audience design, platform affordances, and available input technologies as much as by speaker preference alone (Androutsopoulos, 2015; Herring, 2007).

3.2 Diaspora as a Distinct Digital Public

Because a substantial proportion of Gujarati speakers live outside India, the language's online life is partly organised around diaspora-oriented digital spaces. Comparative work on diaspora websites and forums in Germany found that such spaces develop distinctive patterns of language choice, in which edited or institutional content favours the host-country language while community interaction spaces sustain heritage-language use, code-switching, and identity display through screen names and signatures (Androutsopoulos, 2006b). Gujarati diaspora communities appear to follow a broadly similar pattern, sustaining heritage-language media, community forums, and social media groups that function simultaneously as spaces of linguistic maintenance and as sites where hybrid, code-mixed forms of Gujarati proliferate. This dual function complicates any simple narrative of digital media as either purely protective or purely erosive of heritage languages; instead, diaspora digital practice appears to do both at once, keeping Gujarati visible while accelerating its blending with English.

3.3 Genre Variation within Gujarati Digital Writing

Work on weblog genre and gender in English-language computer-mediated communication demonstrated that stylistic features associated with particular genres, such as the diary-style personal blog versus the filter-style link-sharing blog, cut across author gender in ways that complicate straightforward gender-based explanations of online language variation (Herring & Paolillo, 2006). Although this specific finding was established for English-language blogging, the underlying methodological insight, that digital genre exerts an independent influence on linguistic style, is directly relevant to Gujarati: the register of a Gujarati WhatsApp message, a Gujarati YouTube comment, and a Gujarati news-portal article differ systematically, and any account of

"Gujarati online" that treats these genres as equivalent risks overgeneralisation. A faceted classification approach to computer-mediated discourse, which attends to synchronicity, audience size, anonymity, and message persistence, offers a more precise vocabulary for describing this internal variation than a single undifferentiated category of "digital Gujarati" (Herring, 2007). Reflecting on the same body of sociolinguistic CMC scholarship, an editorial synthesis of this line of work observed that computer-mediated discourse analysis had, by the mid-2000s, matured from a largely descriptive enterprise into one capable of addressing identity, community, and genre with the same analytical seriousness applied to spoken interaction, a maturation that later research on Gujarati and comparable South Asian languages has built upon rather than had to establish afresh (Georgakopoulou, 2006).

4. Script, Transliteration, and Language Technology

4.1 Romanisation as a Transitional and Persistent Practice

For much of the early mobile-phone era, Gujarati speakers, like users of other Indic scripts, faced hardware and software environments poorly equipped to render native script, prompting widespread romanisation, the practice of writing Gujarati words using Latin letters according to informal, non-standardised transliteration conventions (Danet & Herring, 2003). Even as Unicode support for Gujarati script has become near-universal on modern smartphones, romanised Gujarati has persisted rather than disappeared, a pattern consistent with findings from comparable South Asian contexts. Large-scale computational work on romanised South Asian languages, which uses Gujarati script transliteration as one of its illustrative cases, shows that informal romanisation continues to generate substantial volumes of text requiring transliteration back into native script for downstream processing, and that this transliteration is far more tractable at the level of isolated named entities than at the level of full, contextually ambiguous sentences (Kirov et al., 2024). This persistence suggests that romanisation is not merely a stopgap technology but has become an entrenched register choice, associated with informality, speed, and cross-linguistic audience accessibility, that coexists with native-script writing rather than being superseded by it.

4.2 Script, Identity, and the Politics of Writing

Script choice in South Asia carries identity weight beyond simple technical convenience. Comparative work on script politics across the region shows that decisions about which script to use for a given language are frequently entangled with the construction and contestation of ethnic, religious, and national identities, and that these entanglements shape both formal orthographic policy and everyday informal writing practice (Brandt & Sohoni, 2018). The politics-of-authority dimension of this entanglement has been documented in detail for Santali print media, where control over graphic practice and script standards became bound up with community authority and legitimacy rather than remaining a purely technical matter (Choksi, 2017). A closely related but distinct case study of the same language shows how a romanised "trans-script," blending features of multiple writing systems, gained prominence specifically in digital communication as a way of mediating between an official but low-literacy indigenous script and more widely known alternatives, becoming itself a marker of modern digital belonging rather than simply a technical workaround (Choksi, 2020). Gujarati's own script politics are less fraught than those of contested minority scripts such as Ol-Chiki, since Gujarati script enjoys secure official status, but the underlying dynamic, whereby romanisation functions as an index of youth, informality, and transnational orientation, still applies. Community-level graphic practice, including the informal standardisation of specific romanised spellings within particular online communities, further demonstrates that even absent state intervention, digital writing communities develop their own quasi-normative conventions through repeated use (Choksi, 2018).

4.3 Language technology infrastructure for Gujarati

Compared with Hindi, Bengali, Tamil, and Telugu, Gujarati remains a comparatively low-resource language within Indian computational linguistics, a status attributable to the smaller scale of annotated training data and the correspondingly limited academic and industrial investment in Gujarati-specific tools (Baxi & Bhatt, 2025). Recent work has begun to address this gap directly: a bidirectional long short-term memory-based morphological analyser for Gujarati, trained on a purpose-built dataset of 16,234 annotated words, demonstrated that neural sequence models can achieve meaningful accuracy in predicting root forms and grammatical features for Gujarati without relying on hand-crafted suffix rules, a departure from earlier paradigm-based and statistical

approaches (Baxi & Bhatt, 2025). This development matters beyond narrow computational interest, because morphological analysis underlies many downstream applications, including spell-checking, predictive text input, machine translation, and information retrieval, all of which shape the everyday experience of typing and reading Gujarati on digital devices.

At the level of machine translation more broadly, comprehensive reviews of Indian-language systems confirm that while substantial government-sponsored and academic effort has gone into building translation resources for major scheduled languages, resource scarcity, particularly the shortage of large, high-quality parallel corpora, continues to constrain translation quality for many Indian languages, Gujarati among them (Singh et al., 2021). Code-mixed text compounds this difficulty: because a large share of authentic Gujarati social-media writing blends Gujarati, Hindi, and English within single utterances, systems trained on clean monolingual data frequently perform poorly on the messy, code-mixed input that constitutes much of the language's actual digital footprint (Nguyen et al., 2021; Rudra et al., 2019). Work on automatic language identification for Hindi-English code-switched social media text, while not Gujarati-specific, illustrates both the feasibility and the residual error rates involved in disentangling languages at word level within informally romanised, code-mixed streams, a challenge structurally identical to that facing Gujarati-English processing (Nguyen et al., 2021). Table 2 summarises the current state of these resource categories.

Table 2. Summary of Gujarati and Comparable Indic Language Technology Resources

Resource type	Status for Gujarati	Illustrative development
Script rendering	Mature; Unicode support standard on modern devices	Near-universal native-script display
Romanisation and transliteration	Widely used informally; formal context-aware systems emerging	Sentence-level transliteration models for South Asian languages
Morphological analysis	Emerging; limited annotated data historically	Neural morphological analyser trained on 16,234-word dataset
Machine translation	Under-resourced relative to major scheduled languages	Persistent shortage of large parallel corpora
Code-mixed text processing	Early-stage; performance degrades on informal, mixed input	Cross-language code-switching identification systems (Hindi-English exemplar)

Sources: Baxi and Bhatt (2025); Kirov et al. (2024); Nguyen et al. (2021); Rudra et al. (2019); Singh et al. (2021).

5. Code-Mixing, Code-Switching, and Digital Vernacular Practices

5.1 The Normalisation of Mixed-code Writing

Across Indian social media, the blending of a regional language with English and, frequently, Hindi has become the default rather than the exception in informal digital writing. Detailed corpus analysis of English-Hindi code-switching on Twitter identified systematic pragmatic functions served by switching, including narrative-evaluative shifts, emphasis, and clarification, showing that such switching is not haphazard but performs identifiable communicative work (Rudra et al., 2019). Although this evidence base concerns Hindi-English switching specifically, the underlying architecture, a matrix language supplying grammatical structure while embedded elements are drawn from a prestige or technical-register language, is directly applicable to Gujarati-English and Gujarati-Hindi-English trilingual mixing, which is now pervasive in Gujarati WhatsApp messages, Instagram captions, and comment threads.

5.2 Networked Multilingualism and Audience Design

The concept of networked multilingualism, developed to describe how multilingual social-media users select among their full linguistic repertoires in response to perceived and imagined audiences, provides a useful frame for understanding why Gujarati speakers vary their code-mixing intensity across platforms and interlocutors (Androutsopoulos, 2015). A message directed at an extended, semi-public Facebook audience is likely to draw more heavily on widely intelligible English or Hindi elements than a message sent within a tightly bounded family WhatsApp group, where dense Gujarati, including regionally or generationally specific vocabulary, can

be deployed with confidence that it will be understood. This audience-sensitive variation means that Gujarati digital writing cannot be reduced to a single "typical" register; rather, it comprises a continuum from near-monolingual native-script Gujarati in intimate contexts to heavily anglicised, frequently romanised Gujarati in broadcast-style public posts.

5.3 Translanguaging and the Blurring of Named-language Boundaries

Recent applied-linguistics scholarship on translanguaging in digital pedagogical and social contexts argues that treating "Gujarati," "Hindi," and "English" as bounded, mutually exclusive systems increasingly misdescribes how multilingual speakers actually communicate online, since users draw fluidly on their combined linguistic repertoire rather than switching cleanly between discrete codes (Lee & Wei, 2025). This perspective is instructive for Gujarati, where digital messages routinely combine Gujarati script or romanisation, English technical and youth-culture vocabulary, and, particularly in urban and diaspora contexts, Hindi film and media references, in ways that resist tidy segmentation into separate "languages." From a language-maintenance standpoint, this fluid practice raises a genuine question, addressed further in Section 9, about whether such extensive mixing constitutes a healthy adaptive strategy or a gradual erosion of monolingual Gujarati competence among younger, digitally native speakers.

5.4 Linguistic Purism as a Counter-current

Not all Gujarati digital communities embrace mixing uncritically. Comparative sociolinguistic research on minority-language purism elsewhere in Asia demonstrates that where a language's vitality is perceived to be under threat, purist metadiscourse, valorising "unmixed" or "authentic" language use, can itself become a resource for political and cultural resistance, actively promoted and policed within digital community spaces (Baioud & Khuanuud, 2022). While Gujarati is not endangered in the way that some minority languages are, pockets of similar purist sentiment appear within Gujarati literary, educational, and cultural-preservation circles online, where heavy code-mixing is sometimes criticised as a marker of linguistic carelessness or cultural dilution. This tension between pragmatic, audience-driven mixing and normative calls for linguistic purity represents an ongoing, unresolved dynamic within Gujarati's digital public sphere.

6. Digital Divide, Access, and Inequality

6.1 Structural Inequality in Digital Access

Gujarati's digital expansion has not been evenly distributed across the population that speaks the language. National-level research on India's digital divide identifies overlapping axes of exclusion, including caste, income, education, gender, and region, that jointly determine who owns a device, who has reliable connectivity, and who possesses the skills to use digital tools productively (Tewathia et al., 2020). These structural inequalities do not disappear simply because a language has achieved digital visibility in aggregate; rather, they determine which segments of the Gujarati-speaking population are actually producing and consuming the content that constitutes that visibility.

6.2 Caste and Regional Disparities

Ethnographic and survey-based work on caste and digital technology in urban and peri-urban India shows that historically disadvantaged caste groups continue to experience both first-level exclusion, meaning lower rates of device ownership and internet access, and second-level exclusion, meaning lower digital skills and less confident technology use, even where nominal access exists (Kamath, 2018). Gujarat's own social geography, encompassing prosperous urban centres, smaller towns, and rural and tribal regions with markedly different infrastructural endowments, means that caste-linked digital exclusion documented elsewhere in urban India plausibly recurs in Gujarati-speaking regions, although Gujarati-specific quantitative data on this dimension remain scarce in the peer-reviewed literature, a gap noted further in Section 12.

6.3 Gendered Dimensions of Access

Gender constitutes a second major axis of digital inequality relevant to Gujarati speakers. Research on mobile phone ownership and use among women in India documents persistent gaps in ownership rates and identifies

constraints operating simultaneously at individual, family, and community levels, including limits on confidence and digital literacy, restrictions on permitted use imposed by family members, and broader social norms about appropriate contexts for women's technology use (Sonne Vyas, 2025). Complementary economic analysis of mobile-phone ownership barriers among low-income women identifies specific socio-cultural, economic, and health-related factors that jointly restrict ownership even of inexpensive handsets (Potnis, 2016). Because women's participation shapes not only their own access to Gujarati-language digital content but also intergenerational transmission of literacy and digital habits within families, gendered access gaps carry consequences extending beyond individual users to the broader digital vitality of the language.

6.4 Voice Interfaces as a Partial Equaliser

Voice-based technologies have been proposed as a means of extending digital participation to users constrained by literacy or typing proficiency, since spoken interaction removes the requirement to master script input altogether. Ethnographic research on voice user interfaces among diverse user groups in urban India found that such interfaces are enacted through several distinct "technologies-in-practice," including looking up information, learning, and leisure, but also observed that the content and services available through voice assistants remain disproportionately oriented toward more affluent, English-comfortable users, limiting the equalising potential of voice technology for lower socio-economic groups (Kendall et al., 2020). For Gujarati specifically, this suggests that voice interfaces, while promising in principle for extending access to less literate or less typing-confident speakers, have not yet been deployed in ways that fully close existing digital-access gaps. Table 3 summarises these overlapping dimensions of inequality.

Table 3. Dimensions of Digital Inequality Affecting Gujarati-Speaking Populations

Dimension	Nature of disparity	Illustrative evidence
Caste	Lower device ownership and digital skill among historically disadvantaged groups	Peri-urban caste-based exclusion documented in Indian cities
Gender	Lower mobile ownership and more restricted use among women	Individual, family, and community-level constraints on women's phone use
Region	Uneven infrastructure between urban centres and smaller towns or rural areas	National patterns of socio-economic digital divide
Literacy and typing skill	Barriers to text-based participation for less literate users	Voice interfaces proposed as partial remedy, with mixed equalising effect
Income	Affordability constraints on devices and data	Economic barriers to phone ownership among low-income women

Sources: Kamath (2018); Kendall et al. (2020); Potnis (2016); Sonne Vyas (2025); Tewathia et al. (2020).

7. Social Media, Identity, and Diaspora

7.1 Identity Work through Language Choice

Digital platforms function as spaces where Gujarati speakers actively construct and negotiate identity through language and script choice rather than passively transmitting a fixed linguistic repertoire. Foundational computer-mediated communication scholarship established that language choice in online interaction is bound up with the construction of interpersonal relationships and social identities, a claim subsequently elaborated across numerous language communities (Androutsopoulos, 2006a). For Gujarati speakers, the choice between native script and romanisation, or between Gujarati-dominant and English-dominant phrasing, often signals generational identity, educational background, regional origin, or degree of diaspora acculturation, functioning as a form of stylistic self-presentation analogous to clothing or accent choice in offline interaction.

7.2 Diaspora Media and Community Maintenance

Diaspora-oriented Gujarati media, including community news portals, social media groups, and messaging-app networks, serve dual functions of information provision and community cohesion. Research on diaspora websites more generally shows that such platforms structure language use differently across their edited and interactive sections, with institutionally produced content often favouring the host-country language for reach

while user-generated interactive spaces sustain heritage-language and code-switched communication (Androutsopoulos, 2006b). Gujarati diaspora publications appear to follow this pattern, with formally edited news content frequently appearing in English or in relatively standardised Gujarati, while community forums, comment sections, and messaging groups exhibit the denser code-mixing characteristic of informal peer communication.

7.3 Heritage Language Transmission under Digital Conditions

Digital media occupy an ambiguous position with respect to heritage-language transmission among diaspora youth. On one hand, accessible video, messaging, and social platforms provide expanded opportunities for exposure to spoken and written Gujarati outside formal instructional settings, potentially supporting maintenance among second- and third-generation speakers who might otherwise have minimal exposure. On the other hand, the same platforms expose young diaspora Gujarati speakers overwhelmingly to English-dominant global content, and the heavily code-mixed Gujarati that circulates within diaspora digital spaces may itself represent a reduced or restructured variety relative to homeland Gujarati. Comparative minority-language scholarship suggests that social media can simultaneously provide permissive environments for minority-language use while also subjecting that use to competitive and normalising pressure from dominant languages present on the same platforms (Baioud & Khuanuud, 2022), a tension that plausibly applies to diaspora Gujarati as much as to the minority languages more commonly studied in this literature.

8. Platforms, Misinformation, and Media Consumption

8.1 The rise of Regional-language Platform Content

Major digital platforms have increasingly recognised regional Indian languages, Gujarati included, as central to their growth strategies rather than as peripheral markets. Research on YouTube's localisation strategy in India documented how the platform came to treat linguistic and cultural regionality as a core organising principle for content discovery and monetisation, with regional-language creators and audiences becoming central to the platform's expansion beyond English- and Hindi-dominant metropolitan users (Mohan & Punathambekar, 2019). This platform-level recognition has translated into substantial growth in Gujarati-language video, audio, and short-form content, reflecting genuine demand among Gujarati-speaking audiences for material in their own language rather than in English or Hindi alone.

8.2 Vulnerability to Misinformation in Regional Languages

The same conditions that make regional-language content commercially attractive to platforms plausibly create conditions favourable to the spread of misinformation, since fact-checking infrastructure, content-moderation resources, and media-literacy interventions have historically concentrated on English and, to a lesser extent, Hindi, leaving other regional languages comparatively under-served. It should be noted that this specific asymmetry is an inference drawn from the broader pattern of platform prioritisation documented for regional-language content discovery (Mohan & Punathambekar, 2019), rather than a claim that source directly tested; dedicated peer-reviewed research quantifying Gujarati-language misinformation exposure or moderation gaps was not located in the course of this review and is identified as a priority for future work in Section 10. Encrypted messaging applications add a further, largely structural complication: because private group communication is not visible to external platforms, researchers, or automated moderation systems in the way public posts are, it is reasonable to infer, though not directly evidenced in the literature reviewed here, that much everyday Gujarati text exchanged in such groups falls outside the reach of existing verification mechanisms.

8.3 Voice and Multimodal Content as Emerging Vectors

As internet access extends to populations with lower formal literacy, voice notes, video, and image-based content have become increasingly important channels for both legitimate Gujarati-language communication and the circulation of unverified claims. Research on voice interface use in urban India underscores that spoken and multimodal digital interaction is becoming a substantial and distinct mode of engagement, separate from and not fully captured by text-based content-moderation systems designed with written language in mind (Kendall et al., 2020). For Gujarati, whose script rendering and transliteration challenges have historically encouraged voice

and image-based communication as alternatives to typing, this shift toward multimodal content plausibly outpaces the development of corresponding verification and moderation tools. Table 4 summarises these platform-level opportunities and concerns.

Table 4. Platform-Related Opportunities and Concerns for Gujarati Digital Content

Domain	Opportunity	Associated concern
Video platforms	Growing recognition of regional-language audiences and creators	Moderation and fact-checking capacity may lag behind content growth
Messaging applications	Dense, intimate, high-trust Gujarati communication	Limited external visibility for verification, by structural design
Voice and multimodal content	Accessible to lower-literacy users	Verification tools remain oriented toward text-based input
Content localisation strategies	Platform investment in regional-language discovery	Commercial discovery logic does not necessarily extend to information-integrity safeguards

Sources: Kendall et al. (2020); Mohan and Punathambekar (2019).

9. Language Technology Policy, Education, and Institutional Response

9.1 National Policy and Mother-tongue Instruction

India's National Education Policy 2020 marked a significant institutional commitment to mother-tongue and regional-language instruction, advocating that the medium of instruction in early schooling should, where feasible, be the child's home or regional language rather than English (Singh, 2025). Analysis of the policy's implementation trajectory finds that while the policy articulates strong support for multilingual education, translation of this commitment into practice has been uneven, constrained by teacher shortages, limited high-quality instructional materials in regional languages, and persistent social preference for English-medium schooling among aspirational families (Singh, 2025). For Gujarati, this policy environment creates institutional pressure toward greater investment in Gujarati-medium digital educational content, but the gap between policy ambition and classroom reality suggests that digital Gujarati literacy support cannot rely on policy commitment alone.

9.2 Institutional Investment in Language Technology

Government-sponsored efforts to build language technology capacity for Indian languages, including speech recognition, machine translation, and text processing tools, have historically prioritised the most widely spoken scheduled languages, with Gujarati receiving comparatively modest dedicated investment relative to its speaker population (Singh et al., 2021). Academic contributions, including the morphological analyser and associated annotated datasets discussed in Section 4, represent an important but still nascent complement to state-level infrastructure investment, and the field remains reliant on a small number of active research groups rather than a broad, well-resourced ecosystem.

9.3 The Role of Educational Technology in Digital Literacy

Beyond formal medium-of-instruction policy, digital literacy training more broadly shapes whether Gujarati speakers can navigate online environments safely and productively, including recognising misinformation, using privacy settings, and accessing government digital services. Research on multilingual education policy implementation emphasises that effective mother-tongue-based instruction requires coordinated investment across curriculum, teacher training, and digital infrastructure rather than policy statements alone, drawing on comparative international case studies to illustrate the scale of coordination required (Singh, 2025). This suggests that closing Gujarati's digital-literacy gap will depend as much on sustained institutional coordination as on any single technological intervention.

10. Future Directions

Several priorities emerge from this review for future research and practice. First, the development of larger, higher-quality annotated Gujarati corpora, particularly corpora capturing authentic code-mixed and romanised social-media text rather than clean literary or journalistic Gujarati, would substantially improve the performance of downstream natural language processing applications, including predictive text, machine translation, and automated content moderation (Baxi & Bhatt, 2025; Singh et al., 2021). Second, empirical research specifically disaggregating India's national digital-divide patterns by language community would allow researchers to determine whether Gujarati speakers experience distinctive access and skill gaps relative to other linguistic populations, addressing a notable evidentiary gap identified in Section 6 (Tewathia et al., 2020).

Third, context-aware transliteration research of the kind developed for romanised South Asian languages more broadly should be extended and refined specifically for Gujarati, given the persistence of romanised writing even amid widespread native-script support (Kirov et al., 2024). Fourth, platform-level investment in Gujarati-language fact-checking and content-moderation capacity, informed by research on the growth of regional-language platform strategies, represents an urgent applied priority rather than a purely academic one, particularly given that the moderation-asymmetry concern raised in Section 8 currently rests on inference rather than dedicated evidence (Mohan & Punathambekar, 2019). Fifth, longitudinal ethnographic research on diaspora Gujarati communities would help clarify whether current patterns of heavy code-mixing represent a stable, functional bilingual register or a transitional stage toward more substantial language shift among younger diaspora generations, building on frameworks developed for other minority and heritage languages online (Androutsopoulos, 2006b; Baioud & Khuanuud, 2022).

Finally, research attentive to gender should examine whether closing the mobile-ownership and digital-skill gap among Gujarati-speaking women changes the character of Gujarati digital content itself, given that women's participation shapes both content production and intergenerational transmission of digital literacy within families (Potnis, 2016; Sonne Vyas, 2025). Voice-interface research specifically calibrated to lower-literacy Gujarati-speaking users, rather than assuming the affluent, English-comfortable user base documented in existing urban Indian studies, would similarly help determine whether voice technology can fulfil its promised equalising function for this language community (Kendall et al., 2020).

11. Conclusions

Gujarati's movement into digital communication has been rapid, extensive, and technologically improvised rather than centrally planned. Speakers adapted to script-rendering constraints through romanisation long before Unicode support matured, and that romanised register has persisted as a stylistic choice rather than disappearing once native-script typing became convenient. Code-mixing with English and Hindi, itself continuous with Gujarati's long history of linguistic contact, has become the default register of informal digital writing, serving identifiable pragmatic and identity-signalling functions rather than representing careless or degraded language use. At the same time, this expanding digital presence rests on markedly uneven foundations. Language-technology infrastructure for Gujarati remains comparatively underdeveloped relative to major scheduled Indian languages, and access to the devices, connectivity, and skills required to participate fully in Gujarati's digital public sphere is unevenly distributed across caste, gender, income, and regional lines. Diaspora communities sustain the language online while simultaneously accelerating its hybridisation with English, producing a genuinely ambiguous outcome for long-term heritage-language maintenance. Platform and policy responses, from YouTube's regional-content strategies to India's National Education Policy 2020, have begun to recognise Gujarati and comparable regional languages as central rather than peripheral, but implementation continues to lag behind stated commitment. Taken together, the evidence indicates that Gujarati's digital age is characterised less by simple decline or simple flourishing than by an uneven, still-unfolding negotiation between visibility and infrastructure, between adaptive multilingual creativity and the risk of thinning monolingual competence, and between expanding access and persistent exclusion.

12. Limitations

This review is subject to several limitations. The evidentiary base specific to Gujarati remains thin relative to major world languages and even relative to some other Indian languages, requiring the review to draw substantially on closely comparable South Asian and broader Indian-language scholarship rather than Gujarati-

specific studies alone; conclusions extended by analogy from Hindi, Punjabi, or Santali research should be treated as plausible rather than firmly established for Gujarati. The exclusion of non-English-language sources, including Gujarati-medium academic scholarship, means that relevant research published in Gujarati or other Indian languages may have been missed, introducing a language-of-publication bias common to internationally indexed literature reviews. As a narrative rather than systematic review, source selection, while conducted transparently, involved a degree of interpretive judgement in weighing the relevance and quality of heterogeneous studies, and formal risk-of-bias assessment tools appropriate to systematic reviews were not applied. Several important subtopics, including caste-disaggregated digital access data specific to Gujarati-speaking regions and large-scale quantitative studies of Gujarati-language misinformation, could not be substantiated with peer-reviewed evidence and are identified instead as priorities for future research; where the text draws inferences beyond what individual cited studies directly established, this has been flagged explicitly rather than presented as settled fact. Finally, given the rapid pace of change in platform policy, language-technology capability, and government digital-education initiatives, some findings summarised here may already be superseded by developments occurring after the literature search concluded.

Declaration of AI Use

This manuscript was prepared through the combined contributions of all author(s), including contributions to the study design, data, content development, results, interpretation, and related scholarly work. The author(s) acknowledge the use of Grammarly and ChatGPT to assist with grammar checking, language refinement, reference formatting. These AI-assisted tools were not used as authors and did not replace the intellectual contributions or scholarly judgment of the author(s). All AI-assisted outputs, including content, references, and interpretations, were carefully reviewed, revised, verified, and approved by the author(s). The author(s) accept full responsibility for the accuracy, integrity, and final content of the manuscript.

References

Androutsopoulos, J. (2006a). Introduction: Sociolinguistics and computer-mediated communication. *Journal of Sociolinguistics, 10*(4), 419–438. https://doi.org/10.1111/j.1467-9841.2006.00286.x

Androutsopoulos, J. (2006b). Multilingualism, diaspora, and the Internet: Codes and identities on German-based diaspora websites. *Journal of Sociolinguistics, 10*(4), 520–547. https://doi.org/10.1111/j.1467-9841.2006.00291.x

Androutsopoulos, J. (2014). Beyond 'media influence'. *Journal of Sociolinguistics, 18*(2), 242–249. https://doi.org/10.1111/josl.12072

Androutsopoulos, J. (2015). Networked multilingualism: Some language practices on Facebook and their implications. *International Journal of Bilingualism, 19*(2), 185–205. https://doi.org/10.1177/1367006913489198

Baioud, G., & Khuanuud, C. (2022). Linguistic purism as resistance to colonization. *Journal of Sociolinguistics, 26*(3), 315–334. https://doi.org/10.1111/josl.12548

Baxi, J., & Bhatt, B. (2025). A bidirectional LSTM-based morphological analyzer for Gujarati. *Natural Language Processing, 31*(2), 198–214. https://doi.org/10.1017/nlp.2024.14

Brandt, C., & Sohoni, P. (2018). Script and identity—the politics of writing in South Asia: An introduction. *South Asian History and Culture, 9*(1), 1–15. https://doi.org/10.1080/19472498.2017.1411048

Chanda, S., Mishra, A., & Pal, S. (2025). Sentiment analysis of code-mixed Dravidian languages leveraging pretrained model and word-level language tag. *Natural Language Processing, 31*(2), 477–499. https://doi.org/10.1017/nlp.2024.30

Choksi, N. (2017). From language to script: Graphic practice and the politics of authority in Santali-language print media, eastern India. *Modern Asian Studies, 51*(5), 1519–1560. https://doi.org/10.1017/S0026749X16000470

Choksi, N. (2018). Script as constellation among Munda speakers: The case of Santali. *South Asian History and Culture, 9*(1), 92–115. https://doi.org/10.1080/19472498.2017.1411064

Choksi, N. (2020). From transcript to "trans-script": Romanized Santali across semiotic media. *Signs and Society, 8*(1), 62–92. https://doi.org/10.1086/706549

Danet, B., & Herring, S. C. (2003). Introduction: The multilingual Internet. *Journal of Computer-Mediated Communication, 9*(1), Article JCMC9110. https://doi.org/10.1111/j.1083-6101.2003.tb00354.x

Das, S., & Chatterjee, A. (2023). Impacts of ICT and digital finance on poverty and income inequality: A sub-national study from India. *Information Technology for Development, 29*(2–3), 378–405. https://doi.org/10.1080/02681102.2022.2151556

Georgakopoulou, A. (2006). Postscript: Computer-mediated communication in sociolinguistics. *Journal of Sociolinguistics, 10*(4), 548–557. https://doi.org/10.1111/j.1467-9841.2006.00292.x

Green, B. N., Johnson, C. D., & Adams, A. (2006). Writing narrative literature reviews for peer-reviewed journals: Secrets of the trade. *Journal of Chiropractic Medicine, 5*(3), 101–117. https://doi.org/10.1016/S0899-3467(07)60142-6

Heeks, R. (2022). Digital inequality beyond the digital divide: Conceptualizing adverse digital incorporation in the global South. *Information Technology for Development, 28*(4), 688–704. https://doi.org/10.1080/02681102.2022.2068492

Herring, S. C. (2007). A faceted classification scheme for computer-mediated discourse. *Language@Internet, 4*, Article 1. https://scholarworks.iu.edu/journals/index.php/li/article/view/37562

Herring, S. C., & Paolillo, J. C. (2006). Gender and genre variation in weblogs. *Journal of Sociolinguistics, 10*(4), 439–459. https://doi.org/10.1111/j.1467-9841.2006.00287.x

Jain, M., Jindal, R., & Jain, A. (2024). Code-mixed Hindi-English text correction using fuzzy graph and word embedding. *Expert Systems, 41*(7), e13328. https://doi.org/10.1111/exsy.13328

Kamath, A. (2018). "Untouchable" cellphones? Old caste exclusions and new digital divides in peri-urban Bangalore. *Critical Asian Studies, 50*(3), 375–394. https://doi.org/10.1080/14672715.2018.1479192

Kendall, L., Chaudhuri, B., & Bhalla, A. (2020). Understanding technology as situated practice: Everyday use of voice user interfaces among diverse groups of urban users in India. *Information Systems Frontiers, 22*(3), 585–605. https://doi.org/10.1007/s10796-020-10015-6

Kirov, C., Johny, C., Katanova, A., Gutkin, A., & Roark, B. (2024). Context-aware transliteration of romanized South Asian languages. *Computational Linguistics, 50*(2), 475–534. https://doi.org/10.1162/coli_a_00510

Lee, T. K., & Wei, L. (2025). Refashioning linguistic expertise: Translanguaging TESOL in social media. *Applied Linguistics*, Article amaf007. https://doi.org/10.1093/applin/amaf007

Mesthrie, R., & Chavda, V. (2020). Cape Town Gujarati and its relation to Gujarati dialectology: A study of retroflex boosting. *Journal of South Asian Languages and Linguistics, 7*(1), 43–56. https://doi.org/10.1515/jsall-2020-2022

Mohan, S., & Punathambekar, A. (2019). Localizing YouTube: Language, cultural regions, and digital platforms. *International Journal of Cultural Studies, 22*(3), 317–333. https://doi.org/10.1177/1367877918794681

Nguyen, L., Bryant, C., Kidwai, S., & Biberauer, T. (2021). Automatic language identification in code-switched Hindi-English social media text. *Journal of Open Humanities Data, 7*, Article 7. https://doi.org/10.5334/johd.44

Potnis, D. (2016). Inequalities creating economic barriers to owning mobile phones in India: Factors responsible for the gender digital divide. *Information Development, 32*(5), 1332–1342. https://doi.org/10.1177/0266666915605163

Rudra, K., Sharma, A., Bali, K., Choudhury, M., & Ganguly, N. (2019). Identifying and analyzing different aspects of English-Hindi code-switching in Twitter. *ACM Transactions on Asian and Low-Resource Language Information Processing, 18*(3), Article 29. https://doi.org/10.1145/3314935

Singh, A. K. (2025). From policy to practice: Challenges and pathway for advancing multilingual education through India's national education policy 2020. *Policy Futures in Education, 23*(4), 746–765. https://doi.org/10.1177/14782103251320532

Singh, M., Kumar, R., & Chana, I. (2021). Machine translation systems for Indian languages: Review of modelling techniques, challenges, open issues and future research directions. *Archives of Computational Methods in Engineering, 28*, 2165–2193. https://doi.org/10.1007/s11831-020-09449-7

Sonne Vyas, L. (2025). Digital revolution and the gender divide: Factors affecting mobile phone use in India. *Information, Communication & Society, 28*(3), 471–491. https://doi.org/10.1080/1369118X.2024.2420034

Tewathia, N., Kamath, A., & Ilavarasan, P. V. (2020). Social inequalities, fundamental inequities, and recurring of the digital divide: Insights from India. *Technology in Society, 61*, 101251. https://doi.org/10.1016/j.techsoc.2020.101251