Evaluating latent content within unstructured text: an analytical methodology based on a temporal network of associated topics

Journal of Big Data

METHODOLOGY Open Access

Evaluating latent content within unstructured text: an analytical methodology based on a temporal network of associated topics

Edwin Camilleri* and Shah Jahan Miah

*Correspondence:
edwin.camilleri1@gmail.com;
edwin.camilleri@uon.edu.au
Newcastle Business School,
The University of Newcastle,
Hunter St, Newcastle, NSW
2300, Australia

Abstract

In this research various concepts from network theory and topic modelling are combined, to provision a temporal network of associated topics. This solution is presented as a step-by-step process to facilitate the evaluation of latent topics from unstructured text, as well as the domain area that textual documents are sourced from. In addition to ensuring shifts and changes in the structural properties of a given corpus are visible, non-stationary classes of cooccurring topics are determined, and trends in topic prevalence, positioning, and association patterns are evaluated over time. The aforementioned capabilities extend the insights fostered from stand-alone topic modelling outputs, by ensuring latent topics are not only identified and summarized, but more systematically interpreted, analysed, and explained, in a transparent and reliable way.

Keywords: Natural language processing, Topic modelling, Network theory, Text mining

Introduction

Given the mounting availability of unstructured text [1], topic modelling has become an established approach for the evaluation of textual corpora [2]. Topic modelling is an unsupervised process for identifying interpretable latent topics from a collection of documents, and has been utilized by researchers and analytical practitioners for various applications [3, 4]. Examples include the evaluation of online hotel reviews to understand guest experiences [5], monitoring mass media to detect change in public opinion [6], inspecting SMS messages for spam filtering [7], examining financial disclosure statements for the detection of fraud and misreporting [8], and assessing electronic medical records to support the diagnosis of health conditions [9].

The objective of topic modelling differs from other approaches that aide the evaluation of unstructured text. For example, the purpose of document classification and clustering is to allocate documents to a set of discrete categories, based on the similarity of their underlying content [10, 11]. In contrast, topic models distinguish a fixed number of topics that are present within a corpus. Topics are thus characterized as a distribution of words over a fixed vocabulary (i.e., all terms contained within a corpus), and documents represented as a mixture of topics [3, 12]. Topics are therefore interpreted using the highest-ranking terms of their corresponding term-distributions [2], and documents depicted by the most prevalent topics that they embody [1]. With this, topics that are most and least prevalent within a corpus can be distinguished and assessed (e.g., overall or across a time series).

The objective of this paper is to extend the insights fostered from standalone topic modelling outputs, to ensure topics identified from a corpus are better understood and that textual corpora are more comprehensively evaluated. To do so, we utilize various techniques from network theory and natural language processing, to construct a temporal network of associated topics. This study therefore contributes to analytical practice, as it provides a methodological framework that provisions the capability to generate novel insights from unstructured text. Our proposed framework serves as a systematic process that caters for the following capabilities:

(a) Demonstrate how latent topics are structured within a corpus, and how the structural properties of a corpus change over time;
(b) Measure the role, importance, and popularity of each topic over a time-series covered by a corpus;
(c) Based on consistently cooccurring topics within the same set of documents, measure the extent to which latent topics are associated with each other;
(d) Assemble discrete classes of associated topics, and how these evolve over time

To demonstrate the novelty of the insights provisioned by the proposed framework, an experiment is conducted where it is applied to 20 years of academic literature on consumer behaviour—a multidisciplinary field of study encompassing a myriad of topics, emerging trends, varying themes, and developing issues [13, 14]. This experiment is presented as a case study, which not only establishes the utility of the proposed framework, but also contributes new knowledge to the field of consumer behavior research.

This paper is organized as follows. In the next section an overview and background of topic modelling approaches and network theory are provided. Following this, we present our proposed framework, describe each of its components, and how they improve current capability for the interpretation and analyses of latent topics. We then apply the proposed framework to a corpus of academic literature to demonstrate its capability, before providing a general discussion about the proposed solution, its implications, and future improvements.

Background

Topic modelling applications

The scope for topic modelling covers a diverse range of applications. For example, when applied to electronic medical records, the insights provided by topic models have been leveraged for the support of clinical decision making [9]. This was demonstrated in a recent study [9], where medical conditions were depicted as topics, each represented as a distribution of symptoms. Hence, where patient symptoms are consistent with the highest-ranking terms from the vocabulary of a specific topic, the respective topic serves as a medical condition for patient diagnosis [9].

In addition to being utilized within the healthcare domain, the application of topic modelling has also shown value within the finance and airline industries. In a study that applied topic modelling to financial disclosure statements, the ability to detect intentional financial misreporting was improved [8]. Similarly, when topic models have been applied to aviation incident documents, unreported issues and safety conditions have been disclosed [15]. The scope of topic modelling is comprehensive, and is applicable to textual corpora across a breadth of applications [5–7].

Given that the proposed framework extends the insights fostered from standalone topic modelling outputs, the scope for its application is equally diverse. For instance, if applying the proposed framework to electronic medical records, the identification of topic associations can support the diagnosis of related health conditions that often cooccur with a patient's existing conditions. Similarly, if applied to safety and incident reports, the ability to identify safety conditions that are associated with known incidents enables the implementation of preventative safety measures. All in all, the extended capability to evaluate textual corpora demonstrates that the proposed framework has a diverse range of real-world applications.

The development of probabilistic topic modellings

Latent Dirichlet Allocation (LDA) [16] is a well-established method that is widely considered to be the most common approach for topic modelling [17, 18], and is the cornerstone for the development of more recent topic modelling approaches [19]. With LDA, documents are represented as a mixture of latent topics, and each topic as a distribution of words [16]. Hence, the objective of LDA is to infer (or reverse engineer) latent topics as an observed distribution of words from each document of a corpus [17]. To do so, LDA assumes the following generative process for which a document is produced [16]:

1. Choose $N \sim Poisson$ (ξ)
   Select the number of words N for a document, based on a Poisson distribution
2. Choose $\theta \sim$ Dirichlet (α)
   Select the topic mixture for a document, based on a Dirichlet distribution over a fixed set of topics
3. For each of the N words within a document:

(a) Choose a topic based on the selected topic mixture $z_{\mathrm{n} \sim}$ Multinomial (θ)

(b) Choose a word w_{n} from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on topic $\mathrm{z_n}$

Assuming the aforementioned process, the distribution of a selected number of topics for each document within an observed corpus, as well as the distribution of words for each topic, are inferred.[1] As described by Darling [21], this is achieved by first iterating over each document within a corpus, and randomly assigning each word to a specific topic. Then for each document, iterate over each word w, and for each word iterate

[1] Inference in this instance is based on the Collapsed Gibbs Sampling approach [20].

over each of the K topics. Word w is then reassigned to the most probable topic z, based on the topic assignments for all other words. The foregoing steps are repeated multiple times (e.g., until convergence), to ensure an appropriate standard of topic assignments are achieved.

Although highly effective for modelling the latent structure of a given corpus [17], LDA fails to model correlation between topics, prompting an unrealistic assumption that the presence of a given topic is unrelated to the presence of another [22]. This restriction is attributed to independence assumptions implicit in the Dirichlet distribution when modelling variability among topic proportions [22]. Hence, to dispense an improved representation for the latent topic structure of a given corpus, the Correlated Topic Model (CTM) was introduced by replacing the Dirichlet with the logistic normal distribution [23]. The CTM provides an improved representation of latent topics, and is recommended for the exploration of large document collections [22].

The Structural Topic Model (STM) [23] is a more recent topic modelling approach, which extends the CTM by estimating the relationship between latent topics and document meta-data. For the STM, prior structures for topic prevalence are specified as generalized linear models, conditioned on document specific covariates [23]. This is a key innovation of the STM, in that document meta-data are incorporated as covariates into the topic modelling process, which in turn influence the extent to which topic prevalence can vary across a corpus [24]. The STM thus provides a better (and faster) fit to textual data than the CTM [23].

The generative process for each document d over a vocabulary of size V with K topics can be summarized as follows [1]:

1. Document-level attention is drawn towards each topic from a logistic-normal generalized linear model, based on a vector of document covariates X_d

$$\rightarrow_{\theta_{\mathrm{d}}} |\mathrm{Xd}\gamma, \Sigma \sim \mathrm{LogisticNormal}(\mu = \mathrm{Xd}\gamma, \Sigma)$$

   where X_{d} is a 1-by-p vector, γ is a p-by-K—1 matrix of coefficients, and Σ is K—1-by-K—1 covariance matrix

2. Given a document-level content covariate $\mathrm{y_d}$, establish the document-specific distribution over the terms representing each topic (k) using the baseline word distribution (m), the topic specific deviation $\kappa_{\mathrm{k}}^{(\mathrm{t})}$, the covariate group deviation $\kappa_{\mathrm{yd}}^{(\mathrm{c})}$ and the interaction between the two $\kappa^{(\mathrm{i})}{}_{\mathrm{yd,k}}$

$$\beta_{d,k} \propto \exp(m + k_k^{(t)} + k_{yd}^{(c)} + k_{yd,k}^{(i)})$$

   where m, and each $k_k^{(t)}$, $k_{yd}^{(c)}$ and $k_{yd,k}^{(i)}$ are V-length vectors containing one entry per word in the vocabulary.
   when no convent covariate is present, β can be formed as $\beta_{\mathrm{d,k}} \propto \exp(m + \kappa_{\mathrm{k}}{}^{(\mathrm{t})})$

3. For each word in the document, ($n \in 1, \ldots, N_{\mathrm{d}}$):

- Draw the word's topic assignment based on the document-specific distribution over topics

$$z_{d,n}| \rightarrow_{\theta_d} \sim Multinomial\left(\rightarrow_{\theta_d}\right)$$

- Conditional on the topic chosen, draw an observed word from that topic

$$w_{d,n}|z_{d,n},\beta_{d,k=z_{d,n}} \sim Multinomial\left(\beta_{d,k=z_{d,n}}\right)$$

Network theory

Network theory focuses on the representation of interactions (i.e., edges) between a collection of objects (i.e., nodes) for the purpose of evaluating their structure and dynamics [25, 26]. The scope for network representation and analyses has significantly grown in recent years [27], with examples including the evaluation of information propagation through social media channels, describing the transmission of infectious disease, and measuring viral spread of online malware and spam. A salient feature of the foregoing examples is the development and evolution of their underlying structure, as characterized by change in their properties over time [27].

Temporal networks incorporate nodes and/or edges that are encoded with time-based information (e.g., a timestamp or time window) [28, 29]. Thus, given the temporal characteristics of a node and/or edge, the timing for their inclusion within a network are explicitly determined. As such, network structure is non-stationary, and can develop and evolve over time. Temporal networks are particularly useful when embodying textual information, as the propagation of information can be measured and described [30]. This network is commonly referred to as a temporal text network [30].

For the temporal text network, its representation can be manifested in various ways. For instance, network nodes can reflect time-stamped textual objects (e.g., blog posts, email documents, tweets, or academic papers), and edges materialized by a common property or interaction (e.g., a common author, citation, retweet, or email reply) between two nodes that are included in a network within the same window of time. By comparison, textual objects can also be characterized by time-dependant edges (e.g., documents with a publication date), which at a given point in time connect a pair of discrete entities (e.g., email sender and recipient, tweeter and re-tweeter, author and co-author).

Departing from aforementioned representations of the temporal text network, Abuhay et al. [31] constructed time varying networks of latent topics. According to [31], this was the first study to represent a collection of topics as a network. To do so, the authors sourced 5,982 papers from the International Conference on Computational Science (ICCS), and applied Non-Negative Matrix Factorization (NMF) to reduce the corpus vocabulary to 100 dimensions. The reduced dimensions were taken as topics, each represented as a node within a network, and interconnected by edges if included within the same paper.[2] For papers published in each year of the 17-year period covered by the corpus, a series of static networks were constructed to measure variation in topic connections over time.

[2] As described by Abuhay et al. [31], latent features (or topics) with a weighting greater than 0.05 on the same document were considered as a document co-occurrence. The 0.05 threshold was determined by experimenting with multiple candidate threshold values, and selecting the threshold with the most improved results.

This study takes motivation from [31], in that we also model a series of topic networks over time. However, our approach differs to that of [31] in several ways. For the identification of latent topics, we leverage the STM (as opposed to NMF) to exploit its key capabilities for the purpose of constructing a network of topics. Within our constructed network of topics, undirected (weighted) edges are manifested by topic correlations accounted for by the STM, and nodes weighted with reference to the STM conditioning topic prevalence over time. Hence, the properties of our network are appropriately represented, as they are explicitly modelled throughout the topic estimation process of the STM.

The framework presented in this paper is also unique in the approach taken to analyse our network of topics. In this case, our solution draws insight from evaluating the evolution of the structural properties of a given corpus, measures the development of non-stationary classes of associated topics, and evaluates shifts and changes in topic association patterns over time. The presented approach also jointly evaluates topic importance (i.e., centrality) alongside topic prevalence, enabling key themes and developing narratives to be described. Finally, our framework is presented as a standardized approach for the evaluation of latent topics from unstructured text, irrespective of the domain for which textual corpora are derived. This is described in further detail throughout "Proposed method" section.

Recent advances in network theory

Given the recent progression of various methods applied in Network Theory, their overview merits discussion. The Graph Neural Network (GNN) is a notable area of research, which combines node features with network information to produce predictions within a machine learning paradigm [32]. The Simple Graph Convolution (SGC) is a simpler methodological implementation of the GNN, yet maintains competitive predictive accuracy whilst reducing the dimensionality of large, complex datasets in a way that is more interpretable and accessible in terms of computational resource requirements [32, 33]. In a study that applied the SGC to a network of nodes with classification labels, the ability to accurately classify complex entities (e.g., labelling research papers) by producing a smaller, meaningful set of projections of the network was shown [32].

Although the development of GNNs has led to considerable breakthroughs in node classification tasks, their dependence on large volumes of labelled data for training is significant [33]. Active learning research has thus gained much attention for the training of GNNs with less labelled instances, for which only the most informative instances for labelling are selected [33]. In a study that evaluated how model performance is impacted by the way labelled nodes are sampled, it was shown that the careful selection of nodes improves the accuracy of correct label classification [34]. Given that a uniformly best method for node selection across all network topologies does not currently exist, further research in this area is currently being pursued [34].

Proposed method

Methodological framework

This paper proposes a generic framework that can be applied to textual corpora in any context. The presented approach incorporates existing methods and techniques from

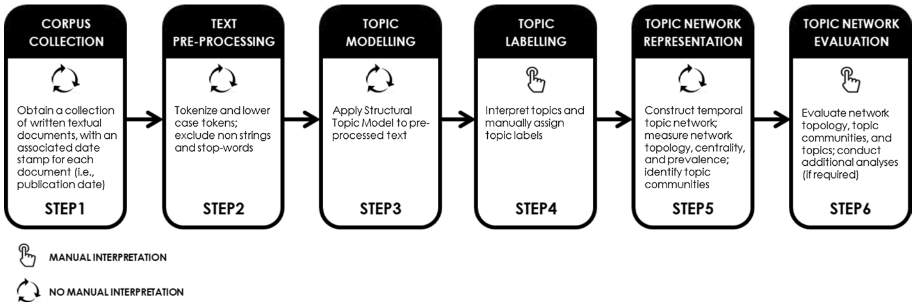

CORPUS COLLECTION
Obtain a collection of written textual documents, with an associated date stamp for each document (i.e., publication date)
STEP1
TEXT PRE-PROCESSING
Tokenize and lower case tokens; exclude non strings and stop-words
STEP2
TOPIC MODELLING
Apply Structural Topic Model to pre-processed text
STEP3
TOPIC LABELLING
Interpret topics and manually assign topic labels
STEP4
TOPIC NETWORK REPRESENTATION
Construct temporal topic network; measure network topology, centrality, and prevalence; identify topic communities
STEP5
TOPIC NETWORK EVALUATION
Evaluate network topology, topic communities, and topics; conduct additional analyses (if required)
STEP6
MANUAL INTERPRETATION
NO MANUAL INTERPRETATION

Fig. 1 Temporal network of associates topics—construction and evaluation

natural language processing and network theory, to provide new ways to evaluate and understand latent topics, as well as the corpus that they are identified from. Given the proposed solution is a methodological framework, we follow Design Science Research (DSR) to evaluate its design.

The DSR evaluation (which is reported in Appendix A) consists of a range of criteria, including validation of the framework's utility. This is specifically demonstrated by conducting an experiment, for which the framework is applied to 20-years of academic literature on consumer behavior. The experiment is thus presented in the form of a case study. In addition to demonstrating the framework's utility, the case study also serves as a blueprint for academic researchers and analytical practitioners to replicate against textual corpora of any kind (e.g., academic literature, blogging material, emails etc.).

As shown in Fig. 1, the proposed framework consists of a series of components, which collectively establish the required process to construct and evaluate a temporal network of associated topics. Each component of the proposed framework is described throughout the sub-sections that follow.

Corpus collection and text pre-processing

To construct a temporal network of associated topics, a corpus with documents encoded with time-based information (e.g., a timestamp for each document) is required. Once collected, textual data are required to be pre-processed for topic modelling to be applied [35]. To do so, text is tokenized into lower case terms, with numbers, punctuation, and whitespace excluded [35]. Stop-words and tokens with a low TF-IDF are also discarded, given the limited relevance that they carry [17, 36]. Furthermore, although often used for text pre-processing, morphological conflation methods such as stemming and lemmatization are avoided, as they have been shown to distort topic modelling results [37]. This is an outcome of terms that share morphological roots being allocated to the same topic, even when varying in meaning [37].

Topic modelling

To identify topics from a given corpus, the STM is applied. Following the approach recommended by Roberts et al. [38], spectral initialization is used in the model fitting process.[3] A covariate is also specified to incorporate temporal information, to ensure measurement of systematic change in topic prevalence can be measured over time [38]. The rank order for each document timestamp (e.g., publication date) is thus taken as a covariate, and specified to have a non-linear relationship in the topic estimation stage by being approximated with a spline.

A common challenge for topic modelling is determining the number of topics to fit [39]. A series of trained models with a different number of topics is therefore required, from which the most parsimonious solution can be selected. In the proposed approach this is based on *semantic coherence* and *exclusivity* metrics. *Semantic coherence* measures co-occurrence between the most probable tokens for each topic, whereas *exclusivity* reflects the most probable tokens for each topic being absent from other topics [40].

Topic labelling

After selecting an appropriate topic modelling solution, each of its topics is then manually labelled. To do so, human intervention is required to interpret the tokens that best characterize each topic, to which topic labels can then be assigned. The metrics reviewed to identify tokens that best characterize topics are a) tokens that account for the highest proportion of a topic's distribution over the corpus vocabulary, b) the *Frex* metric, and c) the *Lift* metric. The *Frex* metric weights tokens by their overall frequency and exclusivity to each respective topic, whereas *Lift* assigns higher weights to tokens appearing less frequently in other topics [23].

Topic network representation

To produce a temporal network of associated topics, a series of static topic networks are produced over equally sized time intervals spanning the corpus. For example, in the experiment conducted and presented in this paper, each document is encoded with a publication date, which span a 20-year period. A network of associated topics is therefore constructed for each year of the 20-year period, with undirected edges established between topics based on the extent to which their prevalence over each document is correlated.,[4],[5] Edge weightings are based on their correlation coefficient, and node weightings based on their combined prevalence over all documents published within each respective time interval.

For the temporal network of associated topics, network topology and node centrality are measured. Topological metrics disclose the structural properties of a network, whereas centrality measures the position, importance, and level of influence that nodes have on network structure [41]. Hence, the time series of the foregoing metrics characterize the evolving structure of information embodied by a corpus, as well as the development for each of its underlying topics over time. The metrics employed to measure network topology and topic centrality are described in Table 1.

[3] The STM recognizes that topic estimation is sensitive to the initialization of model parameters. As such, spectral decomposition based on non-negative matrix factorization using a word co-occurrence matrix is utilized [1].

[4] Topic correlations are based on Spearman's Rank-Order Correlation.

[5] A minimum threshold can also be set, depending on the specific requirements of a given study.

Table 1 Network topology and topic centrality metrics [42–45]

Network topology	
Modularity	Measures non-trivial grouping structure within a network, based on the observed number of edges within a subset of nodes, to the number of edges expected from random assignment $\sum_{k=1}^{k}[f_{kk}(G) - f_{kk}^{*}]^2$ *where* f_{kk}^{*} *is the expected value of* f_{kk} *under some model of random edge assignment*
Transitivity	Measures the extent to which nodes in a network cluster together, based on the ratio of the number of triangles and the number of connected triples $\frac{3\tau_\Delta(G)}{\tau_3(G)}$ *where* $3\tau_\Delta(G)$ *is the number of triangles in the graph, and* $\tau_3(G)$ *is the number of connected triples*
Density	Measures the ratio of the number of edges in a graph to the maximum number of possible edges $\frac{\|E_H\|}{\|V_H\|(\|V_H\|-1)/2}$ *where* $\|E\|$ *is the number of edges and* $\|V\|$ *is the number of nodes in the graph*
Average Path Length	Measures the mean for the shortest paths between all nodes in a network $\frac{1}{n\cdot(n-1)}\cdot\sum_{i\neq j} d(v_i,v_j)$ *where* $d(v_i, v_j)$ *is the shortest path between nodes* v_i *and* v_2*, and n is the number of nodes in the graph*
Diameter	Measures the largest distance between any pair of nodes in a network $max_{u,v}d(u,v)$ *where d(u, v) is the distance between nodes u and v*
Topic centrality	
Betweenness	The fraction of shortest paths that pass through a node $\sum_{s\neq t\neq v\in V}\frac{\sigma(s,t\|v)}{\sigma(s,t)}$ *where* $\sigma(s,t\|v)$ *is the number of shortest paths between s and t that pass through v, and* $\sigma(s,t) = \sum_{v}\sigma(s,t\|v)$
Degree	The number of edges connected to a node $g(v) = \deg(v)$
PageRank	A measure of node importance based on the likelihood of reaching a given node when randomly following links within a network $\alpha\sum_{j}\alpha_{ij}\frac{x_j}{L(j)} + \beta$ *where* $L(j) = \sum_{i} a_{ij}$ *is the number of neighbors of node j, and* α *is a damping factor*

In addition to measuring network topology and topic centrality, community detection is also administered to distinguish discrete subsets of densely connected topics. Within a network of associated topics, communities depict core groups of cooccurring topics, which thereby represent the key subject areas of a corpus. Given that topic associations can vary between different periods of time, community detection is recurrently applied to each time interval spanning the corpus, to ensure change in community membership for each topic is measured. By doing so, the manner in which each subject area evolves over time can be assessed.

To administer community detection several algorithms have been developed, all of which vary in terms of accuracy and computing time for networks with different properties [42]. We therefore refer to the guidelines provided by Yang et al. [42], as they are based on selecting the most appropriate community detection algorithm using the observable properties of a specific network. Further information is provided in Appendix B.

Topic network evaluation

To evaluate the temporal network of associated topics, the time-series for network topology metrics are first assessed, as they illustrate how the structure of information embodied by a corpus are developing over time. This is particularly useful for profiling the development of the domain for which corpora are collected. For example, where a corpus consists of academic papers from a specific disciplinary field, network topological trends can illustrate if the given field is narrowing its scope towards specific areas of study, broadening its coverage by incorporating new topics, or otherwise diverging into specialized areas of research.

To understand why the structure of a corpus is organized and evolving in a particular way, the properties for each topic community (i.e., key subject area) are next measured over a time series. This is achieved by measuring the number of topics assigned to each community, as well as their combined prevalence and centrality at each point in time. In addition to enabling the most prominent and valued subject areas to be recognized, those broadening or reducing their coverage of topics are also identified.

For each topic community, each of their underlying topics are next assessed, by observing shifts and changes in their prevalence and centrality over time. By doing so topics that are emerging and declining in popularity are identified, as is their exposure to other topics. This is particularly useful to characterize the way in which each topic is addressed. For instance, to distinguish exclusive topics that are solely addressed (i.e., low centrality), from those that vary in context and are relevant to other topics and subject areas (i.e., high centrality).

Although topics are effectively characterized by their most characteristic terms (i.e., their content), the circumstances that form the setting for such topics (i.e., their context) are not established in existing topic modelling processes. For example, the term distribution of a topic concerned with smoking will largely skew towards terms such as tobacco, smoke, cigarette, and inhale. Whilst interpretable, the meaning of this topic is largely dependent on its context. If addressed in the context of advertising, sponsorship, and celebrity pastimes, the meaning of this topic varies if otherwise associated with topics such as poor health, cancerous symptoms, and addiction. By establishing the context of topics and how they change over time, key themes and underlying narratives within a corpus can be described. Hence, topics are not only identified and summarized, but are more comprehensively interpreted, evaluated, and explained.

Additional analyses

Following the evaluation of a corpus, topics, and their communities, additional analyses can be conducted. This however is driven by the specific aims, objectives, and overall scope of a particular study. This is demonstrated in the experiment conducted in this paper, which is presented in the next section.

Experiment one: case study

Experimental setting

The objective of this experiment is to establish the utility of the proposed framework, by demonstrating how it extends the insights fostered from standalone topic modelling outputs. Hence, in addition to identifying topics and measuring their prevalence within a corpus (the outputs from standalone topic models), this experiment shows that the proposed solution also ensures changes in the structural properties of a corpus are visible, non-stationary classes of cooccurring topics are measured, and trends in topic prevalence, positioning, and association patterns are evaluated over time. As described in "Proposed method" section, this is achieved by measuring the topological characteristics of a temporal network of topics, identifying and evaluating communities of topics, and then assessing trends, shifts, and changes in prevalence, centrality, and associations between topics over time. By evincing the aforementioned capabilities each component of the proposed approach is tested, and the novelty of the insights that it fosters are shown.

To conduct this experiment, we apply the proposed framework to 20 years of consumer behavior research (hereon also referred to as *'the field'*), as the number of researchers, articles, and topics examined in the field are rapidly growing over time [13]. Moreover, the field is highly exposed to research trends from a myriad of disciplines, which have strengthened its diversity in knowledge and expertise [46]. This field is therefore relevant to a range of topical issues, prominent trends, and disruptive innovations that are continually changing in popularity and interest over time [13, 14]. For the purpose of evaluating the framework proposed in this study, a corpus spanning 20-years of literature from the field lends itself as favourable option. This experiment is presented in the form of a case study, which serves as a blueprint that academic researchers and analytical practitioners can replicate to guide the analyses of textual corpora.

Case study overview

Consumer behavior research is a diverse field of study, with an overwhelming breadth of articles published across a multitude of academic journals [13]. Still, a limited number of studies have applied topic modelling to scholarly work on consumer behavior (an overview of these studies is provided in Appendix C). For the studies that have applied topic modelling to research articles published in the field, corpora have been constrained to articles from a single discipline. Notably, either from marketing or business journals. In contrast, this case study covers the entire realm of consumer behavior from a multidisciplinary perspective, extending previous research by incorporating contributions from technology, psychology, and economics, through to medicine, transport, tourism, and more.

For the studies that have applied topic modelling to research articles published in the field, their analyses are predominantly based on the prevalence of identified topics across an entire corpus, or over a time series covered by a corpus. In contrast, by leveraging the proposed framework, we also measure trends in network topology to demonstrate how the structure of the field is evolving over time, identify topic communities to distinguish the key subject areas that collectively shape the multidisciplinary structure of the field, whist evaluating shifts and changes in topic prevalence, positioning, and associations over time. Hence, the novelty of the insights fostered by the proposed framework ensure new knowledge are contributed to the field, with a more comprehensive understanding of content covered in its underlying literature.

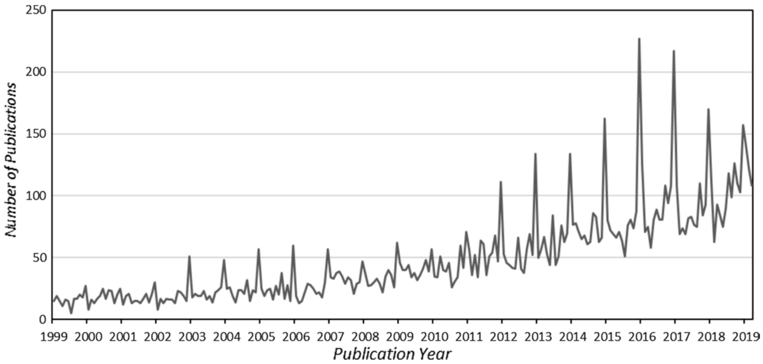


Fig. 2 Monthly time series for consumer behaviour research publications. This figure illustrates the number of consumer behavior research publications each month, from January 1999 to March 2019. In additionto positive trend, seasonality is also visibly present in the time series. The observed peaks each December are largely driven by journalspublishing papers on an annual basis (e.g., Procedia Economics & Finance, Procedia Computer Science)

Data description

To obtain relevant publications for this case study, a search for consumer behavior research was performed on ScienceDirect. The search targeted journal articles that included the term '*consumer*' (or common synonyms for the term '*consumer*') and the term '*behavior*' or '*attitude*' within their title, keywords, or abstract sections.[6] Given the concept of attitude has a significant influence over behavior [47], and hence occupies a central position for consumer behavior research [48], this was also included within the search for publications. A total of 11,841 peer-reviewed articles from 882 journals were retrieved, each published between the 1st of January 1999 to the 31st of March 2019.[7] Given the sample of documents published in 2019 do not cover an entire year, they are retained for analyses applied to the overall corpus, but excluded from any annual time series comparisons. As Fig. 2 shows, consumer behavior research has gained considerable attention in recent years, with over 70% of articles in the corpus published since 2010.

Results

Topic modelling

Following the pre-processing and topic modelling steps described in "Corpus collection and text pre-processing" and "Topic modelling" sections,[8] all text were pre-processed, resulting in a vocabulary of 29,134 unique tokens across 11,841 documents. A series of structural topic models with a different number of topics (ranging from 10 to 120) were then fitted, to which the solution comprised of 70 topics was selected. Details for the evaluation of model fitness that guided the selection of the final solution are reported in

[6] Search Query: (consumer OR customer OR shopper OR buyer OR purchaser OR client) AND (behavior OR behavior OR attitude).

[7] The full body of text for each of the 11,841 articles were retrieved, which were used to establish the corpus.

[8] Text processed using the TidyText package in R [49]; Topic Modelling applied using the STM package in R [1].

Table 2 Characteristics of the temporal network of associated topics

Year	Instances (Number of documents)	Vocabulary (Distinct tokens)	Nodes (Number of topics)	Edges (Topic cooccurrence)
1999	195	15,824	65	468
2000	220	16,927	64	456
2001	211	17,017	60	402
2002	230	17,922	63	420
2003	269	19,323	63	444
2004	303	19,848	58	360
2005	314	20,876	63	393
2006	299	20,704	59	363
2007	401	22,691	65	435
2008	405	22,939	62	435
2009	495	23,975	62	399
2010	509	24,531	65	390
2011	671	25,877	64	408
2012	683	25,897	60	360
2013	793	26,853	61	360
2014	943	27,612	61	339
2015	1019	27,864	62	363
2016	1188	28,262	59	306
2017	1094	28,276	65	372
2018	1229	28,527	63	327

Table 2 lists the characteristics of the temporal network of associated topics, over the 20-year time series. The number of topics (or nodes) listed in Table 2 are those with at least one edge (or cooccurrence) with another topic. Hence, isolated topics not addressed with any of the 70 identified topics are not included in the number of nodes listed in Table 2

Appendix D. The 70 topics are listed within a topic dictionary in Appendix E, alongside their most characteristic terms.

As shown in Appendix E, a diverse range of topics were identified, covering a breadth of knowledge areas, research subjects, methodologies, and applications. Research topics range from *Brand Marketing, Tourism, Impulsive Spending,* and *Loyalty Programs*, to *Green Consumption*, *Queueing Systems*, *Social Media*, and *Health Education*. Several analytical methods were also observed to be modelled as topics, with examples including *Latent Variable Analysis, Statistical Survey Analysis, Data Mining Algorithms*, *Qualitative Research,* and *Choice Modelling.*

Topic network evaluation

With the topics generated from the STM, a temporal network of associated topics was constructed following the process described in "Topic network representation" section. The characteristics of the temporal network of associated topics are provided in Table 2, and its structural properties are presented in Fig. 3.[9] As Fig. 3c shows, the number of edges as a proportion of all potential connections (i.e., network density) continually declined. Thus, given the relationship between reduced density and growth in published research over time ($r = -0.89$, $p = 0.000$), diminishing density appears to be attributed

[9] Network Analysis was conducted using the igraph [50], and tidygraph [51] packages in R.

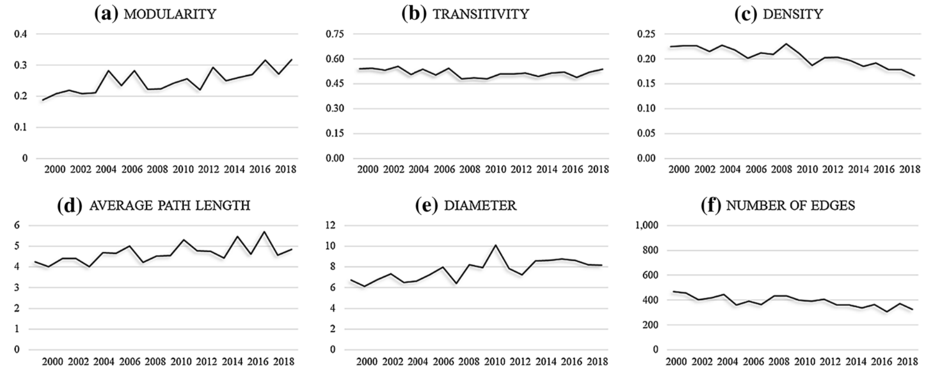


Fig. 3 Time series of network topology metrics. This figure presents the time series for the topological characteristics of the temporal network of associated topics. The decline in networkdensity and total number of edges, together with the rise in modularity, diameter, and average path length, describes and indicates anincrease in clustering of topics over time

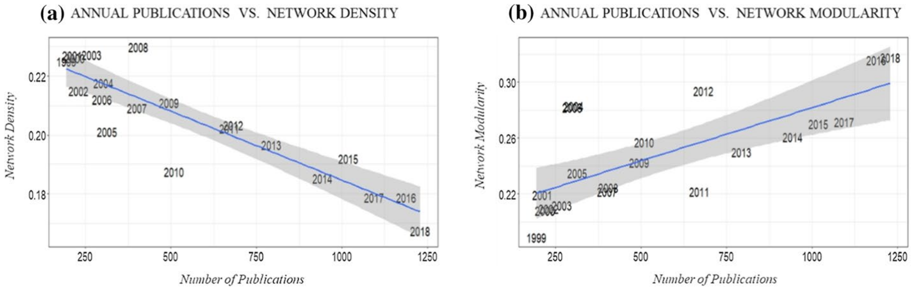


Fig. 4 Reletionship between anuual publications and network topology. This figure presents the relationship between the volume of research publications and network topology (i.e., density and modularity) for eachyear of the 20-year period. As more research articles are published over time, more densely connected subsets of topics (i.e., clustering) arepresent within the network. This demonstrates that the field is continually evolving, with new perspectives and subject areas beingemphasized by researchers on an ongoing basis

to the field continually embracing novel contributions from diverse academic disciplines (Fig. 4a). In this case, research topics addressed from added fields of study are well connected between themselves, yet remain well separated from topics addressed by other academic areas. As a result, both the average path length and network diameter have increased over time (Fig. 3d and e).

To confirm that research topics are increasingly cohesive within discrete topical subsets, as opposed to the network overall, attention is drawn to growth in modularity and consistently high transitivity over time (Fig. 3a and b). High modularity reflects non-trivial grouping structure beyond that expected from random assignment of edges, whereas transitivity measures global clustering by summarizing the relative frequency to which connected triples close to form triangles [43]. For the network representation of the field, clustering is relatively high with more than half of all connected triples closing in this manner, and modularity sustaining positive trend over time.

The clustered structure of the field appears to be manifested by its multidisciplinary disposition. According to [52], multidisciplinarity draws on knowledge from different disciplines to address complex problems. Hence, given the rapidly changing and complex

Table 3 Primary subject areas of consumer behavior research

Subject Area (Topic Community)	Short name	Average number of topics (per year)	Total topic interactions (20-year period)
Consumer Psychology	Psychology	5.6 (±1.7)	12
Marketing	Marketing	10.5 (±2.5)	20
Commercial Strategy	Strategy	3.9 (±1.5)	6
Online & Digital	Digital	5.7 (±0.6)	7
Systems & Technology	Technology	5.8 (±0.8)	8
Sustainability & Preservation	Sustainability	14.8 (±1.7)	22
Health & Wellness	Health	9.7 (±0.7)	12
Economics & Finance	Economics	7.6 (±1.0)	11

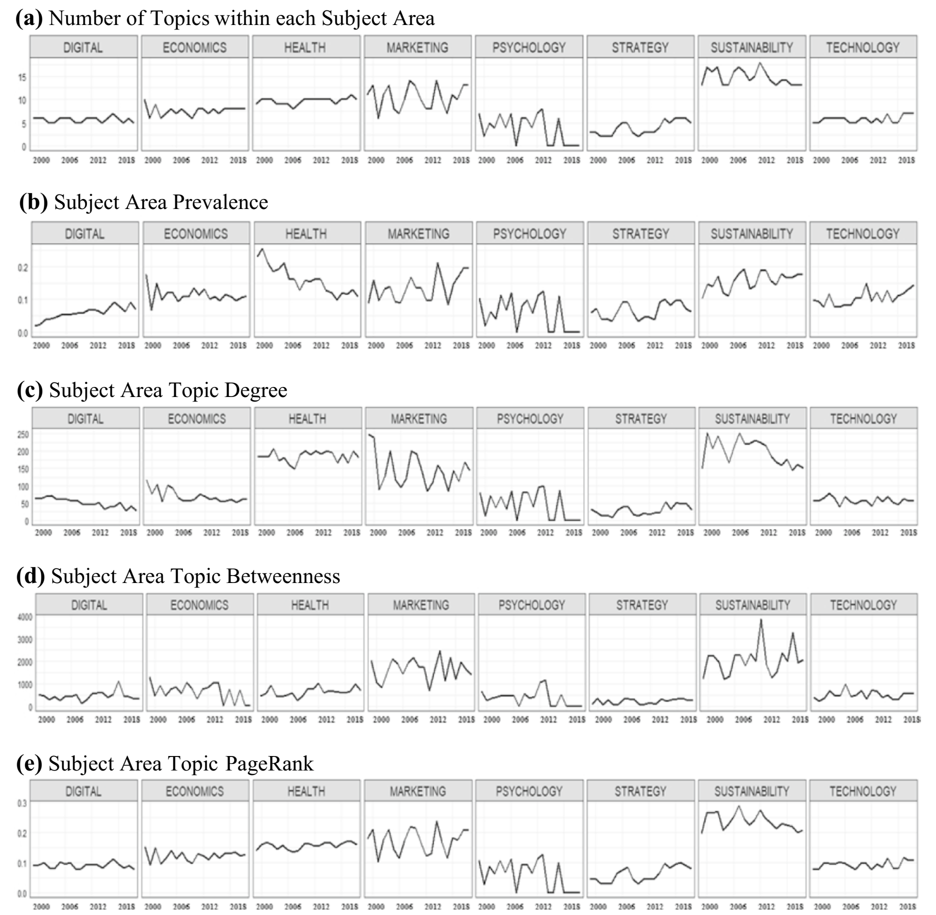
nature of consumer behavior [53], the field continues to sustain dispersed growth over time by fostering research contributions from diverse academic domains. As a result, the field is continually diverging into specialized subject areas.

Topic subject area evaluation

As described in "Topic network representation" section, community detection was administered across the temporal network representation of the field, to distinguish its key subject areas of study. As shown in Table 3, eight major areas of study are embedded within the field, each characterizing a distinct knowledge domain. From *Marketing*, *Economics*, *Technology*, and *Strategy*, to *Sustainability, Health*, *Psychology*, and *Digital*, a diverse range of subject areas are collectively required to encompass the study of consumer behavior.

For each year of the 20-year period, the number of topics assigned to each subject area was measured, as well as their combined prevalence and combined centrality. As shown in Fig. 5a, the context to which research is conducted within the area of *Marketing*, *Consumer Psychology*, and *Sustainability & Preservation* constantly changed, as evidenced by high variability in the number of topics embodied by each area over time. For the aforenamed subject areas, the difference between the total number of topic interactions from the average number of topics per year is evident (see Table 3), reaffirming changing perspectives on consumer behavior.

Overall, *Marketing* and *Sustainability & Preservation* are most centrally positioned within the field. For both subject areas, degree, betweenness, and PageRank are consistently high, emphasising exposure and relevance to a diverse range of topics. Although degree and PageRank are also high for *Health & Wellness* research, the addition of low betweenness indicates intra-topic cohesion with minimal exposure to other subject areas. In other words, health-related topics frequently cooccur together, but are rarely addressed with topics from non-health-related subject areas. Outside the scope of *Consumer Psychology*, which ultimately dissolved as its topics transitioned to other subject areas, *Health & Wellness* is the only subject area that is declining in its prevalence over time (Fig. 5b).



(a) Number of Topics within each Subject Area

(b) Subject Area Prevalence

(c) Subject Area Topic Degree

(d) Subject Area Topic Betweenness

(e) Subject Area Topic PageRank

Fig. 5 Subject area characteristics

Topic evaluation

For each subject area, trends in their centrality and prevalence are largely influenced by the interchange of topics that they each embody. For this reason, all topics within each subject area are next evaluated, to guide the review of their background, development, direction, and status. In this case, each subject area is reviewed in terms of how their topics are connected, and by observing time series curves for the prevalence and centrality of the topics that they each entail. Topic connections for each subject area are presented in Appendix F, and topic time series metrics are displayed in Fig. 6.[10]

Consumer psychology & marketing

As shown in Appendix F.1, *Consumer Psychology* and *Marketing* are two neighbouring fields of study, and are often combined by researchers for the study of consumer behavior. In the latter periods of the 20-year time series, the convergence of the two subject areas embody a unified subset of topics, typifying the developing popularity of customer

[10] Time series curves were smoothed using Loess Regression, to more clearly observe how topic prevalence and centrality measures have changed over time.

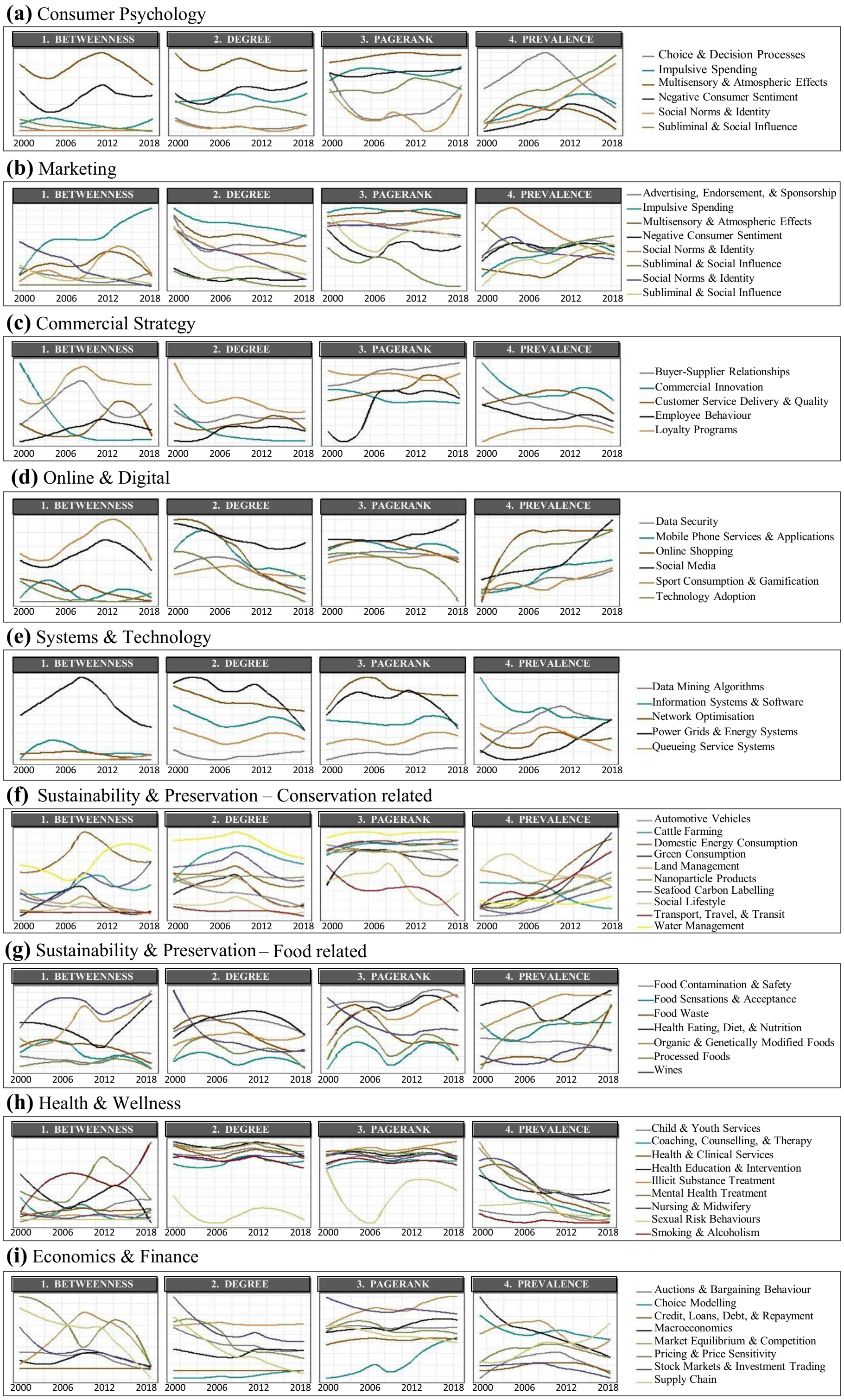


Fig. 6 Topic Prevalence and Centrality, for each Subject Area

centric marketing. Customer centric marketing places the individual at the centre of marketing strategy, design, and delivery, requiring thorough understanding of what makes consumers different and unique [54]. To this end, consumer diversity is addressed by a range of topics within *Consumer Psychology*, including *Consumer Sentiment*, *Choice & Decision Processes*, *Social Influences*, *Impulse Purchases*, *Subliminal Effects*, and *Multisensory Marketing.*

Consumer psychology

As shown in Fig. 6a, over the first decade of the twenty-first century *Choice & Decision Processes* was a growing research topic, and the most prevalent within *Consumer Psychology*. Following its peak in 2010, *Choice & Decision Processes* continually declined over the decade that followed, as researchers increasingly diverted attention toward social influences of consumer behavior. With modern society being shaped by brands, goods and services, as well as advertisements promoting their use and the places that they are displayed, purchased, or subscribed [55], *Social Norms & Identity* and *Subliminal & Social Influences* became the most prevalent topics for *Consumer Psychology* research.

For research in *Consumer Psychology*, topics with low prevalence are most central, as is the case for *Multisensory & Atmospheric Effects*, *Negative Consumer Sentiment*, and *Impulsive Spending.* Whilst the foregoing topics remained affiliated with each other over the 20-year period, their associations with other research topics varied. For *Impulsive Spending*, research was also associated with *Gambling*, *Social Norms & Identity*, and *Subliminal & Social Influence*, and incorporated *Latent Variable Analysis* for their evaluation. In contrast, *Negative Consumer Sentiment* often employed *Qualitative Research Methods*, and was often researched in the context of *Mental Health Treatment* and *Subliminal & Social Influence*. Ultimately, *Multisensory & Atmospheric Effects* was most central, and studied in the context of *Store Retailing*, *Sports Consumption & Gamification*, *Consumer Ethnocentrism*, *Brand Marketing*, and *Advertising, Endorsement & Sponsorship*.

The psychology of how consumers feel, comprehend, and reason between brands, products, and services increases capability to achieve marketing outcomes [56]. By understanding the psychological influences of consumption, customer dynamics are improved by securing more deep, meaningful, and profitable relationships [56]. For firms adopting a customer-centric approach to understand, monitor, and influence consumer behavior, strong market performance has been attained [57]. Hence the increasing popularity of customer-centric marketing, typified by the transition (or merging) of topics initially addressed by *Consumer Psychology*, with those from the *Marketing* subject area.

Marketing

Over recent years both direction and scope of the marketing discipline materially transformed [58]. As topics from *Consumer Psychology* diverged to the area of *Marketing*, those initially embedded within *Marketing* coincided with, or transitioned to other research subject areas, whilst varying in their progression over time. As shown in Fig. 6b, from 2005 research addressing *Store Retailing* and *Multi-Channel Purchase & Promotion* gradually declined over the decade that followed, *Cultural Orientation* and

Advertising, Endorsement & Sponsorship remained relatively consistent, whilst *Tourism*, *Consumer Ethnocentrism*, *Product Presentation*, and *Brand Marketing* became increasingly prevalent. Since 2010 *Brand Marketing* continually gained influence over the flow of information within the *Marketing* domain (i.e., betweenness), despite a reduction in degree over time.

With the rise of globalization [59] the landscape for *Brand Marketing* materially evolved, becoming more complex and important than ever before [60]. Company branding affects all aspects of business [61], and is thus the most central topic within the area of *Marketing* (Fig. 6b). With research addressing the methods that stimulate brand awareness and credibility (e.g., promotions and social influencers), to the factors that impact brand image, trust, and perceived quality (e.g., ethnocentrism and country of origin), *Brand Marketing* is relevant to numerous topics throughout the 20-year period. The most prominent are *Consumer Ethnocentrism*, followed by *Advertising, Endorsement & Sponsorship, Multi-Channel Purchase & Promotion, Loyalty Programs, Multisensory & Atmospheric Effects*, *Wines* and up to 2015, *Store Retailing.*

Much like *Brand Marketing, Cultural Orientation* is also critical to marketers, and highly influenced by the rise of globalization [62]. Globalization portrays the growth of a culturally independent world, exhibited by the presence of a global consumer culture [63]. Nonetheless, unlike *Brand Marketing* taking a central position within the *Marketing* domain, *Cultural Orientation* is an exclusive research topic, as evidenced by remaining consistently prevalent with very low degree. Over the past 20 years *Cultural Orientation* was only observed to cooccur with *Consumer Ethnocentrism*, and on rare occasions with the study of *Tourism*.

Consumer behavior is amongst the major areas of *Tourism* research [64]. Although increasingly prevalent within the Marketing domain, the number of topics associated with Tourism declined from 1999 to 2005 (Fig. 6b2). Since then, *Tourism* was generally researched in the context of *Customer Service Delivery & Quality* and *Multisensory & Atmospheric Effects*, typifying the popularity of customer travel experience. From travel selection and journey activities to the completion of stay, modern consumers expect authentic experiences and instant travel services [65]. Given the influence of personalized services on the tourist experience [66], personalization has become the first priority among modern trends in the travel and tourism industry [67].

Commercial strategy

To remain competitive in the modern customer-led market, organizations require sustained loyalty by strengthening customer relationships from the delivery of high value, personalized experiences [68]. To do so, processes aligning capability, people, and culture with consumer expectations and experiences are required [68]. Customer-centricity therefore extends beyond the area of marketing, entailing continuous engagement between customers, suppliers, employees, and investors [69]. Hence, *Commercial Strategy* is a neighbouring area to the *Marketing* domain, integrating a range of topics. Examples include *Customer Service Delivery & Quality*, *Buyer–Supplier Relationships*, *Employee Behavior*, *Loyalty Programs*, and *Commercial Innovation*.

As shown in Appendix F.3, the number of topics that collectively embody *Commercial Strategy* varied over time. Prior to 2012 *Buyer–Supplier Relationships* and *Commercial*

Innovation transitioned between the *Commercial Strategy* and *Economics & Finance* subject areas, whilst the *Loyalty Programs* and *Customer Service Delivery & Quality* topics were more often affiliated with the *Marketing* domain. Over the years that followed associations between the aforementioned topics remained cohesive, as research addressing commercial strategies that underpin consumer loyalty through improved service culture, quality, and innovation matured. Thus, with service also emerging as a fundamental aspect within the Travel industry, *Tourism* was also strongly affiliated with the *Commercial Strategy* subject area from 2013 to 2017.

Innovation is critical to customer experience, and pertinent to brand reputation, loyalty, and service differentiation [70]. Hence, *Commercial Innovation* is the most popular topic within *Commercial Strategy*, followed by *Customer Service Delivery & Quality, Employee Behavior*, *Buyer–Supplier Relationships*, and then *Loyalty Programs*. Although least prevalent, *Loyalty Programs* ranked highest in betweenness and degree, signifying its relevance to a diverse range of topics. Over the past 20 years loyalty research was predominantly conducted in the context of *Brand Marketing, Multi-channel Purchase & Promotion*, *Customer Service Delivery & Quality*, and *Buyer–Supplier Relationships.*

Online & digital

Over the past 20 years notable trends sequentially materialized within the *Online & Digital* domain. Prevalence for both *Technology Adoption* and *Online Shopping* rapidly increased up to 2005, and then remained constant over the years that followed. Following this, academic researchers focused on *Mobile Phone Services & Applications* through to 2010, and then diverted attention towards *Social Media* (see Fig. 6d). By 2018 *Social Media* was most prevalent, and positioned as a central hub for all *Online & Digital* topics (see Appendix F.3).

As digital technology evolves so too do online security issues [71]. Although the least prevalent topic within the *Online & Digital* area, *Data Security* was consistently accounted for by research addressing *Social Media* and *Mobile Phone Services & Applications*. Meanwhile, *Sports Consumption & Gamification* preserved its position between *Social Media and Multisensory & Atmospheric Effects*, exclusively bridging the broader network with the *Online & Digital* domain. Overall, all *Online & Digital* topics are progressively prevalent and, given their declining degree over time, are increasingly developing into focal topics that are exclusively researched.

Systems & technology

Although a customer-orientated organizational culture is the most important driver of customer-centricity, technological capabilities to personalize customer experiences are similarly important [72]. The *Systems & Technology* subject area is therefore pertinent to the field, and includes a range of topics. Included are *Queueing Service Systems* and *Network Optimisation*, as the impact of waiting time experience is a major factor of consumer satisfaction and brand selection [73]. With consumers also expecting utility service experiences to match those from other industry leaders [74], *Power Grids & Energy Systems* are also embodied within the *Systems & Technology* domain.

From 1999 to 2008 research addressing *Queuing Service Systems* and *Network Optimisation* remained relatively prevalent, *Information Systems & Software* declined, and research on *Data Mining Algorithms* progressively increased. However, research addressing the foregoing topics declined over the decade that followed, as academic interest in *Power Grids & Energy System*s intensified. As shown in Fig. 6e, *Power Grids & Energy Systems* evolved into a focal topic of research, as characterized by growing prevalence and diminishing centrality. By 2018, research on *Power Grids & Energy Systems* was proportionate to *Data Mining Algorithms* and *Information Systems & Software*, co-equally positioned as the most popular topics within the *Systems & Technology* domain.

Sustainability & preservation

Sustainability & Preservation embodies two sub-groups, *Conservation* and *Food* related research. Environmental issues associated with consumer behavior have been extensively examined [75], particularly in relation to lifestyle decisions [76]. As shown in Fig. 6f, whilst *Social Lifestyle* was most prevalent before 2010, researchers have progressively transitioned their attention towards specific issues and activities that underpin pro-environmental behavior. By 2018 *Green Consumption*, *Domestic Energy Consumption*, and *Transport, Travel & Transit* were most prevalent, *Automotive Vehicles*, *Nanoparticle Products*, *Social Lifestyle*, and *Land Management* research then ensued, followed by *Seafood Carbon Labelling*, *Water Management*, and *Cattle Farming*.

Consumer behavioral change is an integral component for the sustainability of natural resources [76]. Notably, as climate change challenges the sustainability of water supplies, the adoption of efficient water management practices remains crucial [77]. Hence, *Water Management* was consistently the most central topic for conservation research, and relevant to a diverse range of topics within the *Sustainability & Preservation* domain. Examples include *Land Management*, *Domestic Energy Consumption*, as well as the processing, waste, contamination, and safety of food.

Much like excessive water consumption, the environmental implications associated with food production and wastage are immense [77]. As shown in Fig. 6g, research addressing *Food Waste* continued to rapidly escalate after 2010, and by 2018 was amongst the most prevalent food-related research topics. *Healthy Eating, Diet & Nutrition* was the most prevalent topic up to 2005, which after a decline between 2006 and 2010 later regained academic interest over the decade that followed. Prevalence for *Organic & Genetically Modified Foods*, *Food Waste*, *Processed Foods*, and *Food Sensations & Acceptance* then succeeded, followed by *Wines* and then *Food Contamination & Safety.*

Although the least prevalent, PageRank for *Food Contamination & Safety* was consistently high, typifying its relevance to pivotal topics of food research (Fig. 6g). Betweenness on the other hand was highest for *Organic & Genetically Modified Foods*, *Wines*, and the *Healthy Eating, Diet & Nutrition* topics. The mixture of research affiliated with the aforementioned topics varied, thus increasing contributions from outside the *Sustainability & Preservation* domain. *Wines* for example was also affiliated with the *Marketing* domain, and often addressed with *Ethnocentrism, Brand Marketing, Multisensory & Atmospheric Effects*, *Tourism*, *Store Retailing*, and *Advertising, Endorsement & Sponsorship*. For *Healthy Eating, Diet & Nutrition*, research also intersected with *Health & Wellness* topics, particularly *Smoking & Alcoholism* and *Health Education & Intervention*.

Health & wellness

Health & Wellness research is relevant to various health-related issues and services, including smoking and alcoholism, counselling, clinical and health education services, mental health, and illicit substance treatment. *Health & Wellness* is remotely positioned within the field, as evidenced by low betweenness for its underlying topics. Still, degree and PageRank were consistently high for most topics, denoting the strong integration between topics within the *Health & Wellness* domain. In 2018 the average degree for *Health & Wellness* topics ($\overline{x}$ =18.0, s = 6.2) was higher than other areas of research, which ranged from 5.4 (±3.9) for *Digital & Online* to 11.5 (±4.9) for *Sustainability & Preservation*.

As shown in Fig. 6h, *Smoking & Alcohol*, *Child & Youth Services*, and *Sexual Risk Behaviors* were less prevalent research topics, yet the extent to which they were studied has remained consistent throughout the 20-year period. Research conducted outside the scope of the aforementioned topics continually declined from 1999 to 2010. Whist the descending trend for these topics persisted over the years that followed, the *Health Education & Intervention* topic regained popularity and later became the most prevalent topic within the *Health & Wellness* domain.

Economics & finance

Throughout the 20-year period dense connections were sustained between *Economics & Finance* topics, during which time this subject area collectively disengaged from the broader network. Figure 6i1 illustrates the decay in betweenness for all topics from 2010, which when converging to zero manifested a disconnected network. The detachment of *Economics & Finance* is attributed to change in context to which its bridging topics were researched. Much like the *Pricing & Price Sensitivity* topic cooccurring with those from *Marketing*,[11] the *Supply Chain* topic was affiliated with those from *Commercial Strategy*[12] and *Systems & Technology*,[13] and the *Stock Markets & Investment Trading* topic with those from the *Commercial Strategy*[14] and *Sustainability & Preservation* subject areas.[15] By 2018 the aforementioned associations diminished, and the context to which the stated topics were addressed was confined to the scope of *Economics & Finance*.

For *Economics & Finance* topics, trends in prevalence varied over time. As shown in Fig. 6i4, *Choice Modelling* and *Macroeconomics* were most prevalent at the onset of the new millennium, and then declined over the years that followed. Despite being increasingly prevalent before 2010, academic interest in both *Price & Pricing Sensitivity* and *Auctioning & Bargaining Behavior* declined thereafter. *Market Equilibrium* also became less prevalent after 2010, however regained popularity from 2015. Whilst *Stock Markets & Trading* and *Credit, Loans & Repayment* were consistently less prevalent, *Supply Chain* remained increasingly popular and by 2018, surfaced as the most popular topic within the *Economics & Finance* domain. *Supply Chain* has become increasingly

[11] Multichannel Purchase & Promotion, Store Retailing and Brand Marketing.

[12] Commercial Innovation.

[13] Power Grids & Energy Systems, Network Optimisation and Queueing Service Systems.

[14] Commercial Innovation and Buyer–Supplier Relationships.

[15] Land Management and Automotive Vehicles.

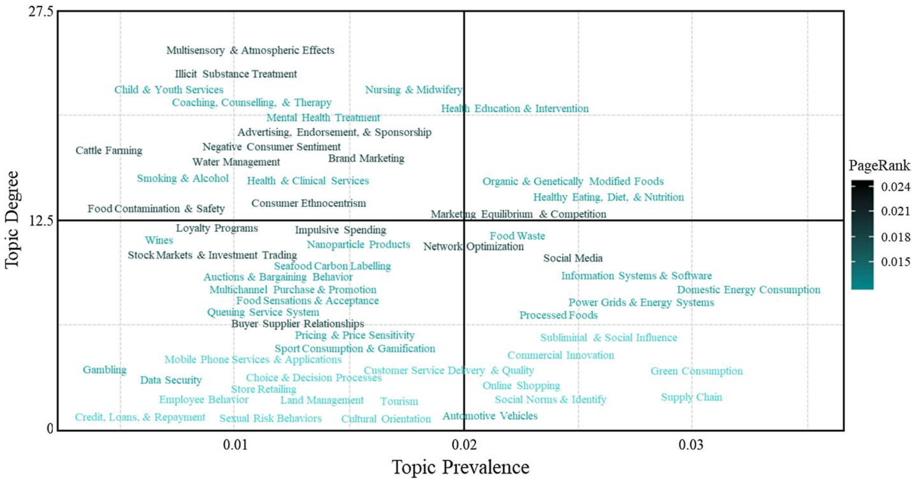


Fig. 7 Topic Prevalence vs Centrality (2018). In this figure topics are positioned based on their prevalence and degree, and their labels coloured based on their PageRank. This enablestopics with varying levels of popularity and exposure (i.e., with other topics) to be distinguished

significant with globalization, and is a major priority in both the manufacturing and service industries [78].

General, transitive & isolated topics

As described in "Topic network representation" section, the field was represented as a temporal network of associated topics, consolidated by edges that manifest topic correlations. For isolated topics with little to no correlation, as well as those excessively transitioning between numerous areas of research or otherwise overly general in nature, their time series for centrality and prevalence were not evaluated with a specific subject area. Centrality and prevalence for such topics are therefore illustrated in Appendix G, which also includes the mean prevalence for all 'other' consumer behavior research topics.

Additional analyses: opportunities for future research

As described in "Additional analyses" section, additional analyses can be performed to address the specific aims, objectives, and overall scope of a particular study. Given this experiment is presented as a literature review, topics can be further evaluated to highlight opportunities for future research. As shown in our analyses, some topics are strategically positioned by connecting different subject areas, others are addressed in isolation, or otherwise highly connected as central or supporting themes. We can therefore distinguish the role of each topic across the field, not only to disclose their current status, but to identify opportunities for how they can be addressed in future study.

Given that the metrics for node centrality vary in the dynamics that they measure, topics are concurrently described in terms of their prevalence, degree, and PageRank in 2018.[16] For all research topics the intersection of the aforementioned metrics are located on Fig. 7, therein conveying the way in which that they have been addressed. For

[16] Only research subjects are evaluated (excluded method-based topics include Latent Variable Analysis, Qualitative Research Methods, Statistical Survey Analysis, Data Mining Algorithms and Choice Modelling).

Table 4 Topic Categorization based on Distinct Prevalence and Centrality

Low degree and low prevalence ($N=4$)	
Given these topics are under-researched, they present opportunity to extend the body of knowledge within the field through further study	
● Credit, Loans, Debt & Repayment (low PageRank) ● Data Security (low PageRank) ● Sport Consumption & Gamification (low PageRank) ● Gambling (moderate PageRank)	
Low degree and high prevalence ($N=7$)	
These topics function as focal content with high popularity and limited associations. Therein lies opportunity to broaden the context to which these topics are addressed, by combining them with other topics	
● Automotive Vehicles (low PageRank) ● Commercial Innovation (low PageRank) ● Green Consumption (low PageRank) ● Online Shopping (low PageRank)	● Social Norms & Identity (low PageRank) ● Subliminal & Social Influence (low PageRank) ● Supply Chain (low PageRank)
High degree and low prevalence *($N=10$)*	
Given their low prevalence but high exposure to other topics, these topics provide added context to more dominant topical subjects. This is particularly the case when PageRank is high	
● Child & Youth Services (moderate PageRank) ● Coaching, Counselling & Therapy (moderate PageRank) ● Food Contamination & Safety (moderate PageRank) ● Health & Clinical Services (moderate PageRank) ● Smoking & Alcoholism (moderate PageRank)	● Cattle Farming (high PageRank) ● Consumer Ethnocentrism (high PageRank) ● Illicit Substance Treatment (high PageRank) ● Multisensory & Atmospheric Effects (high PageRank) ● Water Management (high PageRank)
High degree and high prevalence ($N=4$)	
These topics are focal areas of research that can vary in context. The popularity, influence, and relevance of these topics provide opportunity to elevate topics that are less prevalent and central, by incorporating them for added context in future research	
● Health Education & Intervention (moderate PageRank) ● Healthy Eating, Diet & Nutrition (moderate PageRank) ● Organic & Genetically Modified Foods (moderate PageRank) ● Market Equilibrium & Competition (high PageRank)	

example, within the lower right quadrant are topics exclusively researched, as characterized by high prevalence and low degree. In contrast, less prevalent topics are either infrequently addressed in isolation (i.e., the lower left quadrant) or regularly combined with other topics when researched (i.e., the upper left quadrant). Most *Health & Wellness* topics are located within this quadrant, as they are often addressed with other health-related topics.

Recognizing that several topics within Fig. 7 are positioned along the inner boundaries of each quadrant, those above and below one-half a standard deviation from the mean of each respective metric were distinguished.[17] With topics clearly differentiated by their function, popularity, and level of exposure to other topics, their positioning can guide how they may be appropriately approached in future research. Table 4 presents the four categorical permutations that distinguish topics based on their prevalence and degree. The PageRank classification for each topic is also reported.

[17] Prevalence ($\overline{x} = 0.02$, $s = 0.01$); Degree ($\overline{x} = 10.40$, $s = 6.62$); PageRank ($\overline{x} = 0.02$, $s = 0.003$).

Discission and concluding remarks

Over the past 20 years the field, its subject areas, and their respective topics were observed to vary in the way that they have evolved. As researchers sequentially transitioned their attention from technology adoption to mobile applications and then social media, marketing realigned towards customer centricity, the significance of consumer health education surfaced, customer understanding, experience, and service quality were emphasized, and innovation promoted whilst green and domestic energy consumption were considerably explored.

Outside of the aforementioned trends, the grounds for which topics have been addressed within the field are also diverse. From being strategically positioned to consolidate various topics of study, to remaining isolated or well-connected as central or supporting themes, the positioning of topics within the field present opportunity on how to approach future research. Examples range from drawing attention to topics that are infrequently addressed (e.g., *Gambling Addiction*, *Data Security*, *Debt & Repayment*), to combining under-researched topics with those that are popular and/or relevant to various topical subjects (e.g., *Innovation*, *Subliminal & Social Influence*, *Social Norms & Identity*, *Health Education & Information*).

In this case study, an all-inclusive illustration of the intellectual history, accrued knowledge, development, and direction that the field is progressing towards has been provided. Furthermore, the prevalence, exposure, importance, and context to which topics are addressed have been disclosed. By doing so, new knowledge has been contributed to the field, from the key trends, concealed gaps, and significant issues that impact the study of consumer behavior that have been identified. This is particularly helpful for guiding how future research are selected, planned for, approached, prioritized, and reviewed.

Experiment two: scalability tests

Experimental setting

The solution proposed in this paper is based on a network of latent topics, whereby the size of the network is equivalent to the number of topics identified from a given corpus. Hence, for the experiment presented in "Experiment one: case study" section, the content of 11,841 documents were represented by 70 topics, which were each characterized as a node within the temporal network that was constructed. Recognizing that the number of topics identified from a corpus (and hence the size of a given network) can be vary in size, we performed several tests to evaluate how scalable the proposed solution is across different conditions. In particular, we evaluated how the solution scales across networks of varying size (i.e., the number of nodes) and complexity (i.e., the number of edges). Further, given the proposed solution consists of a series of static topic networks being recursively produced over equally sized time intervals spanning a corpus, this was also considered in our evaluations.

When evaluating the scalability of the proposed solution, all processing was conducted on an 8-core processor with 16 GB RAM. Given that all reported results can be significantly improved by increasing computational resources, it is not the purpose of this experiment to minimise run-time of the proposed solution on the experimental

Table 5 Run-time by number of nodes

Nodes	Edges	Run-time
11,174	23,409	3.092 secs
111,740	234,090	31.09 secs
558,700	1,170,450	115.77 secs
1,117,400	2,340,900	262.63 secs
11,174,000	23,409,000	2,485.04 secs

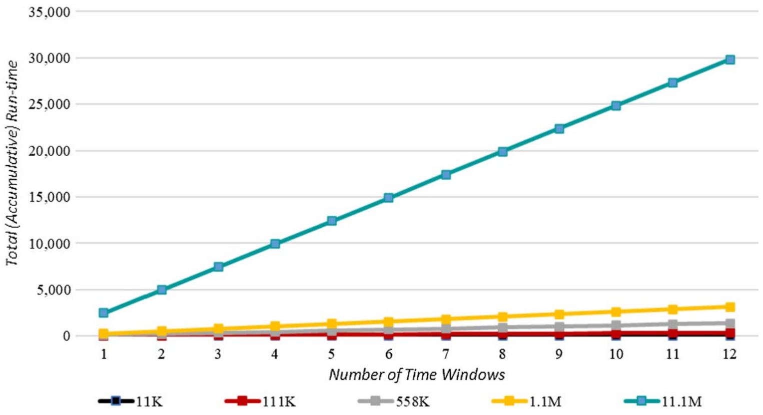


Fig. 8 Run-time by Time-Series Length

datasets, but to instead evaluate the scalability of its application on networks with varying characteristics.

Network size

To assess the application of the proposed solution across networks of varying sizes, we access the Oregon-1 network [79], which consists of 11,174 nodes and 23,409 edges. We then manufactured four additional versions of varying sizes of this network, by iteratively increasing the number its underlying nodes. To ensure the ratio of edges per node remained consistent among the replicated networks, the edges of the replicated nodes within each network were maintained. Table 5 lists the details for each of the five networks, alongside the duration for their construction. The runtime reported in Table 5 also includes computation of network topology and node centrality metrics.

As Table 5 shows, the duration for the construction of each network (including the computation of their topology and node centrality metrics) is magnified for networks of larger size. Moreover, Fig. 8 demonstrates that when computed across a time series, the proposed solution necessitates a linear increase in run-time that is equivalent to the length of the respective time series. This is a result of the recursive nature to which sequences of networks are incrementally computed. As such, even though the reported run-time for each network can be substantially reduced with increased computational resources, run-time is irrespectively extended in proportion to the number of sequences within a given time series.

Table 6 Run-time by network complexity

Dataset	Number of nodes	Number of edges	Edges per node	Run-time (secs)
Oregon-1 [79]	11,174	23,409	2.1	3.092
Oregon-2 [79]	11,461	32,730	2.9	3.476
Gnutella [80]	10,876	39,994	3.7	4.716
Wiki-RFA [81]	10,835	159,388	14.7	7.092

Network complexity

In addition to measuring the run-time for networks of varying size and length (i.e., temporal sequences), we also evaluated the duration to compute networks of varying complexity. To do so we measured the duration to compute four networks that have a similar number of nodes, but vary in their number of edges. Table 6 lists the attributes for each network. As Table 6 shows, run-time is marginally higher for networks with a larger number of edges. Recognizing that the extent to which run-time is influenced by network size, length (i.e., temporal sequences), and complexity, we address this in our discussion within the next section.

Discussion and conclusion

In this paper we combine concepts from network theory and topic modelling to provision a temporal network of associated topics. This solution imparts a systematic process to facilitate the evaluation of latent topics from unstructured text, as well as the domain area that textual documents are sourced from. In addition to ensuring shifts and changes in the structural properties of a given corpus are visible, non-stationary classes of associated topics can be measured, and trends in topic importance (i.e., centrality), prevalence, and association patterns can be evaluated over time. The aforementioned capabilities therefore extend the insights fostered from stand-alone topic modelling outputs, by ensuring latent topics are not only identified and summarized, but more comprehensively interpreted, analysed, and explained.

To evaluate how well the proposed solution facilitates improved understanding of topics identified from unstructured text, it was applied to 20 years of academic literature on consumer behavior, as an experiment that was presented in the form of a case study. The results provisioned by the proposed framework were shown to extend those from standalone topic models, to which the novel insights that it has fostered have contributed new knowledge to this field. In doing so, the utility of the proposed framework was established by the case study, which also serves as a general blueprint that can be replicated to guide the analyses of textual corpora from any domain.

Although facilitating novel insights of topics identified from unstructured text, the solution proposed in this study has its limitations nonetheless. For instance, the interpretation and naming of topics and communities require human intervention, and are based on intuitive judgement. Without the aid of domain knowledge and expertise, and no objective criteria to follow, topic interpretation and conclusions can subjectively vary. Similarly, the interpretation of time series trends for network topology, topic prevalence, and centrality are based on their observation. By incorporating statistical inference to facilitate the evaluation of results, a more objective, rigorous approach for time series analysis can be dispensed, to ensure appropriate conclusions are drawn.

Further, as described in "Proposed method" section the construction of the temporal network of associated topics is based on a series of static topic networks that are produced over equally sized time intervals spanning the corpus. As stated by Michail (2015), incorporating temporal information into a network gives rise to various computational problems and challenges [82]. For the temporal network of associated topics, our experiments showed that the computational requirements induced from recursively constructing a series of networks is equal to the number of time intervals within the applicable time series. Hence the proposed approach necessitates computational complexity when applied to longer-term time series data, particularly for networks of increased size (e.g., number of nodes).

The foregoing limitations suggests that there is scope for improving the capability of the solution presented in this paper. For instance, recent studies have explored solutions to improve the computational efficiency for constructing temporal networks, with examples including the application of distributed computing for efficient network representation [82], as well as designing deep learning frameworks that represent graphs as sequences of timed events [83]. Further, where networks of topics are exceedingly large, the Active Learning and SGC methods described in "Recent advances in network theory" section may also be considered for the purpose of efficiently improving interpretability. The presented framework thus serves as the basis for such capability, to which we invite further contributions for its development in future research.

Appendix A: design methodology evaluation

This study follows the paradigm of Design Science Research (DSR). In comparison to the positivist paradigm emphasising theory development and testing, DSR entails multiple procedures to guide the design, development, and evaluation of innovative solutions, which include method-based models and frameworks [83–86]. The design of the proposed framework is presented with reference to seven guidelines provided by Hevner et al. [85] to effectively conduct DSR.

1. Design as an Artefact: According to [85], a viable artefact takes the form of a construct, model, method, or instantiation. Our proposed solution is thus a method-based artefact, as it provides process guidelines for the evaluation of latent structural processes within unstructured text.
2. Problem Relevance: Given the lack of common methodologies to effectively analyse large volumes of text [41], the provision of such capability provides substantial opportunity for business applications. This opportunity extends to academia alike, as text mining methodologies are under-utilized for the review of academic literature [12].
3. Design Evaluation: Given that the design of a solution must be assessed to demonstrate its quality and effectiveness [85], an observational approach is taken for its evaluation [85]. This is achieved through the application of the solution in an experiment, for which its feasibility and value are demonstrated [87].
4. Research Contributions: This research presents the temporal network of associated topics, which provides a systematic process that ensures shifts and changes in the

structural properties of a given corpus are visible, non-stationary classes of cooccurring topics are measured, and trends in topic prevalence, positioning, and association patterns are evaluated over time. The aforementioned capabilities extend the insights fostered from stand-alone topic modelling outputs, by ensuring latent topics are not only identified and summarized, but more comprehensively interpreted, analysed, and explained.

5. Research Rigor: The proposed design is contingent on rigorous elements from multiple academic fields, including natural language processing, network theory, data mining, and system design. The construction and evaluation of the proposed artefact is thus based on the knowledge base (i.e., theoretical foundations and research methodologies) of the aforementioned fields.
6. Design as a Search Process: To identify the most appropriate components that collectively form the proposed solution, various techniques were iteratively reviewed against the study's research requirements. The search identified a temporal network of associated topics as the most effective solution to reveal the underlying structure and development of a complex system. Various topic modelling approaches were then evaluated, to which the STM was identified as the most appropriate method to support the construction of the proposed approach. Academic research experiments supporting the best approach for text pre-processing and community detection are also taken into account.
7. Communication of Research: The research in this paper is presented as a step-by-step process, and an experiment is presented as a case study to demonstrate how the proposed solution can applied. The case study functions as a blueprint for academic researchers and analytical practitioners to replicate against textual corpora of any kind (e.g., academic literature, blogging material, emails etc.).

Appendix B: community detection algorithm selection

For the selection of the most appropriate community detection algorithm, the recommendations made by [42] are based on the total number of nodes N and mixing coefficient μ of a given network. This mixing coefficient μ measures the extent to which communities are exclusive within a network, and is based on the number of edges connecting a given node to those belonging to a different community, as a proportion of its total degree [42]. Irrespective of N, where $\mu < 0.5$ the *Infomap*, *Label Propagation*, *Multilevel*, and *Walktrap* algorithms are suitable solutions [42]. However, where $\mu > 0.5$ the *Multilevel*, *Walktrap*, and *Springlass* algorithms are suited for $N < 1{,}000$, and the *Multilevel* algorithm for $N > 1{,}000$ [42].

Appendix C: overview of previous review studies

In a review conducted by Wang et al. [88], LDA was applied to 2,031 articles published in the Journal of Consumer Research over a 40-year period. This study identified 16 topics, and the prevalence for each of these were measured over time. By doing so, topics gaining or declining in popularity were distinguished. Recognizing that the review conducted by Wang et al. [88] was applied to a single journal, Cho et al. [18] assembled a corpus of 17,243 research articles from 25 marketing journals, to ensure topic diversity across the entire marketing discipline was covered. After leveraging LDA to identify 100 topics from the corpus, time series trends for the prevalence of each topic were then assessed.

Outside of the aforementioned reviews, the application of topic modelling to academic literature on consumer behavior (or marketing) research have focused on a specific subject area. For example, [89] applied topic modelling to identify 14 topics from 495 articles that focused on the usage large datasets (e.g., big data) on online consumer behavior. Similarly, [90] identified 18 topics from 1,560 articles that address Marketing in the context of Big Data technologies. Given that the study conducted by Cho et al. [18] is the only review to evaluate a corpus of literature covering the entire realm of marketing [18], it is the most comprehensive in terms of corpus size (i.e., number of articles), the number of identified topics (i.e., 100 topics), and domain area coverage (i.e., the entire marketing discipline).

Appendix D: model section & assessment

See Figs. 9 and 10.

Based on the exclusivity and semantic coherence metrics, the STM with K = 70 topics was selected as the most parsimonious solution. Furthermore, to confirm the effectiveness of the STM framework, its results were also compared against those from LDA (also with K = 70 topics). This comparison was conducted as a benchmarking exercise, and was based on the exclusivity and semantic coherence of the topics produced by each of the topic modelling approaches. As described in "Proposed method" section, semantic coherence is maximized when the most probable terms of a topic frequently co-occur. Given that this will naturally occur (by default) when K is low and high-ranking terms are common, semantic coherence should be considered together with topic exclusivity [1]—which is high when high-ranking terms are exclusive to a topic [1]. The two topic modelling approaches were therefore compared using the foregoing metrics with an independent samples t-test. The results showed that the STM outperformed LDA by way of producing more exclusive results $t(123) = 5.40$, $p < 0.0001$, $d = 0.90$ and improved topic coherence $t(135) = 5.36$, $p < 0.0001$, $d = 0.91$.

Appendix E: topic dictionary

Topic Name	Terms
Advertising, Endorsement & Sponsorship	advertising, marketing, sponsorship, celebrity, persuasion, endorsement
Auctions & Bargaining Behavior	Seller, auction, negotiation, price, offer, bidder
Automotive Vehicles	Vehicle, fuel, electric, driving, cars, driverless
Brand Marketing	Brand, marketing, image, extension, product, perceived
Buyer–Supplier Relationships	Relationship, buyer, supplier, exchange, buyersupplier, buyerseller
Cattle Farming	Animal, welfare, milk, farmer, dairy, cattle
Child & Youth Services	Child, family, parents, adolescents, youth, services
Choice & Decision Processes	Decision, search, choice, alternatives, options, prospect
Choice Modelling	Choice, models, utility, probability, attributes, parameters
Coaching, Counselling & Therapy	Coaching, counselor, therapists, interviewing, session, treatment
Commercial Innovation	Entrepreneurship, innovation, creation, capabilities, business, corporate

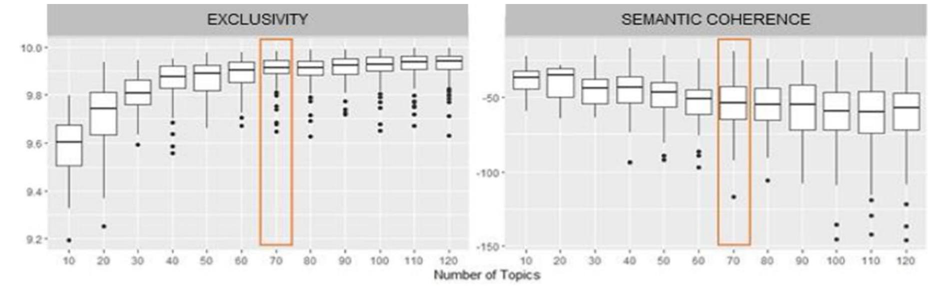


Fig. 9 Model Selection Metrics

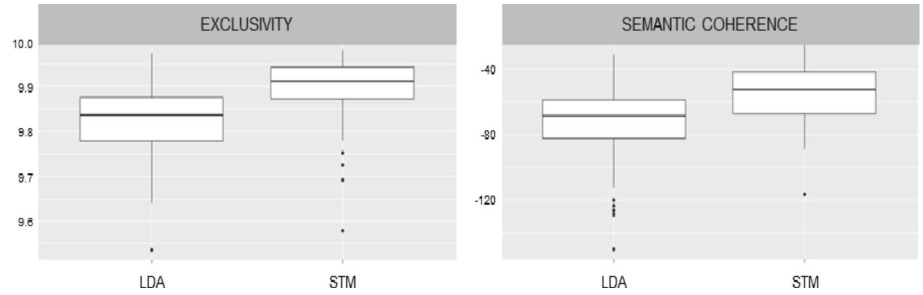


Fig. 10 LDA and STM Comparison

Topic Name	Terms
Consumer Ethnocentrism	Rthnocentrism, clothing, brands, products, luxury, country
Credit, Loans, Debt & Repayment	Credit, card, financial, repayment, debt, loans
Cultural Orientation	Culture, acculturation, ethnic, cross, countries, national
Customer Service Delivery & Quality	Service, satisfaction, quality, relationship, complaint, experience
Data Mining Algorithms	Clusters, fuzzy, mining, algorithm, classifier, neural
Data Security	Privacy, security, personal, disclosure, protection, authentication
Domestic Energy Consumption	Energy, consumption, electricity, efficiency, household, solar
Employee Behavior	Employee, organizational, performance, employee, sales, behaviors
Ethics & Compliance	Public, ethics, audit, compliance, unethical, lawyers
Food Contamination & Safety	Food, safety, temperature, contamination, foodborne, monocytogenes

Topic Name	Terms
Food Sensations & Acceptance	Food, sensory, sensation, liking, taste, neophobic
Food Waste	Food, waste, household, leftovers, wrap, dispose
Gambling	Learning, gambling, habit, casino, lottery, betting
Green Consumption	Green, sustainable, consumption, environmental, recycling, weee
Health & Clinical Services	Health, care, patient, medical, pharmacy, physicians
Health Education & Intervention	Health, intervention, education, program, literacy, screening
Healthy Eating, Diet & Nutrition	Food, nutrition, eating, healthy, diet, calorie
Household Income & Spend	Income, household, spend, home, expenditure, economic
Illicit Substance Treatment	Treatment, drug, substance, abuse, alcohol, addiction
Impulsive Spending	Impulse, compulsive, spend, buying, motivation, control
Information Systems & Software	System, agent, software, cloud, architecture, server
Land Management	Land, market, rural, management, housing, trade
Latent Variable Analysis	Structural, equation, construct, discriminant, convergent, validity
Loyalty Programs	Loyalty, program, marketing, relationship, switching, retention
Macroeconomics	Demand, income, rate, economic, prices, macroeconomics
Market Equilibrium & Competition	Equilibrium, market, firm, competition, optimal, monopoly
Mental Health Treatment	Treatment, therapy, anxiety, cognitive, disorder, depression
Mobile Phone Services & Applications	Mobile, phone, services, devices, smartphone, apps
Multi-Channel Purchase & Promotion	Purchase, multichannel, promotions, channel, sales, coupon
Multisensory & Atmospheric Effects	Music, hedonic, experience, arousal, scent, atmospherics
Nanoparticle Products	Nanoparticles, nano, particle, exposure, cell, concentration
Negative Consumer Sentiment	Emotions, negative, response, anger, embarrassment, guilt
Network Optimisation	Network, problem, algorithm, solution, node, routing
Nursing & Midwifery	Nurses, midwives, nursing, pregnancy, obstetricians, healthcare
Online Shopping	Online, internet, commerce, website, purchase, products
Organic & Genetically Modified Foods	Food, organic, genetically, products, modified, quality
Power Grids & Energy Systems	Power, electricity, load, demand, system, grid
Pricing & Price Sensitivity	Pricing, price, product, purchase, market, discount
Processed Foods	Food, content, properties, processed, emulsifier, fermentation
Product Presentation	Product, visual, images, display, attributes, present
Qualitative Research Methods	Participants, interviews, qualitative, themes, informants, interviewees
Queuing Service Systems	Queueing, theory, queues, lines, system, markov
Seafood Carbon Labelling	Carbon, label, fish, products, seafood, ecolabeling
Sexual Risk Behavior	Risk, safety, sexual, women, workers, condom
Smoking & Alcoholism	Smoking, alcohol, tobacco, drinking, consumption, cigarette
Social Lifestyle	Lifestyle, consumption, social, cultural, society, consumerism
Social Media	Media, social, content, facebook, twitter, youtube
Social Norms & Identity	Norms, social, communities, identity, group, individualizing
Sport Consumption & Gamification	Sport, virtual, engagement, games, experience, gamification
Statistical Survey Analysis	Survey, items, variables, sample, respondents, questionnaire
Stock Markets & Investment Trading	Market, investment, financial, stock, investors, returns
Store Retailing	Store, retail, image, shopping, merchandise, assortment
Subliminal & Social Influence	Cues, influence, implicit, mimicry, subliminal, priming
Supply Chain	Supply, chain, demand, cost, inventory, replenishment
Technology Adoption	Technology, adoption, perceived, acceptance, intention, innovativeness
Tourism	Tourism, tourist, hotel, travel, destination, guests
Transport, Travel & Transit	Travel, transport, trips, transit, taxi, airport
Water Management	Water, management, demand, conservation, wastewater, irrigation
Wines	Wine, authenticity, wineries, cabernet, vino, sauvignon

Appendix F: subject area topic structure

See Fig. 11.

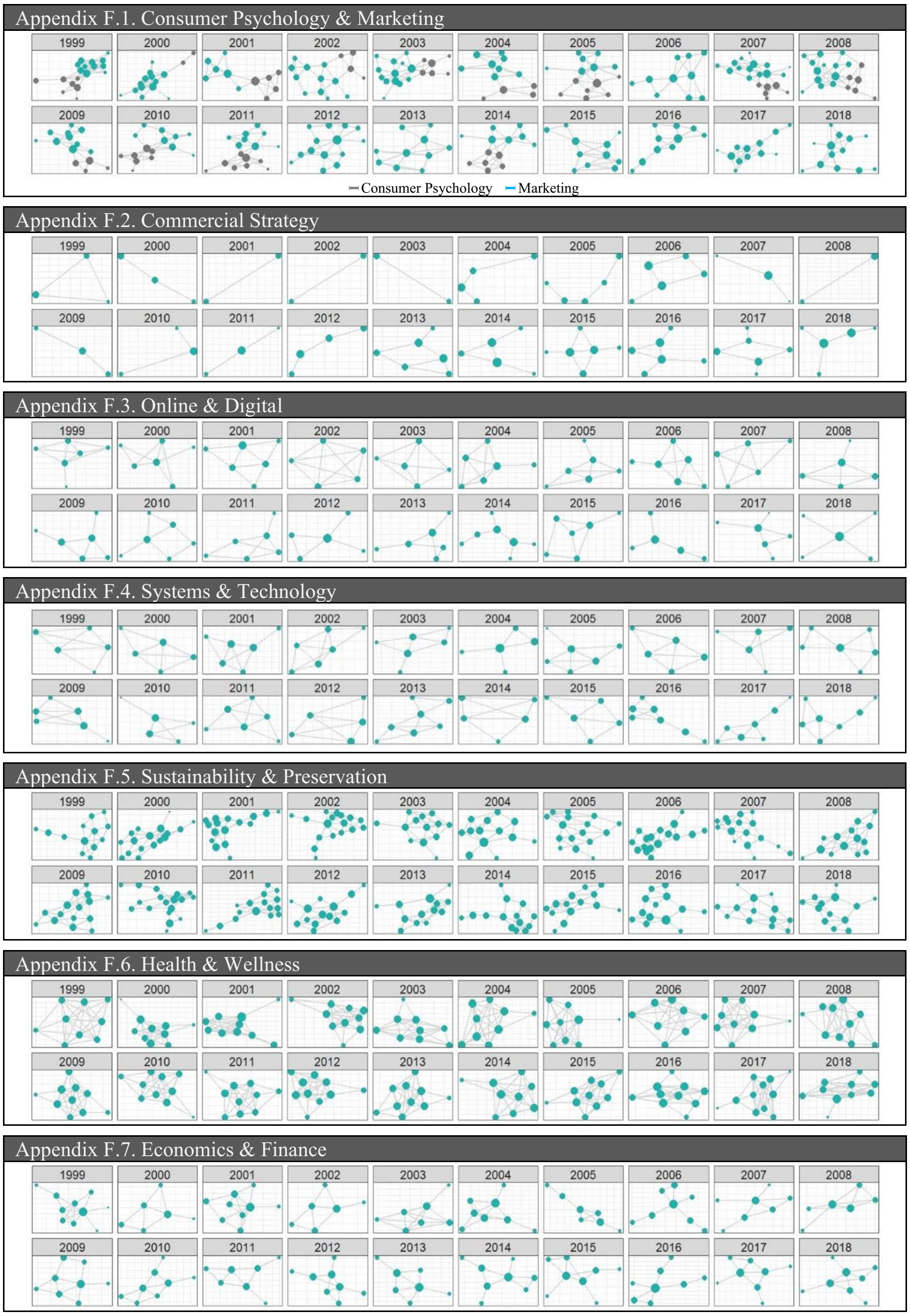


Fig. 11 Topic Community Structure per Subject Area

Appendix G: general, transitive & isolated research topics

See Fig. 12.

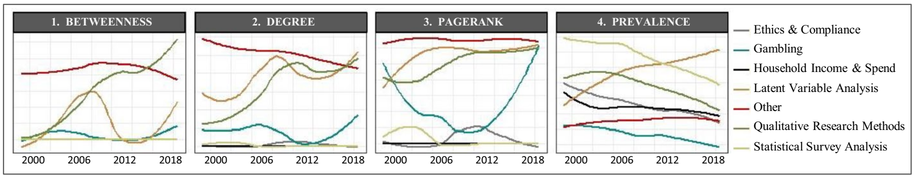


Fig. 12 Topic Prevalence and Centrality, for each Subject Area

Qualitative Research Methods is highly exposed to the flow of information across the field (betweenness) and *Latent Variable Analysis* is the most prevalent. Over the past 20 years *Household Income & Spend*, *Ethics & Compliance*, and *Statistical Survey Analysis* were predominantly isolated from the field with minimal centrality. Their lack of correlation with specific topics suggests broad applicability across all research domains (e.g., *Statistical Survey Analysis*), or proneness to being addressed as an exclusive topic. *Gambling* for example was often isolated from the field, and generally exclusively addressed. When researched with another topic, *Gambling* was combined with *Choice & Decision Processes* and *Impulsive Spending* behavior.

Abbreviations

CTM: Correlated topic model; DSR: Design science research; GNN: Graph Neural Network; LDA: Latent Dirichlet Allocation; NMF: Non-negative matrix factorization; SGC: Simple graph convolution; STM: Structural topic model.

Acknowledgements

Not Applicable

Authors' contributions

EC wrote the paper, designed the framework and executed the case study. SM supervised the research, and contributed to the design and evaluation of the framework. Both authors have revised and approved the final manuscript.

Funding

Not Applicable.

Availability of data and material

The terms and conditions for the use of the dataset applied in this study prohibit the authors from making it publicly available. However, the Elsevier API Services (where the dataset was obtained from) are accessible online. Therefore, by using the same query used in this paper (see paper for details), the same dataset can be reproduced and obtained.

Code availability

Code and aggregated results are available in the following repository: https://github.com/marley1/temporal-network-of-associated-topics

Declarations

Ethics approval and consent to participate

Not Applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 27 April 2021 Accepted: 28 August 2021
Published online: 15 September 2021

References

1. Roberts ME, Stewart BM, Tingley D. stm: An R package for structural topic models. J Stat Softw. 2019;91(2):1–40. https://doi.org/10.18637/jss.v091.i02.
2. O'Callaghan D, Greene D, Carthy J, Cunningham P. An analysis of the coherence of descriptors in topic modelling. Expert Syst Appl. 2015;42(2015):5645–7. https://doi.org/10.1016/j.eswa.2015.02.055.
3. Dieng AB, Ruiz FJR, Blei DM. Topic modeling in embedding spaces. Trans Assoc Comput Linguist. 2020;8(2020):439–53. https://doi.org/10.1162/tacl_a_00325.
4. Li X, Lei L. A bibliometric analysis of topic modelling studies (2000–2017). J Inf Sci. 2019;2019:1–15. https://doi.org/10.1177/0165551519877049.
5. Sutherland I, Sim Y, Lee SK, Byun J, Kiatkawsin K. Topic modeling of online accommodation reviews via latent dirichlet allocation. Sustainability. 2020;12(1821):1–15. https://doi.org/10.3390/su12051821.
6. Yakunin K, Mukhamediev R, Mussabayev R, Buldybayev T, Kuchin Y, Murzakhmetov S, Yunussov R, Ospanova U. Mass media evaluation using topic modelling. In: Alexandrov DA, Boukhanovsky AV, Chugunov AV, Kabanov Y, Koltsova O, Musabirov I, editors. Digital transformation and global society. DTGS 2020. Communications in computer and information science, vol. 1242. Cham: Springer; 2020.
7. Moubayed NA, Breckon T, Matthews P, McGough S. SMS spam filtering using probabilistic topic modelling and stacked denoising autoencoder. In: Villa A, Masulli P, Pons Rivero A, editors. Artificial neural networks and machine learning—ICANN 2016 ICANN 2016. Lecture notes in computer science, vol. 9887. Cham: Springer; 2016. https://doi.org/10.1007/978-3-319-44781-0_50.
8. Brown NC, Crowley RM, Elliot WB. What are you saying? Using topic to detect financial misreporting. J Account Res. 2019;58(1):237–91. https://doi.org/10.1111/1475-679X.12294.
9. Bhattacharya M, Jurkovitz C, Shatkay H. Identifying patterns of associated-conditions through topic models of Electronic Medical Records. In: 2016 IEEE international conference on bioinformatics and biomedicine (BIBM). 2016. p. 466-469. Doi: https://doi.org/10.1109/BIBM.2016.7822561
10. Krishnan. Topic modeling and document clustering; What's the difference? 2016. Retrieved August 18th 2021 from: https://iksinc.online/2016/05/16/topic-modeling-and-document-clustering-whats-the-difference/
11. Mironczuk MM, Protasiewicz J. A recent overview of the state-of-the-art elements of text classification. Expert Syst Appl. 2018;106(2018):36–54. https://doi.org/10.1016/j.eswa.2018.03.058.
12. Asmussen CB, Moller C. Smart literature review: a practical topic modelling approach to exploratory literature review. J Big Data. 2019;6(93):1–18. https://doi.org/10.1186/s40537-019-0255-7.
13. MacInnes D, Folkes V. The disciplinary status of consumer behavior: a sociology of science perspective on key controversies. J Consum Res. 2009;36(6):899–914. https://doi.org/10.1086/644610.
14. Peighambari K, Sattari S, Kordestani A, Oghazi P. Consumer behavior research: a synthesis of the recent literature. SAGE Open. 2016;2016:1–9. https://doi.org/10.1177/2158244016645638.
15. Kuhn K. Using structural topic modeling to identify latent topics and trends in aviation incident reports. Transp Res Part C Emerg Technol. 2018;87(2018):105–22. https://doi.org/10.1016/j.trc.2017.12.018.
16. Blei DM, Ng YA, Jordan IM. Latent dirichlet allocation. J Mach Learn Res. 2003;3(2003):993–1022. https://doi.org/10.5555/944919.944937.
17. Gong J, Abhishek V, Li B. Examining the impact of keyword ambiguity on search advertising performance: a topic model approach. MIS Q. 2018;42(3):805–29. https://doi.org/10.25300/MISQ/2018/14042.
18. Cho YJ, Fu PW, Wu CC. Popular research topics in marketing journals, 1995–2014. J Interact Mark. 2017;40(2017):52–72. https://doi.org/10.1016/j.intmar.2017.06.003.
19. Liu L, Tang L, Dong W, Yao S, Zhou W. An overview of topic modeling and its current applications in bioinformatics. Springerplus. 2016;5(1608):1–22. https://doi.org/10.1186/s40064-016-3252-8.
20. Griffiths T, Steyvers M (2004) Finding scientific topics. In: Proceedings of the National Academy of Sciences of the United States of America, 101, pp. 5228–5235. https://doi.org/10.1073/pnas.0307752101.
21. Darling W. A theoretical and practical implementation tutorial on topic modeling and gibbs sampling. In: Proceedings of the 49th annual meeting of the association for computational linguistics. Human Language Technologies; 2011. p. 642–647
22. Blei DM, Lafferty JD. A correlated topic model of science. Ann Appl Stat. 2007;1(1):17–35. https://doi.org/10.1214/07-AOAS114.
23. Roberts ME, Stewart BM, Tingley D, Airoldi EM. The structural topic model and applied social science. In: Advances in neural information processing systems workshop on topic models: computation, application, and evaluation. 2013. p. 1–4
24. Hu N, Zhang T, Gao B, Bose I. What do hotel customers complain about? Text analysis using structural topic model. Tour Manage. 2019;72(2019):417–26. https://doi.org/10.1016/j.tourman.2019.01.002.
25. Garcia-Robledo A, Diaz-Perez A, Morales-Luna G. Characterization and traversal of large real-world networks. In: Buyya R, Dastjerdi AV, Calheiros RN, editors. Big data, principles and paradigms. Cambridge: Morgan Kaufmann; 2016. p. 119–36.
26. Hamilton WL. Graph representation learning. Synthesis lectures on artificial intelligence and machine learning. Morgan Claypool. 2020;14(3):1–159. https://doi.org/10.2200/S01045ED1V01Y202009AIM046.
27. Rossi RA, Gallagher B, Neville J, Henderson K. Modeling dynamic behavior in large evolving graphs. In: Proceedings of the sixth ACM international conference on web search and data mining (WSDM). 2013. p. 667-676. Doi: https://doi.org/10.1145/2433396.2433479
28. Holme P, Saramaki J. Temporal networks. Phys Rep. 2012;519(3):97–125. https://doi.org/10.1016/j.physrep.2012.03.001.
29. Gao X, Zeng Q, Vega-Oliveros DA, Anghinoni L, Zhao L. Temporal network pattern identification by community modelling. Sci Rep. 2020;10(240):1–12. https://doi.org/10.1038/s41598-019-57123-1.
30. Vega D, Magnani M. Foundations of temporal text networks. Appl Netw Sci. 2018;3(25):1–26. https://doi.org/10.1007/s41109-018-0082-3.

31. Abuhay TM, Kovalchuk SV, Bochenina K, Mbogo GK, Visheratin AA, Kampis G, Krzhizhanovskaya VV, Lees MH. Analysis of publication activity of computational science society in 2001–2017 using topic modelling and graph theory. J Comput Sci. 2018;26(2018):193–204. https://doi.org/10.1016/j.jocs.2018.04.004.
32. Pho P, Mantzaris AV. Regularized Simple Graph Convolution (SGC) for improved interpretability of large datasets. J Big Data. 2020;7(91):1–17. https://doi.org/10.1186/s40537-020-00366-x.
33. Madhawa K, Murata T. Active Learning for Node Classification: An Evaluation. Entropy. 2020;22(1164):1–20. https://doi.org/10.3390/e22101164.
34. Hopwood M, Pho P, Mantzaris AV. Exploring the value of nodes with multicommunity membership for classification with graph convolutional neural networks. Information. 2021;12(4):170. https://doi.org/10.3390/info12040170.
35. Albalawi R, Yeap TH, Benyoucef M. Using topic modeling methods for short-text data: a comparative analysis. Front Artif Intell. 2020;3(42):1–14. https://doi.org/10.3389/frai.2020.00042.
36. Abbasi A, Zhou Y, Deng S, Zhang P. Text analytics to support sense making in social media: a language-action perspective. MIS Q. 2018;42(2):427–64. https://doi.org/10.25300/MISQ/2018/13239.
37. Schofield A, Mimmo D. Comparing apples to apple: the effects of stemmer on topic models. Trans Assoc Comput Linguist. 2016;4(2016):287–300. https://doi.org/10.1162/tacl_a_00099.
38. Roberts ME, Stewart BM, Airoldi EM. A model of text for experimentation in the social sciences. J Am Stat Assoc. 2016;111(515):988–1003. https://doi.org/10.1080/01621459.2016.1141684.
39. Greene D, O'Callaghan D, Cunningham P. How many topics? Stability analysis for topic models. In: Calders T, Esposito F, Hüllermeier E, Meo R, editors. Machine learning and knowledge discovery in databases. ECML PKDD 2014. Lecture notes in computer science, vol. 8724. HBerlin, Heidelberg: Springer; 2014. https://doi.org/10.1007/978-3-662-44848-9_32.
40. Roberts M, Stewart B, Tingley D, Lucas C, Leder-Luis J, Gadarian S, Albertson B, et al. Structural topic models for open ended survey responses. Am J Polit Sci. 2014;58(4):1064–82. https://doi.org/10.1111/ajps.12103.
41. Chau M, Xu J. Business intelligence in blogs: understanding consumer interactions and communities. MIS Q. 2012;36(4):1189–216. https://doi.org/10.2307/41703504.
42. Yang Z, Algesheimer R, Tessone CJ. A comparative analysis of community detection algorithms on artificial networks. Nat Sci Rep. 2016. https://doi.org/10.1038/srep30750.
43. Csardi G, Nepusz T. Statistical network analysis with igraph. New York, NY: Springer; 2016.
44. Wolfram Research, Inc. (www.wolfram.com), Wolfram Language & System, Champaign, IL; 2019
45. Kolaczyk ED. Statistical analysis of network data. methods and models. NY: Springer; 2009.
46. Simonson I, Carmon Z, Dhar R, Drolet A, Nowlis SM. Consumer research: in search of identify. Annu Rev Psychol. 2001;52(2001):249–75. https://doi.org/10.1146/annurev.psych.52.1.249.
47. Cherry K. Attitudes and behavior in psychology. 2019. Retrieved August 10th, 2019 from: https://www.verywellmind.com/attitudes-how-they-form-change-shape-behavior-2795897
48. Foxall GR. Consumer behavior: a practical guide. London: Routledge; 1980.
49. Silge J, Robinson D. tidytext: text mining and analysis using tidy data principles in R. J Open-Source Softw. 2016;1(3):1–3. https://doi.org/10.21105/joss.00037.
50. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal, Complex Systems, 1695. 2006. http://igraph.org
51. Pedersen TL. tidygraph: a tidy API for graph manipulation. R package version 1.1.2. 2019
52. Choi BC, Pak AW. Multidisciplinarity, interdisciplinarity and transdisciplinarity in health research, services, education and policy: definitions, objectives, and evidence of effectiveness. Clin Invest Med. 2006;26(6):351–64 (**PMID: 17330451**).
53. Kennedy B. The challenge of rapidly changing customer behaviour. 2015. Retrieved August 7th, 2019 from https://www.cbsnews.com/news/rapidly-evolving-customer-behavior-to-be-a-game-changer-for-industries/
54. Ernst and Young. The journey toward greater customer centricity. 2013. Retrieved August 22nd, 2019 from https://www.ey.com/Publication/vwLUAssets/The_journey_toward_greater_customer_centricity_-_US/$FILE/Customer_Centricity_Paper_29_April_Final_US.pdf
55. Zukin S, Maguire JS. Consumers and consumption. Ann Rev Sociol. 2004;30(1):173–97. https://doi.org/10.1146/annurev.soc.30.012703.110553.
56. Knilans G. Why it's important to build customer relationships. 2017. Retrieved August 27th, 2019 from: https://www.tradepressservices.com/building-customer-relationships/
57. Verhoef PC, Lemon KN. Successful customer value management: key lessons and emerging trends. Eur Manag J. 2013;31(1):1–15. https://doi.org/10.1016/j.emj.2012.08.001.
58. Kumar V. Evolution of marketing as a discipline: what has happened and what to look out for. J Mark. 2015;79(1):1–9. https://doi.org/10.1509/jm.79.1.1.
59. Ortiz-Ospina E. Is globalization an engine of economic development? 2017. Retrieved November 11th, 2019 from: https://ourworldindata.org/is-globalization-an-engine-of-economic-development
60. Lake L. Why branding is important in marketing. 2019. Retrieved September 9th, 2019 from: https://www.thebalancesmb.com/why-is-branding-important-when-it-comes-to-your-marketing-2294845
61. Stec C. Brand strategy 101: Essentials for strong company branding. 2017. Retrieved September 7th, 2019 from: https://blog.hubspot.com/blog/tabid/6307/bid/31739/7-components-that-comprise-a-comprehensive-brand-strategy.aspx
62. Cleveland M, Mendez JI, Laroche M, Papadopoulos N. Identity, culture, dispositions and behavior: a cross-national examination of globalization and culture change. J Bus Res. 2016;69:1090–102. https://doi.org/10.1016/j.jbusres.2015.08.025.
63. Sobol K, Cleveland M, Laroche M. Globalization, national identity, biculturalism and consumer behavior: a longitudinal study of Dutch consumers. J Bus Res. 2018;82(1):340–53. https://doi.org/10.1016/j.jbusres.2016.02.044.
64. Cohen SA, Prayag G, Moital M. Consumer behaviour in tourism: concepts, influences and opportunities. Curr Issue Tour. 2013;17(10):872–909. https://doi.org/10.1080/13683500.2013.850064.

65. SiteMinder. How consumer behaviour and travel technology are changing each other. n. d. Retrieved September 5th, 2019 from: https://www.siteminder.com/r/trends-advice/hotel-insights/consumer-behaviour-travel-hotel-technology/
66. Buhalis D, Amaranggana A. Smart tourism destinations enhancing tourism experience through personalisation of services. In: Proceedings of the international conference on information and communication technologies in tourism. 2013. p. 553–564. Doi: https://doi.org/10.1007/978-3-319-14343-9_28
67. Stfalcon.com. Top 10 travel industry trends in 2019. 2018. Retrieved September 5th, 2019 from: https://medium.com/swlh/top-10-travel-industry-trends-in-2019-d43d157de7b9
68. Cissowski C. Empowering your people to become a customer-obsessed organisation. 2017. Retrieved September 24th, 2019 from: https://www.ey.com/ie/en/services/advisory/ey-empowering-your-people-to-become-a-customer-obsessed-organisation
69. Ernst & Young. The Digitisation of everything. How organisations must adapt to changing consumer behaviour. 2011. Retrieved June 17th, 2019 from: https://www.ey.com/Publication/vwLUAssets/The_digitisation_of_everything_-_How_organisations_must_adapt_to_changing_consumer_behaviour/$FILE/EY_Digitisation_of_everything.pdf
70. Foroudi P, Jin Z, Gupta S, Melewar TC, Foroudi MM. Influence of innovation capability and customer experience on reputation and loyalty. J Bus Res. 2016;69(2016):4882–9. https://doi.org/10.1016/j.jbusres.2016.04.047.
71. Wakefield R. The influence of user affect in online information disclosure. J Strat Inf Syst. 2013;22(2):157–74. https://doi.org/10.1016/j.jsis.2013.01.003.
72. Yohn DL. 6 Ways to Build a customer-centric culture. 2018. Retrieved September 30th, 2019 from: https://hbr.org/2018/10/6-ways-to-build-a-customer-centric-culture
73. Borges A, Herter MM, Chebat JC. It was not that long!: The effects of the in-store TV screen content and consumers emotions on consumer waiting perception. J Retail Consum Serv. 2015;22(2015):96–106. https://doi.org/10.1016/j.jretconser.2014.10.005.
74. PwC. Customer engagement in an era of energy transformation. 2016. Retrieved September 29th, 2019 from: https://www.pwc.com.au/pdf/web-custtrans-v12-160216.pdf
75. Nguyen TN, Lobo A, Nguyen HL, Phan TTH, Cao TK. Determinants influencing conservation behaviour: perceptions of Vietnamese consumers. J Consum Res. 2016;15(6):560–70. https://doi.org/10.1002/cb.1594.
76. Barr S, Gilg A, Shaw G. Helping people make better choices: exploring the behaviour change agenda for environmental sustainability. Appl Geogr. 2011;31(2):712–20. https://doi.org/10.1016/j.apgeog.2010.12.003.
77. Sun SK, Lu YJ, Gao H, Jiang TT, Du XY, Shen TX, Wu PT, Wang YB. Impacts of food wastage on water resources and environment in China. J Clean Prod. 2018;185(1):732–9. https://doi.org/10.1016/j.jclepro.2018.03.029.
78. Samuel KE, Goury ML, Gunasekaren A, Spalanzani, A.. Knowledge management in supply chain: an empirical study from France. J Strateg Inf Syst. 2011;20(3):283–306. https://doi.org/10.1016/j.jsis.2010.11.001.
79. Leskovec J, Kleinberg J, Faloutsos C. Graphs over time: densification laws, shrinking diameters and possible explanations. In: ACM SIGKDD international conference on knowledge discovery and data mining (KDD). 2005
80. Ripeanu M, Foster I, Iamnitchi A. Mapping the gnutella network: properties of large-scale peer-to-peer systems and implications for system design. IEEE Internet Comput J 2002
81. West R, Paskov HS, Leskovec J, Potts C. Exploiting social network structure for person-to-person sentiment analysis. Trans Assoc Comput Linguist. 2014;2(10):297–310.
82. Michail O. An introduction to temporal graphs: an algorithmic perspective. In: Zaroliagis C, Pantziou G, Kontogiannis S, editors. Algorithms, probability, networks, and games. Lecture notes in computer science, vol. 9295. Cham: Springer; 2015. https://doi.org/10.1007/978-3-319-24024-4_18.
83. Rossi E, Chamberlain B, Frasca F, Eynard D, Monti F, Bronstein M. Temporal graph networks for deep learning on dynamic graphs. 2020. Retrieved from https://arxiv.org/abs/2006.10637
84. Lin YK, Chen H, Brown RA, Li SH, Yang HJ. Healthcare predictive analytics for risk profiling in chronic care: a bayesian multitasking learning approach. MISQ. 2017;41(2):473–95. https://doi.org/10.25300/MISQ/2017/41.2.07.
85. Hevner AR, March ST, Park J, Ram S. Design science in information systems research. MISQ Q. 2004;28(1):75–105. https://doi.org/10.2307/25148625.
86. Gregor S, Hevner AR. Positioning design science research for maximum impact. MIS Q. 2013;37(2):337–55. https://doi.org/10.25300/MISQ/2013/37.2.01.
87. Albert TC, Goes PB, Gupta A. GIST: a model for design and management of content and interactivity of customer-centric web sites. MIS Q. 2004;28(2):161–82. https://doi.org/10.2307/25148632.
88. Wang SX, Bendle TN, Mai F, Cotte J. The journal of consumer research at 40: a historical analysis. J Consum Res. 2015;42(1):5–18. https://doi.org/10.1093/jcr/ucv009.
89. Vanhala M, Lu C, Peltonen J, Sundqvist S, Nummenmaa J, Jarvelin K. The usage of large data sets in online consumer behaviour: a bibliometric and computational text-mining–driven analysis of previous research. J Bus Res. 2020;106(2020):46–59. https://doi.org/10.1016/j.jbusres.2019.09.009.
90. Amado A, Cortez P, Rita P, Moro S. Research trends on Big Data in Marketing: a text mining and topic modelling-based literature analysis. Eur Res Manag Bus Econ. 2018;24(1):1–7. https://doi.org/10.1016/j.iedeen.2017.06.002.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.