INSIGHTS · GENERATING INSIGHT

Insight Extraction from Unstructured Data Ecosystems

How organisations can extract reliable insight from text, images, audio and other unstructured data, covering the main sources, the methods from word counts to topic models and language models, the need for validation, and the biases and privacy questions that arise.

ModQu research article · 2,601 words · 25 references (APA 7th edition)

Most of what customers, employees and markets say is not stored in neat rows and columns. It sits in emails, reviews, chat logs, call recordings, social media posts, photographs, documents and video. This unstructured data records what people think and do in their own words and images, often at the moment it matters. For insight teams it is a vast and growing resource. It is also difficult to use well, because it must be turned into something that can be counted, compared and checked before it can support decisions.

This article describes the sources of unstructured data that organisations typically hold, the methods available for extracting insight from them, the importance of validation, and the biases and ethical questions that arise. The emphasis throughout is on methods that produce evidence a decision maker can trust, not merely impressive displays.

What makes data unstructured

Structured data is organised in a predefined form, such as a table of transactions with fixed fields for date, product and price. Unstructured data has no such form: a customer's complaint email may mention the product, the price, the delivery, a previous complaint and the customer's feelings, in any order and in any words. Gandomi and Haider (2015) described the growth of unstructured and semi-structured data as one of the defining features of "big data", and noted that methods for analysing text, audio, video and social media had become central to extracting value from it.

Balducci and Marinova (2018) reviewed the use of unstructured data in marketing research. They described the main types as text, images, audio and video, and argued that such data is valuable because it captures rich information about how people think, feel and behave, but that it requires careful conversion into variables before it can be analysed. Wedel and Kannan (2016) similarly argued that the data-rich environment facing marketers demands new analytical skills, combining statistical methods with techniques from computer science.

An ecosystem of sources

Unstructured data in most organisations comes from many sources, held by different teams in different systems. Thinking of these sources as an ecosystem, rather than as separate datasets, helps insight teams see how they relate and where they can be combined.

Sources of unstructured data

Examples, typical extraction methods and the main caution

Source	Examples	Methods	Caution
Customer text	Reviews, survey comments, emails, chat logs	Classification, topics, sentiment	Reviewers are not typical customers
Voice	Call-centre recordings	Speech-to-text, then text methods	Transcription errors vary by accent
Images and video	Customer photos, social media posts	Image recognition, logo detection	May miss context and irony
Documents	Contracts, reports, policies	Entity extraction, search, summaries	Layout and scanning errors
Digital traces	Search terms, notes, app feedback	Categorisation, pattern detection	Behaviour without explanation

Diagram by ModQu ModQu.com

Figure 1. Common sources of unstructured data in organisations, with the methods typically used to extract information from each and the main caution that applies. Diagram by ModQu.

Customer text is usually the largest source: product reviews, open-ended survey answers, complaint emails, chat transcripts and social media posts. It captures customers' views in their own words.

Voice data, especially call-centre recordings, records conversations at the moment customers need help. Advances in automatic speech recognition, such as the large-scale models described by Radford et al. (2023), have made it practical to transcribe large volumes of calls and then analyse the transcripts as text.

Images and video include photographs customers post of products in use, social media images that include brands, and video of in-store behaviour. Liu et al. (2020) developed methods for analysing images that consumers post on social media to measure how brands are portrayed, and showed that these images reveal brand associations that differ from those in text.

Documents such as contracts, reports, policies and research papers contain organisational knowledge that is hard to search and compare.

Digital traces such as search terms and notes attached to transactions are partly structured but often contain free text that records what people were looking for.

Methods for extracting insight

Methods for extracting insight from unstructured data range from simple counting to complex machine learning. Each has a place, and the right choice depends on the question, the volume of data and how much accuracy is needed.

Dictionaries and word counts. The simplest methods count words from predefined lists, such as lists of positive and negative words or words associated with particular topics. Humphreys and Wang (2018) described how dictionary-based methods can measure concepts in consumer research, and emphasised that dictionaries must be validated for the context in which they are used. They are transparent and cheap, but miss context: "not bad" contains a negative word but expresses a positive view.

Supervised classification. In supervised methods, people label a sample of documents, for example as complaints about delivery, price or service, and a model learns to apply the labels to the rest. Grimmer and Stewart (2013) described the strengths of supervised methods for measuring categories that researchers have defined in advance, and the need to check their accuracy against held-out human coding.

Topic models. Unsupervised methods discover patterns without predefined categories. Topic models, such as latent Dirichlet allocation (Blei et al., 2003), identify groups of words that tend to occur together, which often correspond to themes. Blei (2012) described how topic models can be used to explore large collections of documents. Tirunillai and Tellis (2014) applied topic models to online product reviews across several markets and identified the dimensions of quality that consumers discussed, showing how these varied between brands and over time.

Sentiment analysis. Sentiment analysis measures the positive or negative tone of text. Pang and Lee (2008) and Liu (2012) reviewed the methods, from word lists to machine learning. Hartmann et al. (2023) compared many sentiment analysis methods on marketing data and found that modern methods based on pretrained language models were substantially more accurate than dictionary methods, and that the choice of method could change the conclusions drawn in marketing research.

Market structure and needs. Text can also reveal how customers see a market. Netzer et al. (2012) mined online forum discussions to map which car brands consumers mentioned together and how they described them, and found that the resulting market maps were broadly consistent with those from conventional survey data. Timoshenko and Hauser (2019) showed that machine learning could help identify customer needs from online reviews, by filtering out uninformative sentences and reducing the amount of text that human analysts needed to read.

Language models. Methods based on representing words and documents as vectors in a shared space, beginning with word embeddings (Mikolov et al., 2013) and developed through models such as BERT (Devlin et al., 2019), allow computers to recognise that different words can express similar meanings. Large language models extend this, and can classify, summarise and answer questions about text with little or no task-specific training. Gilardi et al. (2023) found that a general-purpose language model matched or exceeded crowd workers on several text annotation tasks at much lower cost.

Start with a question

The volume of unstructured data tempts organisations to analyse first and ask questions later, hoping that patterns will emerge. Some exploration is valuable, and topic models are designed for it. But analysis without a question tends to produce long lists of themes and word clouds that nobody acts on. Balducci and Marinova (2018) recommended that the choice of data and method should follow from the research question, not the other way round.

A good question for unstructured data is specific about the decision it serves. "What are customers saying about us?" is too broad to guide the choice of sources, categories or methods. "Why have contacts about our new tariff risen since March, and which customers are affected?" points to particular channels and time periods, suggests categories to look for, and implies a comparison. Starting with such a question also makes validation easier, because it is clear what the method needs to measure accurately.

Exploration still has a role, particularly when an organisation does not know what it does not know. Running a topic model across a year of complaints, or reading a random sample of chat transcripts, can reveal issues that no one thought to ask about. The discipline is to treat such exploration as the start of an investigation, generating questions that are then answered with more focused and validated measures.

A pipeline from data to insight

Extracting insight from unstructured data is a process, not a single analysis. A typical pipeline has five stages.

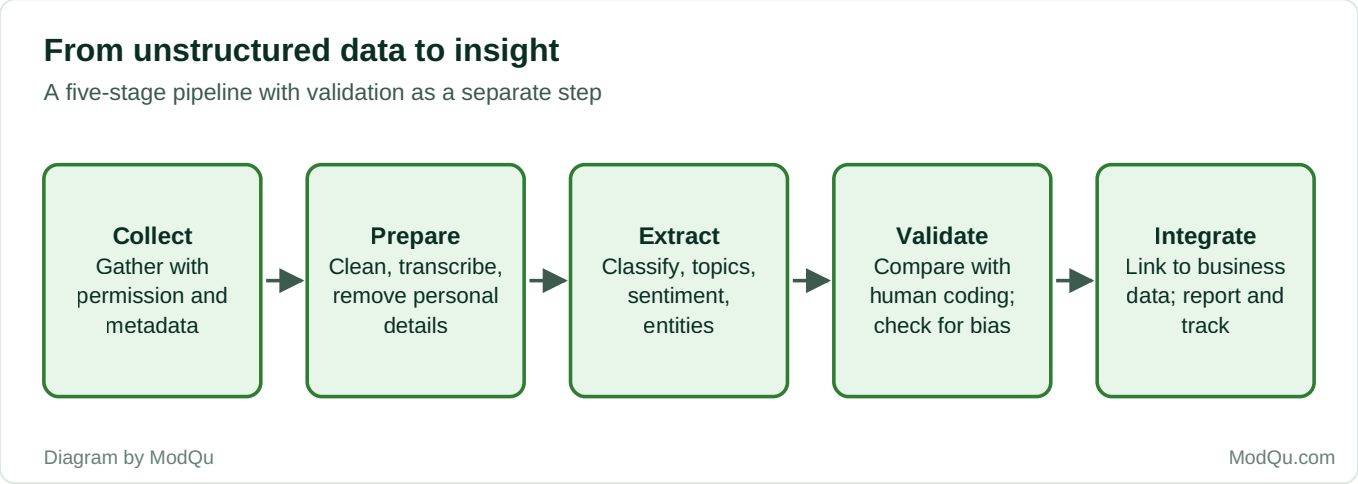


Figure 2. A pipeline for extracting insight from unstructured data. Validation is a separate stage, not an afterthought, because automated methods can be confidently wrong. Diagram by ModQu.

Collect. Gather data from relevant sources, with appropriate permissions and with metadata such as date, channel, product and customer segment, which make later analysis possible.

Prepare. Clean the data by removing duplicates, spam and irrelevant content; transcribe audio; and remove or mask personal details that are not needed.

Extract. Apply methods suited to the question: classification for known categories, topic models for exploration, sentiment measures for tone, and entity extraction to identify products, places or people mentioned.

Validate. Check the results against careful human judgement and look for systematic errors.

Integrate and act. Link the extracted variables to structured data, such as customer value or purchase history, so that themes can be related to outcomes. Report results with representative examples, and track them over time so that changes can be detected.

Validation: the essential step

Automated methods produce numbers that look precise whether or not they are correct. Grimmer and Stewart (2013) stated the principle directly: automated text analysis methods are no substitute for careful reading, and their output must be validated for each application. A classifier trained on one product's reviews may perform poorly on another's; a sentiment dictionary built for movie reviews may misread financial complaints.

Validation means drawing a sample of documents, having people code them carefully, and comparing the automated results with the human coding. Agreement can be measured with statistics such as accuracy or Cohen's kappa. Errors should be examined, not only counted, because patterns in the errors often reveal systematic problems, such as a model that misreads sarcasm or fails on a particular product line. Validation should also be repeated over time, because the language customers use changes, new products and issues appear, and a model that was accurate last year may drift. A small, regular sample of human coding is a cheap way to detect this before it distorts reported trends.

Language models need the same validation, and some additional care. Ji et al. (2023) reviewed research on "hallucination" in language generation, the production of fluent text that is not supported by the source or is factually wrong. A summary of customer feedback produced by a language model may include plausible themes that do not appear in the data, or omit important ones. Lewis et al. (2020) described retrieval-augmented generation, in which a model's answers are grounded in specific retrieved documents, which makes it easier to trace claims back to evidence. For insight work, every important finding produced by a model should be traceable to the documents that support it.

Combining sources

The value of an ecosystem view comes from combining sources. A theme that appears in complaint calls, chat logs, survey comments and public reviews, in similar words, is more credible than one that appears in only one place, because each source has different biases. Combining sources also reveals the path of a problem: a customer may search for help, then use the chat service, then call, then post a review. Linking these contacts, where permissions allow, shows how problems escalate and where they could have been resolved earlier.

Combining unstructured with structured data is often where the most useful insight lies. Text tells an organisation what customers are talking about; transaction and account data tell it what those customers do and how much they are worth. Relating the two shows which issues are associated with cancellations, reduced spending or repeat contacts, and so

which deserve attention first. Wedel and Kannan (2016) argued that integrating data from different sources is one of the main challenges and opportunities of marketing analytics, and noted that it requires attention to data quality, matching and privacy.

Biases and blind spots

Unstructured data is rarely a representative sample of anything. Hu et al. (2009) showed that online product reviews often have a J-shaped distribution, with many very positive and some very negative reviews but few moderate ones, because customers with extreme experiences are more likely to write. Social media users differ from the general population, and complaint channels capture those who complain, not all those who are dissatisfied. Insight from unstructured data should therefore be interpreted with attention to who produced the data and why.

The methods themselves can also be biased. Koenecke et al. (2020) tested five commercial speech recognition systems and found that error rates were substantially higher for Black speakers than for white speakers. Buolamwini and Gebru (2018) found that commercial facial analysis systems classified gender much less accurately for darker-skinned women than for lighter-skinned men. Bender et al. (2021) argued that large language models trained on internet text can reproduce the biases of that text. If an organisation's speech recognition misreads some customers' calls, their complaints may be undercounted, and the resulting insight will be skewed toward those the system hears best.

Privacy and ethics

Unstructured data often contains personal information that people did not expect to be analysed: names and account numbers in emails, faces in photographs, health details in call recordings. Removing or masking personal details before analysis, limiting access, using data only for purposes compatible with those for which it was collected, and being transparent with customers about how their feedback is used are basic safeguards. Data protection laws in many countries impose specific requirements, and research codes such as the ICC/ESOMAR (2016) code require that personal data be protected and used only for agreed purposes. Analysing public social media posts also raises questions, because public is not the same as consenting to analysis.

An example

Consider a hypothetical telecommunications company that receives thousands of customer contacts each week through calls, chat and email, as well as reviews on public sites. Each channel is managed by a different team, and none of them analyses the content systematically.

An insight team brings the sources together. Calls are transcribed automatically, and all text is cleaned and stripped of personal details. The team reads a sample of several hundred contacts and builds a set of categories, then trains a classifier and uses a language model to apply the categories to all contacts. It validates the results on a fresh sample coded by two analysts, finds that agreement is high for most categories but poor for complaints about billing errors, which the model often confuses with questions about price, and refines the category definitions.

The analysis shows that contacts about unexpected charges after the end of promotional offers have tripled over six months, and that customers who make such contacts are far more likely to cancel within three months. The pattern appears in every channel, which gives the team confidence that it is not an artefact of one source. The finding is reported with the numbers, the link to cancellation, and a handful of typical quotations. The company changes how it notifies customers before offers end, and the team tracks the category over the following months to see whether contacts fall.

Key takeaways

Unstructured data, including text, voice, images and documents, records what customers and markets say and do in rich detail, and is a growing source of insight. Methods range from word counts and classifiers to topic models, sentiment analysis and large language models, each suited to different questions. Extracting reliable insight requires a disciplined pipeline in which validation against careful human judgement is a separate and essential step. Insight teams should also remember that unstructured data is rarely representative, that the tools can be biased against some groups, and that the data often contains personal information that must be protected. Used with this care, unstructured data can reveal problems and opportunities that structured data alone would miss.

References

Balducci, B., & Marinova, D. (2018). Unstructured data in marketing. *Journal of the Academy of Marketing Science*, *46*(4), 557–590. https://doi.org/10.1007/s11747-018-0581-x

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

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

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, *81*, 77–91.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences, 120*(30), Article e2305016120. https://doi.org/10.1073/pnas.2305016120

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

Hartmann, J., Heitmann, M., Siebert, C., & Schamp, C. (2023). More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing, 40*(1), 75–87. https://doi.org/10.1016/j.ijresmar.2022.05.005

Hu, N., Zhang, J., & Pavlou, P. A. (2009). Overcoming the J-shaped distribution of product reviews. *Communications of the ACM, 52*(10), 144–147. https://doi.org/10.1145/1562764.1562800

Humphreys, A., & Wang, R. J.-H. (2018). Automated text analysis for consumer research. *Journal of Consumer Research, 44*(6), 1274–1306. https://doi.org/10.1093/jcr/ucx104

ICC/ESOMAR. (2016). *ICC/ESOMAR international code on market, opinion and social research and data analytics*. ESOMAR.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), Article 248. https://doi.org/10.1145/3571730

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences, 117*(14), 7684–7689. https://doi.org/10.1073/pnas.1915768117

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*, 9459–9474.

Liu, B. (2012). *Sentiment analysis and opinion mining*. Morgan & Claypool. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Liu, L., Dzyabura, D., & Mizik, N. (2020). Visual listening in: Extracting brand image portrayed on social media. *Marketing Science, 39*(4), 669–686. https://doi.org/10.1287/mksc.2020.1226

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems, 26*, 3111–3119.

Netzer, O., Feldman, R., Goldenberg, J., & Fresko, M. (2012). Mine your own business: Market-structure surveillance through text mining. *Marketing Science, 31*(3), 521–543. https://doi.org/10.1287/mksc.1120.0713

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1–2), 1–135. https://doi.org/10.1561/1500000011

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. *Proceedings of Machine Learning Research*, *202*, 28492–28518.

Timoshenko, A., & Hauser, J. R. (2019). Identifying customer needs from user-generated content. *Marketing Science*, *38*(1), 1–20. https://doi.org/10.1287/mksc.2018.1123

Tirunillai, S., & Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent Dirichlet allocation. *Journal of Marketing Research*, *51*(4), 463–479. https://doi.org/10.1509/jmr.12.0106

Wedel, M., & Kannan, P. K. (2016). Marketing analytics for data-rich environments. *Journal of Marketing*, *80*(6), 97–121. https://doi.org/10.1509/jm.15.0413