International Journal of Latest Research in Humanities and Social Science (IJLRHSS)
Volume 09 - Issue 08, 2026
www.ijlrhss.com || PP. 54-60

Measuring Firm-Level Digital Transformation: A Methodological Review of Annual-Report Text Indicators, Construct Validity, and Reporting Practices

Sun Jiangbaihe[1*], Badrul Hisham Bin Kamaruddin[2]
[1]*City University Malaysia, City Graduate School, Menara City U, No. 8, Jalan 51A/223, 46100 Petaling Jaya, Selangor Darul Ehsan, Malaysia*
[2]*City University Malaysia, City Graduate School, Menara City U, No. 8, Jalan 51A/223, 46100 Petaling Jaya, Selangor Darul Ehsan, Malaysia*
**Corresponding Author: Sun Jiangbaihe*

Abstract: Firm-level digital transformation is widely studied but cannot be observed through a single, universally accepted measure. Researchers therefore rely on annual report language, surveys, digital investment, patents, recruitment records, and composite maturity indices. These indicators are useful for different purposes, yet they capture different parts of the transformation process. This focused methodological review examines firm-level measurement with particular attention to annual report text. The review shows that text-based indicators remain attractive because they permit large samples, longitudinal analysis, and reproducible construction. However, the interpretation of these indicators should remain more limited than the broader organisational construct they are intended to represent. A keyword score may reflect implemented projects, managerial attention, strategic communication, policy response, or a mixture of these elements. Its validity also depends on dictionary design, report length, duplicate text, document sections, industry vocabulary, and changes in terminology over time. Recent studies improve measurement by restricting the corpus, using theoretically informed dictionaries, applying word-embedding models, and separating capability dimensions. Based on this evidence, the paper proposes a seven-part reporting framework covering construct definition, data source, dictionary development, text cleaning, indicator construction, validation, and limitations. The proposed framework is especially relevant to archival studies in which digital-transformation measures are linked to firm performance, innovation investment, or capital allocation efficiency. From the limited body of research reviewed, no indicator is consistently superior. The appropriate measure depends on the research question, and the conclusions should not extend beyond what the chosen data can reasonably show.

Keywords: Digital Transformation, Annual Report Text Analysis, Construct Validity, Reporting Framework

1. Introduction

Digital transformation has become a common subject in management, accounting, information systems, operations, and innovation research. The term usually refers to more than the adoption of individual technologies. It describes coordinated changes in strategy, organisational design, business processes, decision routines, and value creation [1]–[4]. This broad meaning is useful, but it creates a practical problem. Revenue, employment, and investment can be read directly from financial records. Digital transformation must usually be inferred from incomplete traces left by the firm.

Recent firm-level studies use several kinds of evidence. Some count digital terms in annual reports. Others use surveys, software and information technology expenditure, digital patents, recruitment records, or multidimensional maturity scales. These measures are not interchangeable. A survey can capture leadership support and employee skills, while a patent records an invention. Expenditure shows resource commitment, but not whether systems were integrated or used well. Annual report language records formal disclosure. It does not provide a direct audit of implementation.

The distinction matters because measurement choices can change empirical results. A maturity scale may include data governance, process integration, and organisational capability. A keyword count usually records the frequency or breadth of disclosed digital language. Both indicators may rise in the same firm, but they need not do so. Differences between studies can therefore reflect different stages of the underlying process rather than contradictory economic effects [5]–[8].

Text-based measurement has become common in research on Chinese listed firms [9], [10], [13] and has also been applied to United States 10-K filings [11], [12]. The method is practical, transparent, and suitable for long panels. It is also sensitive to decisions that are easy to overlook, including the document section, dictionary scope, repeated language, report length, and changing terminology. More sophisticated algorithms do not remove these choices. They move the choices into training data, model design, and validation.

This paper addresses three questions. First, how is firm digital transformation measured in recent firm-level research? Second, what validity and comparability problems arise when annual report text is used? Third, what information should be reported so that readers can judge and reproduce the indicator? The paper does not estimate the effect of digital transformation on performance. Instead, it provides a methodological foundation for firm-level archival studies that examine performance and mediating mechanisms, including innovation investment and capital allocation efficiency. Performance studies are discussed only when they illustrate how a measurement decision affects interpretation.

A methodological distinction is therefore necessary between a measure that is easy to reproduce and a measure that covers the full organisational meaning of transformation [1], [3], [8]. A stable keyword count may be coded consistently across years, yet it can still represent disclosure rather than operational integration [8]–[11]. This difference is small in wording but substantial when results are interpreted for managers or policymakers [5], [8].

2. Review Approach

This study adopts a focused narrative methodological review rather than a systematic literature review. Studies were selected purposively when they provided a clear definition of a firm-level digital construct and sufficient information about its measurement approach. Recent digital-transformation studies formed the main evidence base, while earlier foundational textual-analysis studies were included when they provided directly relevant methodological guidance.

Conceptual reviews, empirical applications, survey scales, machine-learning measures, and statistical measurement discussions were compared on five points: construct definition, data source, unit of analysis, construction procedure, and stated limitations.

The focused design has a clear boundary. It identifies recurring measurement choices and reporting problems, but it does not claim to cover every scale, dictionary, or industry index. The value of the review lies in comparing how different measures work and what they permit researchers to conclude.

3. Construct Boundaries

Measurement should begin with the construct. Digitisation is the conversion of analogue information into digital form. Digitalisation refers to the use of digital technologies to improve existing activities. Digital transformation is broader. It involves coordinated organisational change that may alter structures, capabilities, processes, products, services, and value propositions [1]–[5]. A firm can digitalise one production line without transforming the organisation as a whole.

Technology adoption represents the most narrowly defined and directly observable aspect of digital transformation. A firm may introduce cloud software, an industrial internet platform, or an artificial intelligence application. The project may remain within one department or end after a pilot. In that setting, technology adoption is an accurate label. A claim of enterprise-wide transformation would require evidence of wider integration, changed routines, or altered decision rights.

Digital orientation, digital intensity, and digital maturity describe different conditions. Orientation concerns strategic emphasis. Intensity concerns the breadth or depth of digital activity. Maturity concerns the extent to which technology, data, skills, governance, and processes have been integrated over time. Nasiri, Saunila, and Ukko [7]treat these as related but distinct constructs. Their findings also show why a single label should not be used for all three.

Digitalisation capability is different again. It concerns what the firm can do with digital resources. Recent reviews identify integration, platform, and innovation capabilities, while machine-learning work separates sensing, seizing, and reconfiguring dimensions [6], [13]. These measures require a theory that links words or survey items to a defined capability. A list of popular technologies is not, by itself, a capability model.

Text-based studies should therefore use restrained wording. For longitudinal firm-year analysis, disclosed digital-transformation intensity is a more defensible label than digital maturity when the empirical proxy is annual-report keyword frequency. The theoretical construct remains broader than the observable text-based indicator. A survey that covers leadership, data use, employee competence, and process integration may justify a maturity label. The name of the variable should follow the evidence rather than the preferred theory.

4. Firm-Level Measurement Approaches

4.1 Annual Report Text Indicators

Annual report text analysis is one of the most widely used firm-level methods. A typical procedure builds a dictionary containing terms such as artificial intelligence, big data, cloud computing, blockchain, digital platform, industrial internet, and smart manufacturing. The terms are counted in each report. Researchers may use the raw total, the number of represented categories, a count per thousand words, or ln(1 + frequency). For firm-year research, an illustrative specification is ln(1 + annual-report digital-keyword frequency), which reduces skewness and retains observations with zero mentions. However, the resulting variable should still be interpreted as disclosed digital-transformation intensity rather than complete digital maturity. When this measure is used to explain firm performance or mediation pathways, researchers should report the dictionary, corpus, text-cleaning rules, timing, and robustness alternatives.

Chinese studies have used annual reports to construct longitudinal indicators and relate them to stock liquidity, core-business performance, and other firm outcomes [9], [10]. In the United States, Chen and Srinivasan [11] restrict the corpus to the business-description section of 10-K reports. Zareie [12] and colleagues use a broader 10-K score and examine how organisational capital changes its association with corporate value. These applications confirm that the method can be adapted across disclosure systems. They do not establish that all corpora measure the same phenomenon.

The main strengths are coverage, repeated observation, and reproducibility. Annual reports can be linked to accounting, market, ownership, innovation, and governance data. A dictionary and code file can also be archived. The central weakness is indirectness. The score depends on what management discloses, where it discloses it, and how the text is cleaned. These decisions belong in the main method section, not only in an appendix.

As a practical reporting aid, researchers can show the original sentence, the matched term, the category assigned to that term, and its resulting contribution to the firm-year score. Such a worked example does not replace code disclosure, but it can help readers detect ordinary problems such as partial matches, duplicated passages, and words used in a non-digital sense.

4.2 Survey and Maturity Scales

Survey instruments capture internal conditions that are rarely visible in public records. Items may cover digital strategy, leadership support, data integration, employee skills, process redesign, customer interfaces, and organisational agility. When the research question concerns maturity or capability, a carefully validated multi-item scale can offer broader content coverage than a simple word count.

Survey evidence is also perceptual. Respondents can overstate progress, interpret the same project differently, or answer in a way that reflects current performance. Common-method bias becomes more serious when both digital transformation and performance are collected from the same person at the same time. Samples are usually smaller and cross-sectional. Their advantage is depth; their limitation is historical and population coverage.

4.3 Digital Investment and Asset Measures

Accounting and operational measures include software expenditure, information technology spending, digital equipment, digital-related intangible assets, and the share of digital projects in total investment. These variables show whether management has committed resources. They are less dependent on narrative language, but they remain input measures. High spending can coexist with duplicated systems, poor data quality, or failed implementation.

Comparability is difficult because firms classify software, licences, platforms, embedded systems, and consulting costs differently. Some items are expensed, while others are capitalised. A digital component may also be included in a larger production project. For this reason, asset-based variables should normally be described as digital investment or digital capital rather than complete firm transformation.

4.4 Patents, Recruitment, and Composite Indicators

Patents provide evidence of digital inventions. Recruitment records show demand for digital skills. Procurement, website activity, platform use, and technology partnerships offer additional behavioural traces. These sources reduce dependence on management narrative, but each has a narrow scope. Patents favour firms that invent rather than adopt. Recruitment data miss transformation achieved through consultants, acquisitions, or retraining.

Composite indices combine several inputs, processes, and outcomes. Their coverage can be wider, but weighting decisions become more important and data availability may vary by industry. Composite and behavioural indicators can be particularly useful as complementary or validation measures, although they may also serve as primary measures when their construction is theoretically justified and transparent.

5. Validity of Text-Based Indicators

5.1 What Text Indicators Measure

A text score is an observable feature of a document. It is not the transformation itself. Depending on the corpus and dictionary, the score may capture implemented projects, managerial attention, planned investment, perceived risk, or communication with investors. The measure is still useful when its meaning is stated precisely. Problems arise when a disclosure indicator is interpreted as a complete maturity assessment.

Recent methods move beyond an undifferentiated count. Yang and colleagues [13]use word embeddings and a theory-based dictionary to estimate sensing, seizing, and reconfiguring capabilities from annual reports. Bellstam, Bhagat, and Cookson [15]show in a different setting that text can identify forms of corporate innovation not captured by patents or R&D. These studies demonstrate the value of richer text methods. They also show that a text measure becomes more credible when its categories are tied to a clear construct.

5.2 Disclosure and Implementation

The largest validity concern is the gap between language and practice. One firm may discuss artificial intelligence because it has integrated the technology into quality inspection. Another may use the same words because competitors use them or policy documents encourage them. The frequency can be similar even though the organisational change is not.

Annual reports are regulated documents, but they are also communication instruments. Digital language may combine actual activity, strategic emphasis, forward-looking plans, and impression management. Studies based on 10-K disclosure find meaningful associations with valuation and performance [11], [12]. Those associations support the economic relevance of the measure. They do not prove that every mention represents a completed project.

A practical response is to code different types of statements separately. Action verbs, completed projects, quantified investment, expected benefits, and future plans need not carry the same weight. A small manual sample can test whether high scores reflect operational evidence or only general discussion. Although modest in scope, such manual validation can provide direct evidence about content validity that additional outcome regressions cannot supply.

Manual checking can be particularly informative near the upper and lower tails of the score distribution. A high-scoring report may reveal a detailed account of completed projects, but it may also contain repeated strategic language across several sections. A low-scoring report may describe one technically important system in concise terms, so frequency alone should not be treated as a direct ranking of implementation quality.

5.3 Dictionary Design and Semantic Ambiguity

The dictionary determines what the algorithm can see. A narrow list misses relevant language. A broad list admits ambiguous words such as platform, cloud, intelligence, and network. The same term can also have different meanings across industries. Dictionary construction should therefore be treated as a substantive research task rather than a mechanical preliminary step.

Common starting points include policy documents and terminology adopted in prior studies [8], [9]. Researchers may then supplement these sources with expert review or corpus-based expansion. Each source has a potential bias. Policy documents may encourage standardised wording. Borrowed dictionaries can reproduce earlier omissions. Purely data-driven expansion can identify linguistic similarity without conceptual relevance.

A defensible procedure begins with theory-based seed terms, expands them through corpus evidence, manually reviews proposed additions, and records exclusions. Word-embedding methods can improve coverage [13], but they still require human decisions about the seed words, training corpus, similarity threshold, and final categories. The full dictionary and its version should be disclosed.

5.4 Document Cleaning, Length, and Section Choice

Long reports naturally contain more terms. Raw frequency therefore reflects both digital content and document length. Researchers can use term density, logarithmic transformation, or a report-length control. These choices are not equivalent. Density measures concentration, while a logarithm only compresses extreme counts. The selected transformation should match the intended interpretation.

Repeated text creates another problem. The same project may appear in the chairperson's statement, management discussion, risk section, and notes. Entire paragraphs may also be copied from the previous year. Counting every occurrence can overstate change, while aggressive deduplication can remove legitimate evidence from different business units. Researchers should state the duplicate-handling rule and examine whether the indicator is sensitive to alternative treatments.

Section choice is equally important. A term in the business description is more likely to describe current operations. The same term in a risk section may refer to uncertainty. Chen and Srinivasan's [11]restricted 10-K corpus offers a clear example of matching the document section to the construct. Full-report analysis remains reasonable, but section-level sensitivity checks can show whether one part of the report drives the score.

Text preparation should report file format, optical character recognition, header and footer removal, treatment of tables, tokenization, abbreviations, duplicate detection, and comparative prior-year text. Financial documents require domain-specific rules because ordinary dictionaries can misread terms that have technical accounting meanings [14].

5.5 Industry and Time Comparability

The informational value of a term varies by industry. Smart manufacturing may signal substantial change in a traditional industrial firm but little in a technology-native company. Cloud computing can be a strategic initiative in one sector and routine infrastructure in another. General dictionaries improve formal comparability, yet they can favour industries whose vocabulary resembles the chosen terms.

Firm size also affects disclosure. Large firms publish longer reports and employ specialist investor-relations teams. Smaller firms may implement important changes but describe them briefly. Reasonable responses include industry-year standardisation, report-length controls, technology-sector exclusions, and industry-specific validation. No single adjustment is correct for every design.

Time creates a separate problem. Digital vocabulary changes quickly. Terms that indicated advanced technology in 2015 may describe ordinary infrastructure ten years later. A fixed dictionary improves consistency but misses emerging technologies. An updated dictionary improves coverage but can weaken year-to-year comparability. A stable core dictionary with rule-based additions offers a practical compromise. Year fixed effects absorb a general rise in digital language, but they do not correct semantic drift.

5.6 Validation and Interpretation

Validation should use more than one test. Expert review and manual coding examine content validity. Correlation with software assets, digital patents, recruitment, technology partnerships, or survey evidence can assess convergence. Discriminant tests should show that the score is not merely report length, general optimism, innovation language, or policy imitation.

A significant coefficient does not validate the construct by itself. A poorly measured indicator can still be statistically related to performance. Validation asks whether the score represents the intended concept, not only whether it predicts an outcome. Recent measurement discussions in China reach the same conclusion: the preferred route depends on the research objective and the available data [8].

Interpretation should remain within the observable range. A high annual-report score can support a conclusion about disclosed digital activity or strategic emphasis. It cannot, without additional evidence, establish enterprise-wide implementation quality. This limited claim is not a weakness. It is a more accurate statement of what the data show.

6. Reporting Framework

A text-based study becomes easier to evaluate when the construction process is visible. This review proposes seven reporting elements: construct definition, data source, dictionary development, text cleaning, indicator construction, validation, and limitations. The sequence is deliberately simple. It can be used in a short empirical paper without turning the method section into a technical manual.

First, define the construct before presenting the words. State whether the variable represents adoption, orientation, disclosed intensity, maturity, or capability. Second, identify the corpus. Report document type, years, language, source, sections, and treatment of revised reports. Third, disclose the full dictionary, categories, seed terms, expansion rules, manual screening, and excluded ambiguous terms.

Fourth, describe text cleaning. At minimum, report format conversion, optical character recognition, headers, footers, tables, tokenization, abbreviations, duplicates, and report-length treatment. Fifth, provide the exact formula. The reader should know whether the study uses raw counts, density, category breadth, a binary indicator, ln(1 + frequency), term frequency-inverse document frequency, or a model-based score.

Sixth, report validation. A useful package combines manual review with at least one external measure where data permit. Seventh, state operational limits. Instead of writing that textual analysis has limitations, explain the likely direction of error. Random noise may weaken estimates. Strategic disclosure may inflate the score when stronger firms communicate more actively.

Version control should accompany these seven elements. Dictionaries, cleaning scripts, and coding rules often change during a project. The submitted paper should identify the version used for the reported results. Where confidentiality permits, the code, dictionary, and a small worked example should be archived.

7. Implications and Future Research

Measurement differences have substantive consequences. A maturity scale, an investment measure, and a keyword score observe different stages of digital change. Their coefficients may differ because attention, spending, integration, capability development, and economic returns do not occur at the same time. Apparent disagreement between studies may therefore be a measurement issue before it is a theoretical contradiction.

Three research directions appear especially useful. The first is multi-source validation. Annual report text can be compared with patents, software assets, recruitment, procurement, or survey evidence. The second is semantic separation. Models can distinguish completed systems from plans, risks, and general discussion. The third is controlled adaptation. Core terms can remain stable while industry-specific and emerging terms are added under published rules.

More complex models should be judged by the information they add. Word embeddings can recover context and build multidimensional measures [13]. Composite indices can cover more of the transformation process. Both approaches also introduce additional assumptions. From the limited sample of studies reviewed, interpretability and reproducibility remain at least as important as algorithmic complexity.

Future comparative work should separate two questions. One asks whether firms differ in digital transformation. The other asks whether researchers measured different aspects of that transformation. Making this distinction explicit would improve cumulative evidence and reduce avoidable disputes over the sign or size of reported effects.

Building on the measurement concerns highlighted in [8] and the disclosure-based approaches in [11]–[13], future studies could examine whether disclosure-based measures behave differently before and after a firm begins to report quantified investment, completed deployment, or organisational redesign. This staged comparison would connect language with observable milestones without assuming that all digital terms represent the same degree of change. It would also allow researchers to test whether an early communication signal later develops into a more durable capability or remains mainly a statement of intent.

Measurement choices may also alter the estimated relationship between digital transformation and firm performance and the apparent strength of mediating pathways. Future studies should examine whether the roles of innovation investment and capital allocation efficiency remain stable across alternative text dictionaries, section-restricted scores, report-length adjustments, and non-text measures of digital transformation.

8. Conclusion

Firm digital transformation cannot be represented fully by one indicator. Surveys, digital assets, patents, recruitment records, composite indices, and annual report text each reveal a different part of the process. Text-based indicators remain useful because they permit broad coverage, long time series, and reproducible construction. Their main limitation is also clear: language can reflect implementation, attention, plans, signalling, and policy response at the same time.

Sound measurement begins by matching the construct to the proxy. It continues with transparent reporting of the data, dictionary, cleaning rules, formula, validation, and limitations. The central conclusion is restrained but important. Better measurement does not begin with a more elaborate algorithm. It begins with a precise statement of what the indicator can and cannot represent. For firm-level performance research, annual-report text provides a defensible proxy for disclosed digital-transformation intensity when its construction and limitations are transparent, but it should not be treated as direct evidence of complete organisational transformation.

9. References

[1] C. Gong and V. Ribiere, “Developing a Unified Definition of Digital Transformation,” Technovation, CII, Article 102217, 2021.

[2] P. C. Verhoef, T. Broekhuizen, Y. Bart, A. Bhattacharya, J. Q. Dong, N. Fabian, and M. Haenlein, “Digital Transformation: A Multidisciplinary Reflection and Research Agenda,” Journal of Business Research, CXXII, pp. 889-901, 2021.

[3] A. Hanelt, R. Bohnsack, D. Marz, and C. Antunes Marante, “A Systematic Review of the Literature on Digital Transformation: Insights and Implications for Strategy and Organizational Change,” Journal of Management Studies, LVIII (5), pp. 1159-1197, 2021.

[4] L. Wessel, A. Baiyere, R. Ologeanu-Taddei, J. Cha, and T. Blegind Jensen, “Unpacking the Difference Between Digital Transformation and IT-Enabled Organizational Transformation,” Journal of the Association for Information Systems, XXII (1), pp. 102-129, 2021.

[5] J. Dąbrowska et al., “Digital Transformation, for Better or Worse: A Critical Multi-Level Research Agenda,” R&D Management, LII (5), pp. 930-954, 2022.

[6] A. Annarelli, C. Battistella, F. Nonino, V. Parida, and E. Pessot, “Literature Review on Digitalization Capabilities: Co-Citation Analysis of Antecedents, Conceptualization and Consequences,” Technological Forecasting and Social Change, CLXVI, Article 120635, 2021.

[7] M. Nasiri, M. Saunila, and J. Ukko, “Digital Orientation, Digital Maturity, and Digital Intensity: Determinants of Financial Success in Digital Transformation Settings,” International Journal of Operations & Production Management, XLII (13), pp. 274-298, 2022.

[8] Y. Yang and M. Gao, “Enterprise Digital Transformation: Concepts, Technical Routes of Statistical Measurement, and Ideas for Improvement,” Statistical Research, XLI (3), pp. 62-73, 2024.

[9] F. Wu, H. Hu, H. Lin, and X. Ren, “Enterprise Digital Transformation and Capital Market Performance: Empirical Evidence from Stock Liquidity,” Journal of Management World, XXXVII (7), pp. 130-144, 2021.

[10] L. Yi, F. Wu, and X. Chang, “Enterprise Digital Transformation Process and Main Business Performance: Empirical Evidence from the Text Recognition of the Annual Reports of Listed Companies in China,” Modern Finance and Economics, XLI (10), pp. 24-38, 2021.

[11] W. Chen and S. Srinivasan, “Going Digital: Implications for Firm Value and Performance,” Review of Accounting Studies, XXIX (2), pp. 1619-1665, 2024.

[12] M. Zareie, N. Attig, S. El Ghoul, and I. Fooladi, “Firm Digital Transformation and Corporate Performance: The Moderating Effect of Organizational Capital,” Finance Research Letters, LXI, Article 105032, 2024.

[13] J. Yang, J. Liu, Z. Yao, and C. Ma, “Measuring Digitalization Capabilities Using Machine Learning,” Research in International Business and Finance, LXX, Article 102380, 2024.

[14] T. Loughran and B. McDonald, “When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks,” The Journal of Finance, LXVI (1), pp. 35-65, 2011.

[15] G. Bellstam, S. Bhagat, and J. A. Cookson, “A Text-Based Analysis of Corporate Innovation,” Management Science, LXVII (7), pp. 4004-4031, 2021.

Author Profile

Sun Jiangbaihe obtained a bachelor's degree from Shanghai Lixin University of Accounting and Finance in China in 2019, and a master's degree from Harbin Institute of Technology in China in 2021. Currently, she is pursuing a doctoral degree in Business Administration at City UniversityMalaysia. Her research interests include digital transformation, firm performance, innovation investment, and capital allocation efficiency.

Badrul Hisham Bin Kamaruddin received his Ph.D. from Universiti Utara Malaysia in 2007. He is currently an Adjunct Professor at the City Graduate School, City University Malaysia. His research interests include corporate finance, Islamic finance, supply chain finance, production efficiency, and Islamic entrepreneurship.