RESEARCH · QUANTITATIVE AND MIXED METHODS

Content Analysis: Quantifying Text and Media

How content analysis turns text, images and media into data that can be counted and compared, covering its history, quantitative and qualitative forms, sampling, coding schemes, intercoder reliability, computer-assisted and automated methods, and validity.

ModQu research article · 2,811 words · 25 references (APA 7th edition)

Every day, organisations and people produce enormous quantities of text and media: news stories, advertisements, social media posts, customer reviews, speeches, reports and open-ended survey answers. **Content analysis** is a family of methods for analysing such material systematically. Krippendorff (2018) defined it as a research technique for making replicable and valid inferences from texts, or other meaningful matter, to the contexts of their use. In its classic form it counts how often particular categories appear, turning communication into numbers that can be compared over time, across sources or between groups. In other forms it interprets meaning more closely. This article explains how content analysis works, how to make it reliable and valid, and how computers have changed it.

A short history

Systematic analysis of texts has a long history, but content analysis became a defined method in the mid-twentieth century. During the Second World War, analysts studied enemy propaganda to infer intentions and morale. Lasswell (1948) summarised the study of communication in a famous question: who says what, in which channel, to whom, with what effect? Content analysis addresses the "says what" part.

Berelson (1952) gave an influential early definition: a research technique for the objective, systematic and quantitative description of the manifest content of communication. The emphasis on **manifest content**, what is visibly present in a text, distinguished the method from looser interpretation. Holsti (1969) broadened the definition to include inferences about the sources and audiences of messages, and to cover any technique for making inferences by objectively and systematically identifying specified characteristics of messages.

A famous early application of statistics to text came from Mosteller and Wallace (1963), who studied the disputed authorship of several of the Federalist Papers. By counting how often Hamilton and Madison used common words such as "upon" and "whilst", they concluded that Madison had very probably written the disputed essays. Gerbner and Gross (1976) used content analysis on a large scale to measure violence on American television year by year, as part of their **cultivation** research into how television shapes viewers' beliefs.

Quantitative and qualitative content analysis

Content analysis comes in two broad forms. **Quantitative content analysis** applies predefined categories to a sample of material, codes each unit, and analyses the resulting counts statistically. Neuendorf (2017) described it as a summarising, quantitative analysis of messages that follows the standards of the scientific method, including attention to objectivity, reliability, validity, generalisability and replicability.

Qualitative content analysis focuses on meaning and context. Mayring (2000) described it as an approach that keeps the systematic, rule-guided character of quantitative content analysis but develops categories through close reading and interpretation. Hsieh and Shannon (2005) distinguished three approaches: **conventional**, in which categories are derived from the data; **directed**, in which an existing theory provides initial categories; and **summative**, which counts keywords and then interprets their use in context. Elo and Kyngäs (2008) described the process of inductive and deductive qualitative content analysis, and Graneheim and Lundman (2004) clarified key concepts such as meaning units, condensation and the distinction between manifest and **latent content**, the underlying meaning that requires interpretation. Schreier (2012) set out practical steps for building a coding frame in qualitative content analysis.

The two forms are not sharply separated. Many studies begin qualitatively, reading material to develop categories, and then apply those categories quantitatively to a larger sample.

What content analysis can infer

Content analysis describes texts, but researchers usually want to learn about something beyond them. Krippendorff (2018) described several kinds of inference. Researchers may infer the **intentions or characteristics of senders**, as wartime analysts did when they used changes in propaganda to anticipate military plans. They may infer **audience attitudes**, for example by treating the content of best-selling books or popular programmes as a reflection of what audiences want. They may infer **social trends**, such as changing attitudes to a group, from shifts in how it is portrayed over decades. Or they may describe **the content itself** as a feature of the environment people live in, as Gerbner and Gross (1976) did with television violence.

Each kind of inference needs justification. A rise in news coverage of crime does not show that crime has risen; it may reflect editorial choices. A rise in negative reviews may reflect a change in who writes reviews rather than a change in the product. Holsti (1969) emphasised that inferences about the causes or effects of communication require evidence beyond the content, such as data on senders, audiences or events. Good content analysis states which inference it is making and what supports it.

The process

Quantitative content analysis follows a series of steps. The diagram below summarises them.

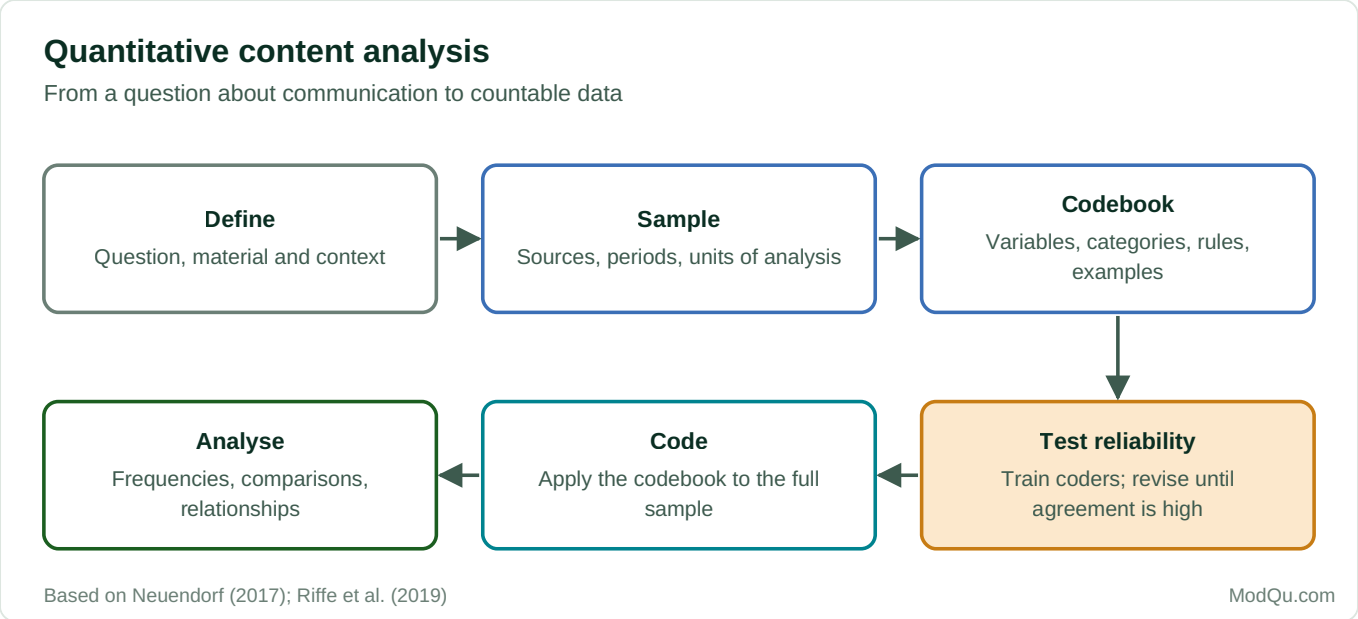


Figure 1. The main steps in quantitative content analysis. Based on Neuendorf (2017) and Riffe et al. (2019); diagram by ModQu.

Defining the question and the content

The researcher begins with a clear research question and identifies the body of material that can answer it: all front-page news stories in three newspapers over five years, all customer reviews of a product, all advertisements in a set of magazines. Krippendorff (2018) emphasised that content analysis always involves inference from texts to some context, so the researcher should be clear about what that context is: the intentions of senders, the attitudes of audiences, or the state of public debate.

Sampling

The full population of relevant material is often too large to code by hand. Riffe et al. (2019) described sampling strategies for media content, including random samples, stratified samples by source or period, and **constructed weeks**, in which days are sampled so that each day of the week is represented, avoiding bias from the weekly cycles of news. The **unit of analysis** must also be defined: the whole article, the paragraph, the sentence, the image or the advertisement.

Choosing units and categories

Decisions about units and categories have large effects on results. Coding at the level of whole articles is quick but loses detail when an article covers several topics; coding at the sentence level is precise but slow. Categories can be **descriptive**, recording features such as topic, source or format, or **evaluative**, recording tone or stance. Evaluative categories are more interesting but harder to code reliably, so they need especially clear definitions and examples. Riffe et al. (2019) advised that researchers decide in advance how to handle units that fit no category or more than one, and record these rules in the codebook so that all coders treat such cases the same way.

Developing a codebook

A **codebook** defines each variable to be coded, its categories and rules for applying them. Good categories are exhaustive, so that every unit can be coded, and mutually exclusive, so that each unit fits only one category for a given variable. Each category needs a clear definition with examples and guidance for borderline cases. Neuendorf (2017) recommended developing the codebook iteratively, testing it on sample material and refining unclear categories.

Training and testing coders

Coders are trained using the codebook, then code a common sample independently. Their agreement is measured to assess **intercoder reliability**. If agreement is too low, the codebook is revised and coders retrained. Only when reliability is acceptable does full coding begin.

Coding and analysis

Coders apply the codebook to the full sample. The resulting data are analysed using descriptive and inferential statistics: frequencies of categories, comparisons between sources or periods, and relationships between variables.

Intercoder reliability

Reliability is central to content analysis, because coding involves human judgement. If coders cannot agree on how to classify material, the results reflect the coders as much as the content. Lombard et al. (2002) reviewed content analyses published in communication journals and found that many did not report reliability at all, and that those that did often used inadequate measures. They recommended reporting a chance-corrected index for each variable.

Percentage agreement, the simplest measure, overstates reliability because some agreement occurs by chance, especially when one category is very common. Cohen (1960) proposed **kappa**, which corrects for chance agreement between two coders. Krippendorff's **alpha** extends the idea to any number of coders, any level of measurement and missing data. Hayes and Krippendorff (2007) argued that alpha should be the standard measure of reliability for coding data. Krippendorff (2018) suggested that values of 0.80 or above indicate acceptable reliability, with values from 0.667 acceptable only for tentative conclusions.

Lacy et al. (2015) reviewed current practices and recommended that researchers report reliability for each variable, describe how the reliability sample was chosen, and test reliability on a sample large enough to give a meaningful estimate.

Validity

Reliability is necessary but not sufficient. A codebook can be applied consistently and still fail to capture what the researcher intends. Krippendorff (2018) discussed several forms of validity in content analysis, including whether categories capture the meanings that matter, whether results correlate with other evidence about the context, and whether inferences from texts to contexts are justified.

A common validity problem is the gap between **manifest** and **latent** content. Counting the word "risk" in company reports is reliable, but the word may be used in very different ways. Coding whether a report expresses concern about risk is closer to the meaning, but requires more judgement and is harder to code reliably. Kolbe and Burnett (1991) reviewed content analyses in consumer research and found that many paid insufficient attention to the objectivity of coding procedures and to reporting the details needed to judge validity.

Computer-assisted and automated methods

Computers have transformed content analysis. Weber (1990) described early computer-assisted methods based on **dictionaries**: lists of words assigned to categories, such as positive and negative emotion words, which software counts automatically. Tausczik and Pennebaker (2010) described Linguistic Inquiry and Word Count, a widely used dictionary program, and reviewed evidence that word counts reveal psychological states such as attention, emotion and social relationships.

Machine learning has added further methods. In **supervised learning**, human coders label a training sample, and an algorithm learns to classify the remaining material. In **unsupervised learning**, algorithms find patterns without predefined categories. **Topic models**, such as latent Dirichlet allocation (Blei et al., 2003), identify groups of words that tend to appear together, which can be interpreted as topics. Roberts et al. (2014) developed structural topic models, which allow topics to vary with characteristics of the documents, and applied them to open-ended survey responses.

The diagram below places these approaches on a spectrum from manual to automated.

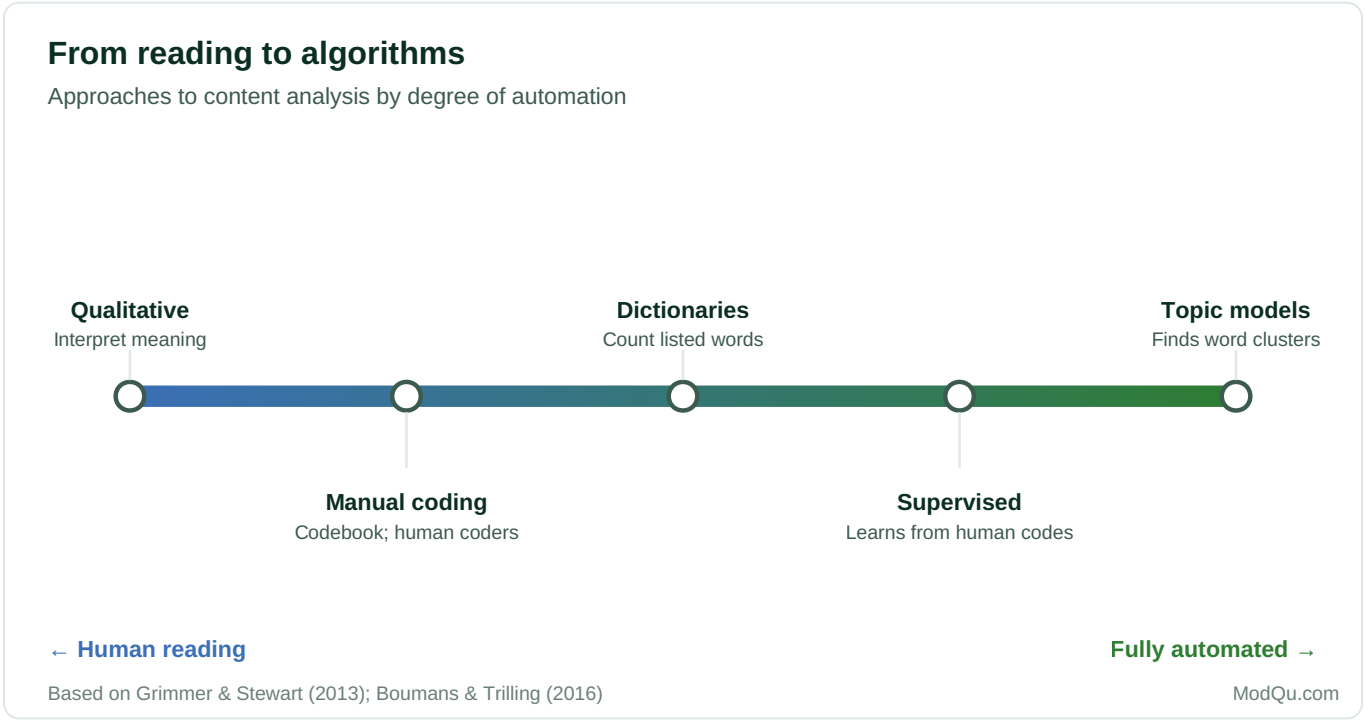


Figure 2. Approaches to content analysis, from manual to automated. Based on Grimmer and Stewart (2013) and Boumans and Trilling (2016); diagram by ModQu.

Grimmer and Stewart (2013) set out four principles for automated text analysis. All quantitative models of language are wrong, but some are useful. Automated methods amplify and augment careful reading, but do not replace it. There is no single best method for every task. And every automated method must be validated against human judgement for the task at hand. Validation usually means comparing the algorithm's output with careful human coding on a sample that the algorithm has not seen, and reporting measures such as accuracy, precision and recall for each category. A classifier that performs well overall can still perform poorly on rare categories, which are often the ones of most interest. Boumans and Trilling (2016) reviewed automated approaches for journalism research and recommended choosing methods according to whether categories are known in advance and how much human coding is available. van Atteveldt and Peng (2018) warned that computational methods bring their own risks, including reliance on data that platforms choose to provide, and errors that are hard to detect at scale.

Combining content analysis with other methods

Content analysis is often strongest when combined with other data. Linking what media say to what audiences believe, measured by surveys, allows researchers to study effects rather than only content; this was the design of the cultivation research by Gerbner and Gross (1976), which compared heavy and light television viewers' beliefs with the content they watched. Linking the content of company communications to financial or operational data can show whether what organisations say matches what they do. In mixed methods studies, qualitative content analysis of a small sample can develop categories that are then applied quantitatively to a large sample, and interviews can explain patterns that the counts reveal.

Ethics of analysing published and online material

Most content analysis uses material that is already public, so it may seem to raise few ethical questions. Online material is more complicated. Social media posts and forum messages are public in a technical sense, but their authors may not expect them to be collected and analysed, and quoting them can make authors identifiable through a simple search. Researchers should consider whether the material was intended for a public audience, whether the topic is sensitive, and whether direct quotations are needed or paraphrase would serve. Platform terms of service may also limit how data can be collected and shared. For material such as customer reviews or complaints held by an organisation, the organisation's privacy commitments to its customers apply.

Strengths and limitations

Content analysis has several strengths:

- it is **unobtrusive**: the material already exists, so analysing it does not affect the people who produced it;
- it can handle **large volumes** of material, especially with computer assistance;
- it allows **comparison** across time, sources and contexts;
- its procedures can be **replicated** by other researchers using the same codebook.

It also has limitations:

- it describes what was communicated, not why, or what effect it had, without other evidence;
- the codebook shapes what can be found, so important features outside its categories may be missed;
- reliable coding of subtle meaning is difficult;
- automated methods can be opaque and may perform poorly on material different from their training data.

Common mistakes

Several mistakes recur in content analysis. The first is **sampling convenience material**: analysing whatever is easiest to obtain, such as the first pages of search results or the posts a platform makes available, without asking how they relate to the wider body of content. The second is **vague categories**, which coders interpret differently, producing low reliability or, worse, high agreement on a misunderstood definition. The third is **reporting percentage agreement only**, which can look high even when coders agree little more than chance would predict (Lombard et al., 2002). The fourth is **over-interpreting counts**: treating the frequency of a word or theme as a measure of its importance, when a single prominent statement may matter more than many passing mentions. The fifth, in automated work, is **trusting an off-the-shelf tool** without checking that it works on the material at hand; a sentiment dictionary built for product reviews may misread irony, specialist vocabulary or another variety of English. Each of these can be avoided by planning the sample, piloting the codebook, using chance-corrected reliability measures, reading examples behind the numbers and validating automated methods against human coding.

Uses in business and applied research

Content analysis has many applied uses. Companies analyse customer reviews and complaints to identify common problems. Marketers analyse competitors' advertising to understand positioning. Communications teams track media coverage of their organisation. Researchers analyse open-ended survey responses to supplement closed questions.

A hypothetical example: a hotel group wants to know what drives negative reviews. It samples 2,000 reviews stratified by hotel and season. A team reads 200 reviews to develop a codebook with categories such as cleanliness, staff, noise, value and booking problems. Two coders apply the codebook to a reliability sample of 150 reviews, revising two categories until Krippendorff's alpha exceeds 0.80 for each. A supervised classifier is then trained on the human-coded reviews and applied to the full sample, with its accuracy checked against a further human-coded set. The analysis shows that noise complaints are concentrated in a few city-centre hotels and rise in summer. The numbers are illustrative, but the process combines the strengths of human judgement and automated scale.

Reporting content analysis

A good report describes the material and how it was sampled, defines the unit of analysis, summarises the codebook and makes the full version available, describes coder training, reports intercoder reliability for each variable using a chance-corrected measure, and explains how automated methods were validated. Neuendorf (2017) emphasised that replicability is a defining feature of content analysis, and that it depends on reporting these details.

Conclusion

Content analysis turns communication into data that can be counted, compared and tested. Its quantitative form depends on careful sampling, clear codebooks and demonstrated reliability; its qualitative form on systematic interpretation of meaning. Computers have made it possible to analyse material at a scale that manual coding could never reach, but they have not removed the need for careful definitions, human judgement and validation. Used well, content analysis provides systematic evidence about what people, organisations and media say, and how that changes over time.

References

Berelson, B. (1952). *Content analysis in communication research*. Free Press.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Boumans, J. W., & Trilling, D. (2016). Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism, 4*(1), 8–23. https://doi.org/10.1080/21670811.2015.1096598

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Elo, S., & Kyngäs, H. (2008). The qualitative content analysis process. *Journal of Advanced Nursing, 62*(1), 107–115. https://doi.org/10.1111/j.1365-2648.2007.04569.x

Gerbner, G., & Gross, L. (1976). Living with television: The violence profile. *Journal of Communication, 26*(2), 172–199. https://doi.org/10.1111/j.1460-2466.1976.tb01397.x

Graneheim, U. H., & Lundman, B. (2004). Qualitative content analysis in nursing research: Concepts, procedures and measures to achieve trustworthiness. *Nurse Education Today, 24*(2), 105–112. https://doi.org/10.1016/j.nedt.2003.10.001

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1*(1), 77–89. https://doi.org/10.1080/19312450709336664

Holsti, O. R. (1969). *Content analysis for the social sciences and humanities*. Addison-Wesley.

Hsieh, H.-F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research, 15*(9), 1277–1288. https://doi.org/10.1177/1049732305276687

Kolbe, R. H., & Burnett, M. S. (1991). Content-analysis research: An examination of applications with directives for improving research reliability and objectivity. *Journal of Consumer Research, 18*(2), 243–250. https://doi.org/10.1086/209256

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology* (4th ed.). SAGE.

Lacy, S., Watson, B. R., Riffe, D., & Lovejoy, J. (2015). Issues and best practices in content analysis. *Journalism & Mass Communication Quarterly, 92*(4), 791–811. https://doi.org/10.1177/1077699015607338

Lasswell, H. D. (1948). The structure and function of communication in society. In L. Bryson (Ed.), *The communication of ideas* (pp. 37–51). Harper and Brothers.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research, 28*(4), 587–604. https://doi.org/10.1111/j.1468-2958.2002.tb00826.x

Mayring, P. (2000). Qualitative content analysis. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research, 1*(2), Article 20. https://doi.org/10.17169/fqs-1.2.1089

Mosteller, F., & Wallace, D. L. (1963). Inference in an authorship problem. *Journal of the American Statistical Association, 58*(302), 275–309. https://doi.org/10.1080/01621459.1963.10500849

Neuendorf, K. A. (2017). *The content analysis guidebook* (2nd ed.). SAGE. https://doi.org/10.4135/9781071802878

Riffe, D., Lacy, S., Watson, B. R., & Fico, F. (2019). *Analyzing media messages: Using quantitative content analysis in research* (4th ed.). Routledge.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science, 58*(4), 1064–1082. https://doi.org/10.1111/ajps.12103

Schreier, M. (2012). *Qualitative content analysis in practice*. SAGE.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

van Atteveldt, W., & Peng, T.-Q. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. *Communication Methods and Measures, 12*(2–3), 81–92. https://doi.org/10.1080/19312458.2018.1458084

Weber, R. P. (1990). *Basic content analysis* (2nd ed.). SAGE. https://doi.org/10.4135/9781412983488