

Digital Content Patterns in Pune's Real Estate Market: Information Disclosure and Marketing Content Convergence

Dr. Swapnil Sitaram Phadtare

Post-Doctoral Fellow, SR University Warangal, Telangana, IND

Dr Geetha Manoharan

Associate Professor, School of Business, SR University, Warangal, TG, IND

Abstract

Digital property marketing research requires a distinction between project information, shared website text and buyer responses. This study examines information disclosure, enquiry architecture and lexical similarity in a dated corpus of 49 project webpages from two developer websites serving Pune. Public HTML was collected on 5 October 2026, with URL manifests, response hashes and deterministic dictionary coding. Project main text was compared with whole-body text, and 1,176 page pairs were assessed using unique five-word-shingle Jaccard similarity. Main-text BHK, area, RERA-format, numerical-price and possession markers occurred on 40, 18, 8, 27 and 30 pages, respectively. Mean disclosure breadth was 2.51 of five markers. Median within-developer lexical overlap was 0.0393, compared with 0.0000 between developers. Whole-body extraction increased sustainability-marker prevalence from 5 to 39 pages and investment-marker prevalence from 4 to 49 pages. The median body-to-main word ratio was 16.83. These results show that communication descriptions are sensitive to extraction scope and that enquiry patterns coincide with website templates. The contribution is a traceable audit linking content observations to their extraction definitions. Results apply to the selected source frame and do not establish consumer overload, trust, memory, conversion or sales effects.

Keywords: digital real estate marketing; information disclosure; lexical similarity; website content analysis; extraction sensitivity; Pune

1 Introduction

Property websites organize information that buyers may use before seeking further details. Configuration, area, pricing, registration identifiers and possession language provide different kinds of information, while brochure and visit prompts define routes for continuing an enquiry. Describing these elements is a necessary preliminary step before attributing changes in trust or sales to digital marketing. A count of content features can be reproduced from webpages, whereas a buyer response requires separate observation.

This article integrates disclosure, similarity and extraction analyses within one empirical study. It asks how decision-information markers and enquiry prompts are distributed, how closely project texts overlap, and how the extraction boundary changes recorded prevalence. The analytical unit is a dated webpage, not a buyer or a transaction. The common sampling frame permits direct comparison of substantive content measures and extraction sensitivity.

2 Conceptual foundation and research questions

Signaling theory distinguishes communicated signals from receiver interpretation (Connelly et al., 2011). A numerical price or registration-format string can be recorded as a potential information signal without demonstrating credibility or accuracy. Information-overload research similarly distinguishes the quantity of available information from the capacity and task of its recipient (Eppler and Mengis, 2004). Information structure may affect decision processes (Lurie, 2004), but HTML word counts do not directly measure processing difficulty. Cognitive-load and mediated-message theories motivate future consumer studies rather than supplying behavioural outcomes for this audit (Sweller, 1988; Lang, 2000).

Advertising repetition studies examine repeated encounters by consumers, including effects that depend on familiarity and frequency (Campbell and Keller, 2003; Schmidt and Eisend, 2015). The present similarity measure instead concerns overlap between documents. A high pairwise overlap may reflect related development phases or common wording. It supplies a possible stimulus-selection variable, but does not reveal how often an individual encountered an advertisement.

RQ1 concerns the prevalence and breadth of five decision-information markers in project main text. RQ2 concerns enquiry prompts and literal HTML media elements. RQ3 concerns lexical and amenity-category overlap within and between the two developer websites. RQ4 concerns changes in measured prevalence when extraction expands from main text to the whole page body. These descriptive questions are answered by the results below; no retrospective behavioural hypotheses are assigned support labels.

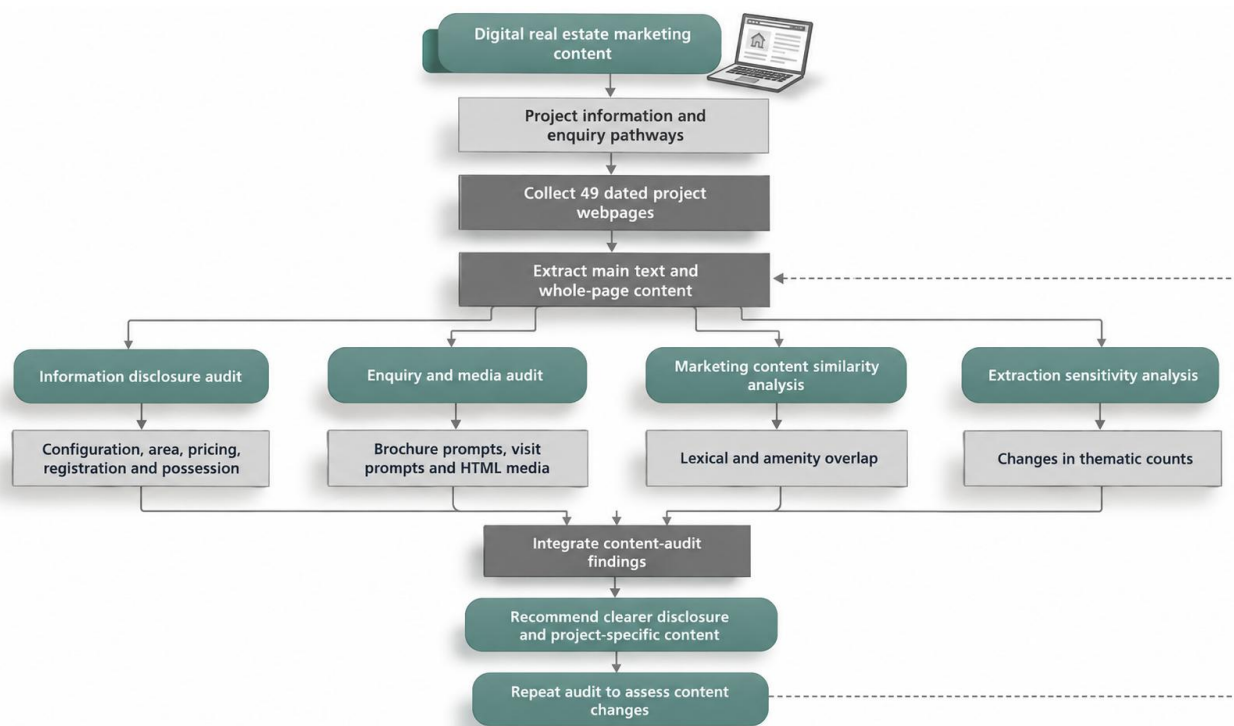


Figure . Digital marketing content-audit workflow for Pune's real estate market.

3 Materials and methods

3.1 Sampling frame

The source frame consists of the recorded project links from <https://www.koltepatil.com/pune/residential-properties/ongoing> and <https://www.godrejproperties.com/pune>. Every eligible distinct URL in these two frozen inventories was selected. Kolte-Patil configuration-category pages labelled 1BHK, 2BHK and 3BHK and the generic Life Republic parent were excluded. Project-specific Life Republic child pages were retained. Godrej phase-specific URLs were retained as separate page units. The resulting frame contains 15 Kolte-Patil and 34 Godrej URLs.

This is a census of a selected two-website link frame, not a census of Pune projects, developers or homebuyers. Different phases and related projects may share a development and template. A page appearing under an ongoing-project path is not independently verified as currently available. Residential and plotted project language can occur within the frame; commercial and agricultural property markets are not represented sufficiently to support conclusions about them. A third candidate developer inventory failed during the earlier discovery and is outside the frozen frame.

All 49 selected project requests returned HTTP 200 and contained a main element with an h1 heading. URLs, redirects, UTC request timestamps, response size and SHA256 hashes are preserved. The collection uses public HTML responses and does not log into websites, submit enquiry forms, contact developers or collect private account data. Acquisition elapsed time is retained as a collection diagnostic and cannot be reported as browser page-load speed.

Table 1 Corpus composition and acquisition outcomes

Measure	Recorded result
Kolte Patil project URLs	15
Godrej project URLs	34
Total selected and retrieved pages	49
HTTP 200 responses	49
Prior pilot URL overlap	16
Additional project URLs	33
Page pairs	1176
First match evidence records	1011

3 2 Extraction and coding

The primary extraction selects the innermost main element containing h1. Nested script, style, navigation, header, footer, noscript and SVG elements are removed. Nonblank text nodes are stripped, whitespace-normalized and joined with newlines. Unicode word-token counts include numbers. No JavaScript execution, image OCR, CSS visibility filtering or browser viewport measurement is performed. Consequently, the extracted text is HTML-available text rather than a verified record of everything seen by a visitor.

The sensitivity extraction uses the entire body and removes script, style, noscript and SVG but deliberately retains navigation, headers and footers. This scope can capture corporate material and repeated menus. A main-versus-body comparison measures sensitivity to the extraction boundary; it does not establish that one representation is a complete gold standard. An additional exact text-node deduplication analysis identifies repeated strings inside the main extraction, not repeated advertisements experienced by a buyer.

Deterministic case-insensitive dictionaries detect five decision-information markers, seven additional marketing-language categories, ten amenity categories and four enquiry-prompt categories in each scope. A first-match evidence row records the short matched token and its zero-based character offset. The offset is a location in a normalized string, not on-screen position or attention. Full dictionaries and field definitions are provided in the workbook and archive. Expanded dictionary version 2 places word boundaries around Rs and INR to prevent incidental letter fragments from being treated as price markers. Direct comparison with the earlier pilot price counts requires recoding under this corrected definition.

Disclosure breadth equals the sum of five binary information flags. Exact-node repetition equals one minus the deduplicated main word count divided by the original main word count. Text Jaccard equals the number of shared unique five-word shingles divided by the union count. Amenity Jaccard uses category sets in the same way; an empty union is undefined and remains missing rather than zero. Body-to-main ratio is computed separately for each page before taking its median. These measures are descriptive indices, not reflective psychological scales.

3 3 Analysis and quality control

The analysis links every record to a requested URL, final URL and response hash. The corpus has no failed selected-page retrievals, although the source frame itself remains narrow. Dictionary zero means that a pattern was not detected in the relevant extraction. It does not mean a fact is absent from the rendered site. Short HTML extracts, synonyms, text embedded in images and dynamic sections can all produce false negatives. A

RERA-format match identifies a string pattern only; registrations, prices, amenities and possession dates have not been validated externally.

No independent human coding panel was involved, so Cohen kappa, Krippendorff alpha and inter-rater agreement are not available. Reproducible automated coding does not establish content validity. Human review should check false positives and negatives against rendered pages and brochures, with disagreements resolved through a documented rubric. Consumer-scale reliability and convergent validity statistics require participant responses. Cronbach alpha, composite reliability, AVE and SEM paths must not be assigned to this website corpus as though it were a buyer survey.

Descriptive statistics characterize the frozen frame. No population confidence intervals, hypothesis-support labels, developer performance rankings or causal regression estimates are reported. In the similarity dataset, 1,176 rows arise from 49 pages; a page participates in many pairs, making the pair observations dependent. Statistical precision must be based on page or development sampling rather than treating all pairs as independently sampled people.

All reported results use expanded dictionary version 2 and the same saved responses. The corrected price dictionary prevents incidental terminal letters in words such as developers from matching Rs followed by a number. Earlier pilot price counts require recoding before comparison. Descriptive counts and rounded statistics were checked against the record-level JSON and workbook summaries. No independent human validation or rendered-browser accuracy estimate is implied by these computational checks.

4 Results

4 1 Text quantity and observed breadth

Table 2 Page level quantities across 49 pages

Measure	Mean	Median	Minimum	Maximum
Main words	324.16	99.00	59	1635
Deduplicated main words	311.41	94.00	54	1606
Whole body words	1609.37	1652.00	945	2280
Amenity categories	2.51	1.00	0	9
Enquiry prompt categories	1.31	1.00	1	2

Main extracts had a median of 99 words and a mean of 324.16, indicating a skewed distribution. Whole-body text averaged 1,609.37 words. The mean of page-level exact-node repetition fractions was 5.39%; this is distinct from the percentage reduction obtained by dividing pooled word totals. Website structure contributes to these observations, so longer extraction should not be interpreted as better marketing.

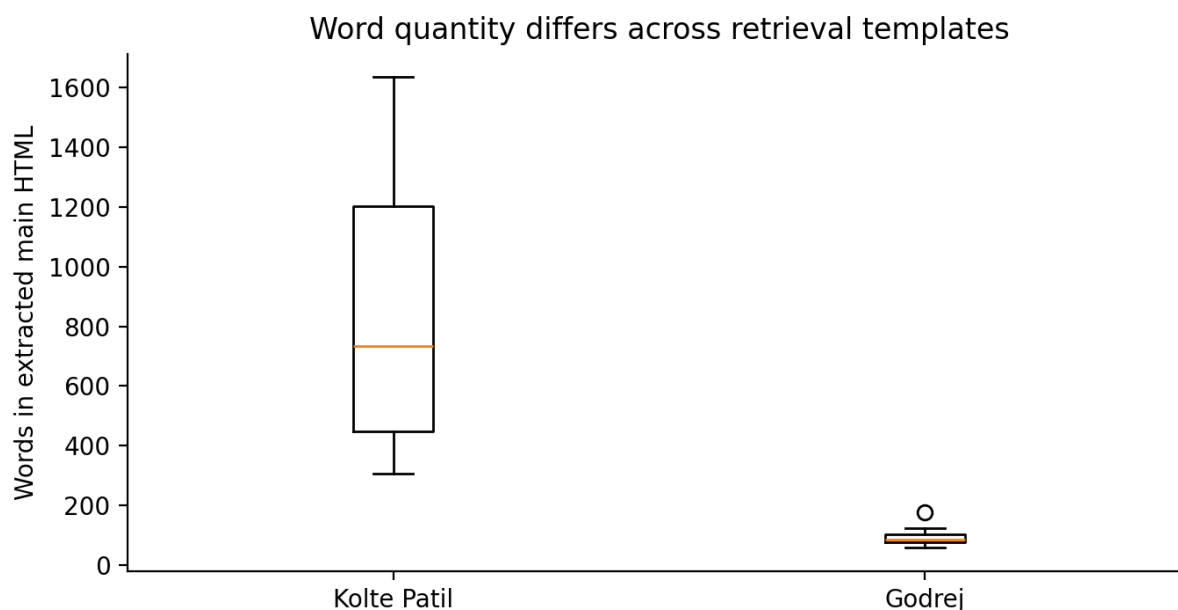


Figure 1 Distribution of main text word quantity in the selected corpus

4 2 Information disclosure

Table 3 Decision information markers by extraction scope

Marker	Main n	Main %	Body n	Body %
BHK configuration	40	81.6%	49	100.0%
Area unit	18	36.7%	44	89.8%
RERA format	8	16.3%	8	16.3%
Numerical price	27	55.1%	27	55.1%
Possession	30	61.2%	30	61.2%

Table 4 Disclosure breadth distribution

Markers present	Pages	Share
0	6	12.2%
1	3	6.1%
2	8	16.3%
3	24	49.0%
4	8	16.3%
5	0	0.0%

Mean disclosure breadth was 2.51 markers. Twenty-four pages contained three markers, while six contained none and no page contained all five. The index counts selected detected categories and is not a verified completeness rating. Area-unit matches may describe site acreage rather than apartment carpet area. Possession wording may not specify a usable date, and RERA-format strings were not checked against a registry.

4 3 Enquiry architecture and media

Table 5 Pages containing enquiry or HTML features

Feature	Pages	Share
brochure	15	30.6%
enquiry	15	30.6%
callback	0	0.0%
visit	34	69.4%
pdf links	2	4.1%
iframe elements	12	24.5%
video elements	0	0.0%
form elements	15	30.6%

Brochure and enquiry markers occurred on the 15 Kolte-Patil extracts, while visit markers occurred on the 34 Godrej extracts. This alignment with developer templates prevents separating a prompt effect from the website in this frame. PDF, iframe and form counts describe literal main-element markup. They do not verify accessibility, functionality, views, downloads or submissions.

Table 6 Image alt attributes in main HTML

Attribute observation	Image elements	Share of 770
All images	770	100.00%
Nonblank alt	720	93.51%
Blank alt	50	6.49%
Absent alt	0	0.00%

All 770 image elements had an alt attribute; 720 were nonblank and 50 blank. These element-level counts are not page-level accessibility scores. Empty alt may be appropriate for decorative images, and the appropriateness of nonblank alternatives requires image-context review (World Wide Web Consortium Web Accessibility Initiative, 2026).

4 4 Content repetition and lexical overlap

Table 7 Completed content repetition and similarity results

Measure	Calculated value	Unit
Mean exact node repetition	5.39%	49 pages
Median within developer text Jaccard	0.0393	666 dependent pairs
Median between developer text Jaccard	0.0000	510 dependent pairs
Maximum text Jaccard	0.4324	64 shared / 148 union
Identical five word shingle sets	0	1,176 pairs

Table 7 summarizes recorded textual repetition and overlap. They answer a content question without assigning memory benefits, fatigue trajectories or spacing effects to unobserved consumers.

Table 8 Similarity by pair group

Pair group	Pairs	Text median	Text maximum	Amenity median	Defined amenity pairs
------------	-------	-------------	--------------	----------------	-----------------------

Pair group	Pairs	Text median	Text maximum	Amenity median	Defined amenity pairs
All	1176	0.0191	0.4324	0.0000	1005
Kolte Patil within	105	0.0290	0.1924	0.6667	105
Godrej within	561	0.0437	0.4324	0.0000	390

The highest text overlap occurred between Godrej Meadows and Godrej Meadows 2. The highest Kolte-Patil pair shared 142 shingles out of a union of 738, yielding 0.1924. The Kolte-Patil median amenity overlap of 0.6667 exceeds its lexical median of 0.0290, showing that category overlap and repeated phrasing describe different properties of the content. There were 171 undefined amenity comparisons because both pages had empty category sets. No pair was counted as identical at the unique five-word-shingle level.

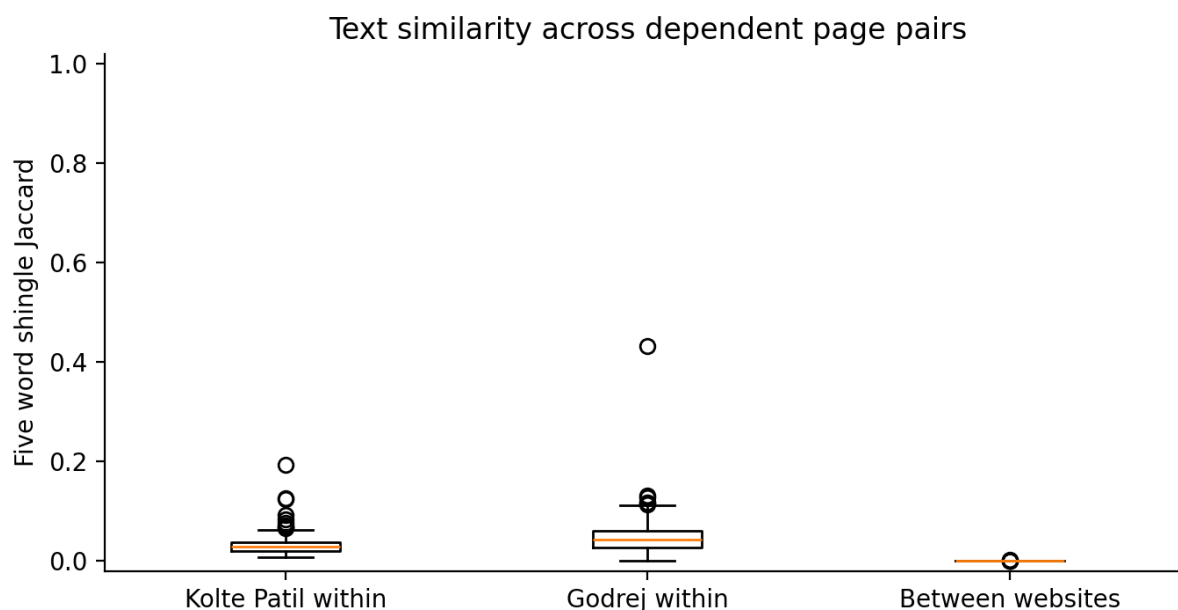


Figure 2 Distribution of pairwise lexical similarity with dependent page pairs

4.5 Sensitivity to extraction scope

Table 9 Marketing language by extraction scope

Category	Main n	Body n	Body only n	Increase pp
Qualifier	1	35	34	69.4
Scarcity	0	0	0	0.0
Sustainability	5	39	34	69.4
Luxury	19	49	30	61.2
Connectivity	16	49	33	67.3
Investment	4	49	45	91.8
Wellbeing	13	13	0	0.0

Table 10 Extraction quantities by website

Group	Pages	Mean main words	Mean body words	Median body main ratio
All	49	324.16	1609.37	16.83

Group	Pages	Mean main words	Mean body words	Median body main ratio
Kolte Patil	15	850.93	1497.40	1.87
Godrej	34	91.76	1658.76	19.12

Whole-body extraction increased sustainability prevalence by 69.4 percentage points and investment prevalence by 91.8 percentage points. The median body-to-main ratio was 16.83 overall, 1.87 for Kolte-Patil and 19.12 for Godrej. In the inspected Godrej example E016, Sustainability appeared in navigation outside the selected main text. This demonstrates how a literal whole-page match can have a different context from a project description; it does not demonstrate misleading sustainability claims. The observed website differences combine template, content and extraction effects.

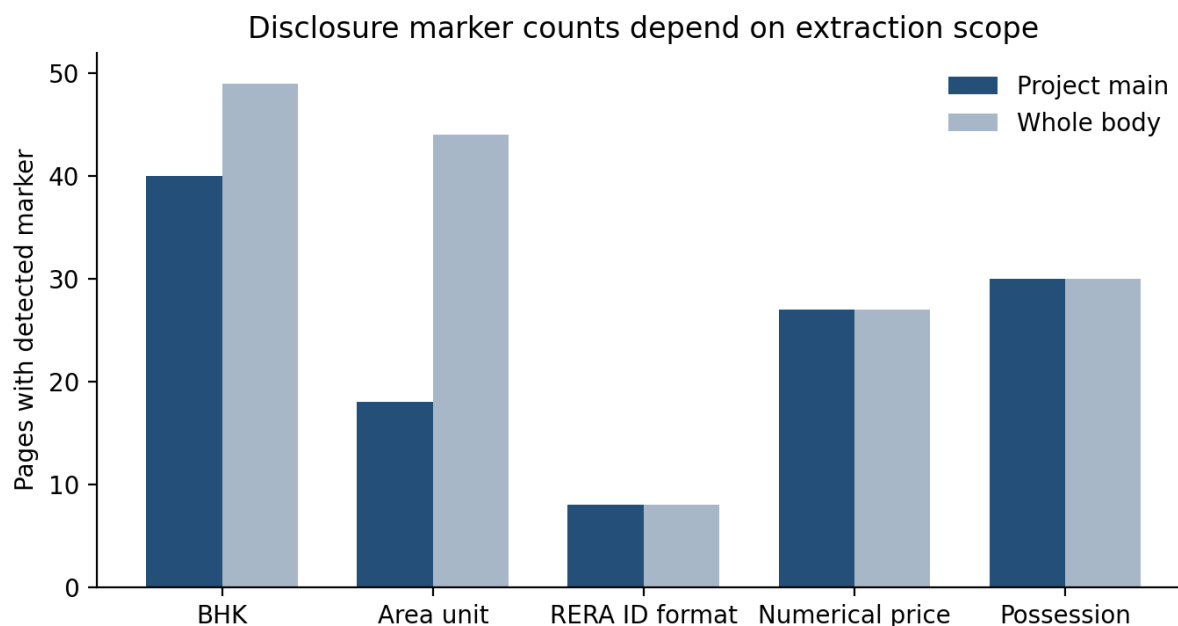


Figure 3 Decision information marker presence under paired extraction scopes

5 Discussion

The findings connect a substantive marketing description to a measurement choice. Decision-information markers are unevenly distributed, but the frequency of detected language depends on whether shared navigation and corporate material are included. Treating body-level language as project-specific information can therefore alter the characterization of a developer portfolio without a corresponding change in project copy. The paired design makes this consequence observable within the same pages.

The disclosure results identify categories that can be audited before evaluating buyer responses. BHK language was common, while RERA-format matches were less frequent in the main extracts. These observations should guide a contextual review of communication, rather than a claim that the detected markers constitute transparency or legal compliance. Signaling theory is useful precisely because the presence of a potential signal and its interpretation remain different empirical questions.

Similarity also requires a clear unit of interpretation. Low median shingle overlap does not imply unique positioning, because texts may share meaning while using different words. Conversely, overlap between phase pages can be appropriate when factual project information is shared. The amenity and lexical measures reveal complementary aspects of communication. Neither supports an effective-frequency threshold or a conclusion about advertising wear-out.

5.1 Managerial implications

Developers can use the audit protocol to check whether decision information is recoverable from project sections and whether enquiry prompts are placed alongside it. A useful next review distinguishes carpet area from total site area, verifies identifiers through authoritative routes, and checks whether possession and price statements are specific and current. These recommendations follow from measurement limitations, not demonstrated conversion improvements.

Marketing analysts should define extraction scope before comparing projects, maintain separate corporate and project fields, and retain short match evidence for review. Similarity analysis can identify candidate project pairs for editorial inspection or future stimulus selection. HTML media counts should be supplemented with rendered-page checks before evaluating virtual experiences or accessibility. Website-group summaries must not be used as causal developer rankings.

5.2 Limitations and future research

The frame covers two developer websites on one collection date and cannot represent all Pune property markets. Related phases and shared templates constrain independence. No JavaScript rendering, OCR or viewport measurement was used, and dictionary omissions or ambiguous matches may affect counts. Independent adjudicated labels are needed to estimate coding precision and recall. Source hashes preserve acquisition identity but do not certify developer claims. The literature foundation is selective rather than systematic.

A subsequent consumer experiment could vary content redundancy, exposure count and spacing while measuring delayed brand-linked recall and ad-specific fatigue. A separate disclosure experiment could measure comprehension, credibility and voluntary information seeking. These are future designs, not missing numerical cells in the completed website results. Campaign performance analysis additionally requires authorized impressions, clicks, spend and lead outcomes. The present corpus supplies neither participant scales nor a basis for Cronbach alpha, AVE, composite reliability or buyer SEM paths.

6 Conclusion

This study provides evidence of uneven information disclosure and selective marketing content convergence within a corpus of 49 real estate project pages relevant to Pune's market. Within-developer lexical similarity was generally low (median = 0.0393), although individual comparisons exhibited greater overlap (maximum = 0.4324). Shared enquiry structures therefore coexist with differentiated wording, suggesting that structural consistency and textual similarity represent distinct dimensions of digital property communication.

The higher frequency of sustainability and investment matches under whole-body coding demonstrates that extraction scope materially influences thematic measurement. Reliable interpretation requires explicit page boundaries, documented source provenance and contextual validation of keyword matches. The 1,176 dependent pairwise comparisons should consequently be understood as relationships within the selected corpus rather than independent observations.

The study contributes a reproducible basis for evaluating disclosure coverage and content similarity, supporting consistent disclosure criteria, project-specific information and periodic content review. Its conclusions remain bounded by the selected corpus and content-analysis design. Further research should examine broader samples and integrate consumer and commercial evidence to establish whether observed content characteristics influence buyer understanding, trust and sales outcomes.

References

- Campbell, M. C., & Keller, K. L. (2003). Brand familiarity and advertising repetition effects. *Journal of Consumer Research*, 30(2), 292–304. <https://doi.org/10.1086/376800>
- Connelly, B. L., Certo, S. T., Ireland, R. D., & Reutzel, C. R. (2011). Signaling theory: A review and assessment. *Journal of Management*, 37(1), 39–67. <https://doi.org/10.1177/0149206310388419>

- Eppler, M. J., & Mengis, J. (2004). The concept of information overload: A review of literature from organization science, accounting, marketing, MIS, and related disciplines. *The Information Society*, 20(5), 325–344. <https://doi.org/10.1080/01972240490507974>
- Lang, A. (2000). The limited capacity model of mediated message processing. *Journal of Communication*, 50(1), 46–70. <https://doi.org/10.1111/j.1460-2466.2000.tb02833.x>
- Lurie, N. H. (2004). Decision making in information-rich environments: The role of information structure. *Journal of Consumer Research*, 30(4), 473–486. <https://doi.org/10.1086/380283>
- Schmidt, S., & Eisend, M. (2015). Advertising repetition: A meta-analysis on effective frequency in advertising. *Journal of Advertising*, 44(4), 415–428. <https://doi.org/10.1080/00913367.2015.1018460>
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4
- World Wide Web Consortium Web Accessibility Initiative. (2026). Images tutorial. Updated 8 April 2026. <https://www.w3.org/WAI/tutorials/images/> (accessed 5 October 2026).
- Kolte-Patil Developers. (2026). Pune ongoing residential project inventory and linked project pages. <https://www.koltepatil.com/pune/residential-properties/ongoing> (retrieved 5 October 2026; sampled URLs in accompanying manifest).
- Godrej Properties. (2026). Pune project inventory and linked residential project pages. <https://www.godrejproperties.com/pune> (retrieved 5 October 2026; sampled URLs in accompanying manifest).
- Phadtare, S. S., & Manoharan, G. (2026). Digital marketing strategies for real estate sales in Pune: The role of interactive content, virtual experiences and buyer trust. *Ochrona Przed Korozja*, 8(9), 239–249. <https://doi.org/10.10124/OPKJ/2608922>