Check for updates

Quantitative text analysis

Kristoffer L. Nielbo ®[1] ✉, Folgert Karsdorp[2], Melvin Wevers ®[3], Alie Lassche ®[4], Rebekah B. Baglini ®[5], Mike Kestemont[6] & Nina Tahmasebi ®[7]

Abstract

Text analysis has undergone substantial evolution since its inception, moving from manual qualitative assessments to sophisticated quantitative and computational methods. Beginning in the late twentieth century, a surge in the utilization of computational techniques reshaped the landscape of text analysis, catalysed by advances in computational power and database technologies. Researchers in various fields, from history to medicine, are now using quantitative methodologies, particularly machine learning, to extract insights from massive textual data sets. This transformation can be described in three discernible methodological stages: feature-based models, representation learning models and generative models. Although sequential, these stages are complementary, each addressing analytical challenges in the text analysis. The progression from feature-based models that require manual feature engineering to contemporary generative models, such as GPT-4 and Llama2, signifies a change in the workflow, scale and computational infrastructure of the quantitative text analysis. This Primer presents a detailed introduction of some of these developments, offering insights into the methods, principles and applications pertinent to researchers embarking on the quantitative text analysis, especially within the field of machine learning.

Sections

[1]Center for Humanities Computing, Aarhus University, Aarhus, Denmark. [2]Meertens Institute, Royal Netherlands Academy of Arts and Sciences, Amsterdam, The Netherlands. [3]Department of History, University of Amsterdam, Amsterdam, The Netherlands. [4]Institute of History, Leiden University, Leiden, The Netherlands. [5]Department of Linguistics, Aarhus University, Aarhus, Denmark. [6]Department of Literature, University of Antwerp, Antwerp, Belgium. [7]Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg, Gothenburg, Sweden. ✉e-mail: kln@cas.au.dk

Primer

Introduction

Qualitative analysis of textual data has a long research history. However, a fundamental shift occurred in the late twentieth century when researchers began investigating the potential of computational methods for text analysis and interpretation[1]. Today, researchers in diverse fields, such as history, medicine and chemistry, commonly use the quantification of large textual data sets to uncover patterns and trends, producing insights and knowledge that can aid in decision-making and offer novel ways of viewing historical events and current realities. Quantitative text analysis (QTA) encompasses a range of computational methods that convert textual data or natural language into structured formats before subjecting them to statistical, mathematical and numerical analysis. With the increasing availability of digital text from numerous sources, such as books, scientific articles, social media posts and online forums, these methods are becoming increasingly valuable, facilitated by advances in computational technology.

Given the widespread application of QTA across disciplines, it is essential to understand the evolution of the field. As a relatively consolidated field, QTA embodies numerous methods for extracting and structuring information in textual data. It gained momentum in the late 1990s as a subset of the broader domain of data mining, catalysed by advances in database technologies, software accessibility and computational capabilities[2,3]. However, it is essential to recognize that the evolution of QTA extends beyond computer science and statistics. It has heavily incorporated techniques and algorithms derived from corpus linguistics[4], computer linguistics[5] and information retrieval[6]. Today, QTA is largely driven by machine learning, a crucial component of data science, artificial intelligence (AI) and natural language processing (NLP).

Methods of QTA are often referred to as techniques that are innately linked with specific tasks (Table 1). For example, the sentiment analysis aims to determine the emotional tone of a text[7], whereas entity and concept extraction seek to identify and categorize elements in a text, such as names, locations or key themes[8,9]. Text classification refers to the task of sorting texts into groups with predefined labels[10] — for example, sorting news articles into semantic categories such as politics, sports or entertainment. In contrast to machine-learning tasks that use supervised learning, text clustering, which uses unsupervised learning, involves finding naturally occurring groups in unlabelled texts[11]. A significant subset of tasks primarily aim to simplify and structure natural language. For example, representation learning includes tasks that automatically convert texts into numerical representations, which can then be used for other tasks[12]. The lines separating these techniques can be blurred and often vary depending on the research context. For example, topic modelling, a type of statistical modelling used for concept extraction, serves simultaneously as a clustering and representation learning technique[13–15].

QTA, similar to machine learning, learns from observation of existing data rather than by manipulating variables as in scientific experiments[16]. In QTA, experiments encompass the design and implementation of empirical tests to explore and evaluate the performance of models, algorithms and techniques in relation to specific tasks and applications. In practice, this involves a series of steps. First, text data are collected from real-world sources such as newspaper articles, patient records or social media posts. Then, a specific type of machine-learning model is selected and designed. The model could be a tree-based decision model, a clustering technique or more complex encoder–decoder models for tasks such as translation. Subsequently, the selected model is trained on the collected data, learning to make categorizations or predictions based on the data. The performance of the model is evaluated using predominantly intrinsic performance metrics (such as accuracy for a classification task) and, to a lesser degree, extrinsic metrics that measure how the output of the model impacts a broader task or system.

Three distinct methodological stages can be observed in the evolution of QTA: feature-based models, representation learning models and generative models (Fig. 1). Feature-based models use efficient machine-learning techniques, collectively referred to as shallow learning, which are ideal for tabular data but require manual feature engineering. They include models based on bag-of-words models, decision trees and support vector machines and were some of the first methods applied in QTA. Representation learning models use deep learning techniques that automatically learn useful features from text. These models include architectures such as the highly influential transformer architecture[17] and techniques such as masked language modelling, as used in language representation models such as Bidirectional Encoder Representations from Transformers (BERT)[18]. BERT makes use of the transformer architecture, as do most other large language models after the introduction of the architecture[17]. This shift towards automatic learning representations marked an important advance in natural language understanding. Generative models, trained using autoregressive techniques, represent the latest frontier. These models, such as generative pre-trained transformer GPT-3 (ref. 19), GPT-4 and Llama2 (ref. 20), can generate coherent and contextually appropriate responses and are powerful tools for natural language generation. Feature-based models preceded representation learning, which in turn preceded generative models.

Although these models are temporally ordered, they do not replace each other. Instead, each offers unique methodological features and is suitable for different tasks. The progress from small models with limited computing capacity to today's large models with billions of parameters encapsulates the transformation in the scale and complexity of the QTA.

The evolution of these models reflects the advancement of machine-learning infrastructure, particularly in the emergence and development of tooling frameworks. These frameworks, exemplified by platforms such as scikit-learn[21] and Hugging Face[22], have served as essential infrastructure for democratizing and simplifying the implementation of increasingly sophisticated models. They offer user-friendly interfaces that mask the complexities of the algorithms, thereby empowering researchers to harness advanced methodologies with minimal prerequisite knowledge and coding expertise. The advent of high-level generative models such as GPT-3 (ref. 19), GPT-4 and Llama2 (ref. 20) marks milestones in the progression. Renowned for their unprecedented language understanding and generation capabilities, these models have the potential to redefine access to the sophisticated text analysis by operating on natural language prompts, effectively bypassing the traditional need for coding. It is important to emphasize that these stages represent an abstraction that points to fundamental changes to the workflow and underlying infrastructure of QTA.

This Primer offers an accessible introduction to QTA methods, principles and applications within feature-based models, representation learning and generative models. The focus is on how to extract and structure textual data using machine learning to enable quantitative analysis. The Primer is particularly suitable for researchers new to the field with a pragmatic interest in these techniques. By focusing on machine-learning methodologies, a comprehensive overview of several key workflows currently in use is presented. The focus consciously

Primer

excludes traditional count-based and rule-based methods, such as keyword and collocation analysis. This decision is guided by the current dominance of machine learning in QTA, in terms of both performance and scalability. However, it is worth noting that machine-learning methods can encompass traditional approaches where relevant, adding to their versatility and broad applicability. The experiments in QTA are presented, including problem formulation, data collection, model selection and evaluation techniques. The results and real-world applications of these methodologies are discussed, underscoring the importance of reproducibility and robust data management practices. The inherent limitations and potential optimizations within the field are addressed, charting the evolution from basic feature-based approaches to advanced generative models. The article concludes with a forward-looking discussion on the ethical implications, practical considerations and methodological advances shaping the future of QTA. Regarding tools and software, references to specific libraries and packages are omitted as they are relatively easy to identify given a specific task. Generally, the use of programming languages that are well suited for QTA is recommended, such as Python, R and Julia, but it is also acknowledged that graphical platforms for data analysis provide similar functionalities and may be better suited for certain disciplines.

Experimentation

In QTA, the term experiment assumes a distinct character. Rather than mirroring the controlled conditions commonly associated with randomized controlled trials, it denotes a structured procedure that aims to validate, refine and compare models and findings. QTA experiments provide a platform for testing ideas, establishing hypotheses and paving the way for advancement. At the heart of these experiments lies a model — a mathematical and computational embodiment of discernible patterns drawn from data. A model can be considered a learned function that captures the intricate relationship between textual features and their intended outcomes, allowing for informed decisions on unseen data. For example, in the sentiment analysis, a model learns the association between specific words or phrases and the emotions they convey, later using this knowledge to assess the sentiment of new texts.

The following section delineates the required steps for a QTA experiment. This step-by-step description encompasses everything from problem definition and data collection to the nuances of model selection, training and validation. It is important to distinguish between two approaches in QTA: training or fine-tuning a model, and applying a (pre-trained) model (Fig. 1). In the first approach, a model is trained or fine-tuned to solve a QTA task. In the second approach, a pre-trained model is used to solve a QTA task. Finally, it is important to recognize that experimentation, much like other scientific pursuits, is inherently iterative. This cyclic process ensures that the devised models are not just accurate but also versatile enough to be applicable in real-world scenarios.

Problem formulation

Problem formulation is a crucial first step in QTA, laying the foundation for subsequent analysis and experimentation. This process involves several key considerations, which, when clearly defined beforehand, contributes to the clarity and focus of the experiment. First, every QTA project begins with the identification of a research question. The subsequent step is to determine the scope of the analysis, which involves defining the boundaries of the study, such as the time period, the type of texts to be analysed or the geographical or demographic considerations.

An integral part of this process is to identify the nature of the analytical task. This involves deciding whether the study is a classification task, for example, in which data are categorized into predefined classes; a clustering task, in which data are grouped based on similarities without predefined categories; or another type of analysis. The choice of task has significant implications for both the design of the study and the selection of appropriate data and analytical techniques. For instance, a classification task such as sentiment analysis requires clearly defined categories and suitable labelled data, whereas a clustering task might be used in the exploratory data analysis to uncover underlying patterns in the data.

After selecting data to support the analysis, an important next step is deciding on the level of analysis. QTA can be conducted at various levels, such as the document-level, paragraph-level, sentence-level or even word-level. The choice largely depends on the research question, as well as the nature of the data set.

Classification. A common application of a classification task in QTA is the sentiment analysis. For instance, in analysing social media comments, a binary classification might be employed in which comments are labelled as positive or negative. This straightforward example

Table 1 | Common quantitative text analysis tasks

Task	Description
Sentiment analysis	Analysing the emotional tone behind a text to understand attitudes, opinions and emotions[99]
Emotion detection	Going beyond basic sentiment analysis to detect specific emotions such as happiness, anger or sadness in text[100]
Text classification	Categorizing text into predefined groups or classes, such as spam filtering or news categorization
Document clustering	Grouping similar documents or texts together using unsupervised learning, useful for organizing large data sets
Topic modelling	Identifying topics or themes in large volumes of text methods such as latent Dirichlet allocation[13] and non-negative matrix factorization[101] or, more recently, neural network-based algorithms[14]
Named entity recognition	Identifying and classifying key information in text into predefined categories (such as names and places)[102]
Entity relation extraction	Identifying and classifying relationships between named entities in a text, crucial for building knowledge graphs and understanding complex text structures[103]
Co-occurrence analysis	Identifying terms or concepts that frequently appear together in a text[104]
Trend analysis	Analysing changes in textual data over time to identify patterns or trends
Concept extraction	Identifying key concepts in texts and mapping their relationships, aiding in knowledge discovery and ontology development[9]
Content summarization	Creating a concise summary of large text, maintaining key information and overall meaning[105]

Today, most of these tasks rely on machine learning, although several have equivalents that rely on more traditional count-based and rule-based methods.
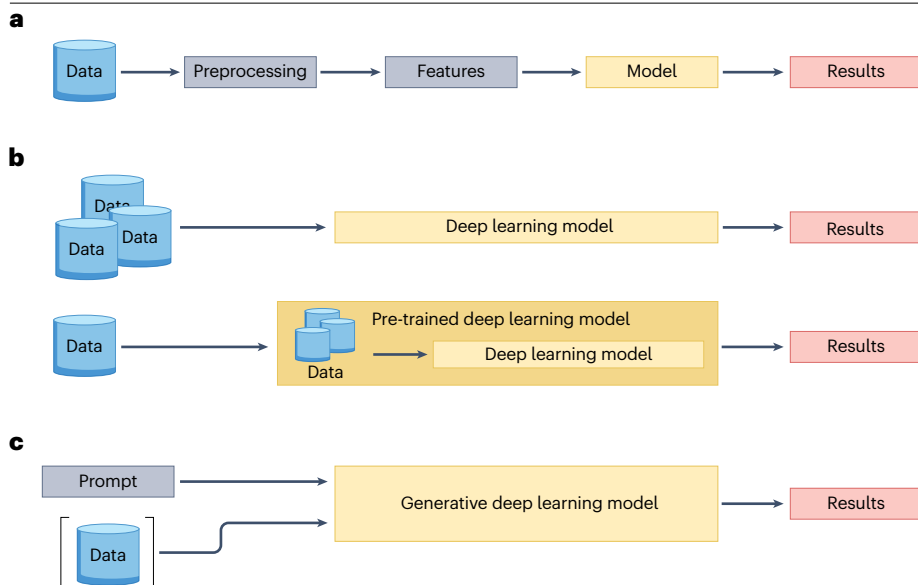
Primer

Fig. 1 | Schematic representation of three predominant approaches in the quantitative text analysis. a, Feature-based models in which data undergo preprocessing to generate features for model training and prediction. b, Representation learning models that can be trained from scratch using raw data or leverage pre-trained models fine-tuned with specific data. c, Generative models in which a prompt guides the generative deep learning model, potentially augmented by external data, to produce a result.

showcases the formulation of a problem in which the objective is clear-cut classification based on predefined sentiment labels. In this case, the level of analysis might be at the sentence level, focusing on the sentiment expressed in each individual comment.

From this sentence-level information, it is possible to extrapolate to general degrees of sentiment. This is often done when companies want to survey their products or when political parties want to analyse their support, for example, to determine how many people are positive or negative towards the party[23]. Finally, from changing degrees of sentiment, one can extract the most salient aspects that form this sentiment: recurring positive or negative sentiments towards price or quality, or different political issues.

Modelling of themes. The modelling of themes involves the identification of prevalent topics, for example, in a collection of news articles. Unlike the emotion classification task, here the researcher is interested in uncovering underlying themes or topics, rather than classifying texts into predefined categories. This problem formulation requires an approach that can discern and categorize emergent topics from the textual data, possibly at the document level, to capture broader thematic elements. This can be done without using any predefined hypotheses[24], or by steering topic models towards certain seed topics (such as a given scientific paper or book)[25]. Using such topic detection tools, it can be determined how prevalent topics are in different time periods or across genre to determine significance or impact of both topics and authors.

Modelling of temporal change. Consider a study aiming to track the evolution of literary themes over time. In this scenario, the problem formulation would involve not only the selection of texts and features but also a temporal dimension, in which changes in themes are analysed across different time periods. This type of analysis might involve examining patterns and trends in literary themes, requiring a longitudinal approach to text analysis, for example, in the case of scientific themes or reports about important events[26] or themes as proxy for meaning change[27]. Often, when longitudinal analysis is considered,

additional challenges are involved, such as statistical properties relating to increasing or decreasing quantity or quality of data that can influence results, see, for example, refs. 28–31.

In similar fashion, temporal analysis of changing data happens in a multitude of disciplines from linguistics, as in computational detection of words that experience change in meaning[32], to conceptual change in history[33], poetry[34], medicine[35], political science[36,37] and to the study of ethnical biases and racism[38–40].

Data

The GIGO principle, meaning 'garbage in, garbage out', is ever present in QTA because without high-quality data even the most sophisticated models can falter, rendering analyses inaccurate or misleading. To ensure robustness in, for example, social media data, its inherently informal and dynamic nature must be acknowledged, often characterized by non-standard grammar, slang and evolving language use. Robustness here refers to the ability of the data to provide reliable, consistent analysis, despite these irregularities. This requires implementing specialized preprocessing techniques that can handle such linguistic variability without losing contextual meaning. For example, rather than discarding non-standard expressions or internet-specific abbreviations, these elements should be carefully processed to preserve their significant role in conveying sentiment and meaning. Additionally, ensuring representativeness and diversity in the data set is crucial; collecting data across different demographics, topics and time frames can mitigate biases and provide a more comprehensive view of the discourse if this is needed. Finally, it is important to pay attention to errors, anomalies and irregularities in the data, such as optical character recognition errors and missing values, and in some cases take steps to remediate these in preprocessing. More generally, it is crucial to emphasize that the quality of a given data set depends on the research question. Grammatically well-formed sentences may be high-quality data for training a linguistic parser; social media could never be studied as people on social media rarely abide by the rules of morphology and syntax. This underscores the vital role of data not just as input but also as an essential component that dictates the success and validity of the analytical endeavour.

Primer

Data acquisition. Depending on the research objective, data sets can vary widely in their characteristics. For the emotion classifier, a data set could consist of many social media comments. If the task is to train or fine-tune a model, each comment should be annotated with its corresponding sentiment label (labels). If the researcher wants to apply a pre-trained model, then only a subset of the data must be annotated to test the generalizability of the model. Labels can be annotated manually or automatically, for instance, by user-generated ratings, such as product reviews or social media posts, for example. Training data should have sufficient coverage of the phenomenon under investigation to capture its linguistic characteristics. For the emotion classifier, a mix of comments are needed, ranging from brief quips to lengthy rants, offering diverse emotional perspectives. Adhering to the principle that there are no data like more data, the breadth and depth of such a data set significantly enhance the accuracy of the model. Traditionally, data collection was arduous, but today QTA researchers can collect data from the web and archives using dedicated software libraries or an application programming interface. For analogue data, optical character recognition and handwritten text recognition offer efficient conversion to machine-readable formats[41]. Similarly, for auditory language data, automatic speech recognition has emerged as an invaluable tool[42].

Data preprocessing. In feature-based QTA, manual data preprocessing is one of the most crucial and time-consuming stages. Studies suggest that researchers can spend up to 80% of their project time refining and managing their data[43]. A typical preprocessing workflow for feature-based techniques requires data cleaning and text normalization. Standard procedures include transforming all characters to lower case for uniformity, eliminating punctuation marks and removing high-frequency functional words such as 'and', 'the' or 'is'. However, it is essential to recognize that these preprocessing strategies should be closely aligned with the specific research question at hand. For example, in the sentiment analysis, retaining emotive terms and expressions is crucial, whereas in syntactic parsing, the focus might be on the structural elements of language, requiring a different approach to what constitutes 'noise' in the data. More nuanced challenges arise in ensuring the integrity of a data set. For instance, issues with character encoding require attention to maintain language and platform interoperability, which means resorting to universally accepted encoding formats such as UTF-8. Other normalization steps, such as stemming or lemmatization, involve reducing words to their root forms to reduce lexical variation. Although these are standard practices, their application might vary depending on the research objective. For example, in a study focusing on linguistic diversity, aggressive stemming may erase important stylistic or dialectal markers. Many open-source software libraries exist nowadays that can help automate such processes for various languages. The impact of these steps on research results underscores the necessity of a structured and well-documented approach to preprocessing, including detailed reporting of all preprocessing steps and software used, to ensure that analyses are both reliable and reproducible. The practice of documenting preprocessing is crucial, yet often overlooked, reinforcing its importance for the integrity of research.

With representation learning and generative techniques, QTA has moved towards end-to-end models that take raw text input such as social media comments and directly produces the final desired output such as emotion classification, handling all intermediate steps without manual intervention[44]. However, removal of non-textual artefacts such as HTML codes and unwanted textual elements such as pornographic material can still require substantial work to prepare data to train an end-to-end model.

Annotation and labelling. Training and validating a (pre-trained) model requires annotating the textual data set. These data sets come in two primary flavours: pre-existing collections with established labels and newly curated sets awaiting annotation. Although pre-existing data sets offer a head-start, owing to their readymade labels, they must be validated to ensure alignment with research objectives. By contrast, crafting a data set from scratch confers flexibility to tailor the data to precise research needs, but it also ushers in the intricate task of collecting and annotating data. Annotation is a meticulous endeavour that demands rigorous consistency and reliability. To ensure inter-annotator agreement (IAA)[45], for example, annotations from multiple annotators are compared using metrics such as Fleiss' kappa (κ) to assess consistency. A high IAA score not only indicates annotation consistency but also lends confidence in the reliability of the data set. There is no universally accepted manner to interpret κ statistics, although $\kappa \geq 0.61$ is generally considered to indicate 'substantial agreement'[46].

Various tools and platforms support the annotation process. Specialized software for research teams provides controlled environments for annotation tasks. Crowdsourcing is another approach, in which tasks are distributed among a large group of people. This can be done through non-monetized campaigns, focusing on volunteer participation or gamification strategies to encourage user engagement in annotation tasks[47]. Monetized platforms, such as Amazon Mechanical Turk, represent a different facet of crowdsourcing in which microtasks are outsourced for financial compensation. It is important to emphasize that, although these platforms offer a convenient way to gather large-scale annotations, they raise ethical concerns regarding worker exploitation and fair compensation. Critical studies, such as those of Paolacci, Chandler and Ipeirotis[48] and Bergvall-Kåreborn and Howcroft[49], highlight the need for awareness and responsible use of such platforms in research contexts.

Provenance and ethical considerations. Data provenance is of utmost importance in QTA. Whenever feasible, preference should be given to open and well-documented data sets that comply with the principles of FAIR (findable, accessible, interoperable and reusable)[50]. However, the endeavour to harness data, especially online, requires both legal and ethical considerations. For instance, the General Data Protection Regulation delineates the rights of European data subjects and sets stringent data collection and usage criteria. Unstructured data can complicate standard techniques for data depersonalization (for example, data masking, swapping and pseudonymization). Where these techniques fail, differential privacy may be a viable alternative to ensure that the probability of any specific output of the model does not depend on the information of any individual in the data set[51].

Recognition of encoded biases is equally important. Data sets can inadvertently perpetuate cultural biases towards attributes such as gender and race, resulting in sampling bias. Such bias compromises research integrity and can lead to models that reinforce existing inequalities. Gender, for instance, can have subtle effects that are not easily detected in textual data[52]. A popular approach to rectifying biases is data augmentation, which can be used to increase the diversity of a data set without collecting new data[53]. This is achieved by applying transformations to existing textual data, creating new and diverse examples. The main goal of data augmentation is to

improve model generalization by exposing it to a broader range of data variations.

Model selection and design

Model selection and design set the boundaries for efficiency, accuracy and generalizability of any QTA experiment. Choosing the right model architecture depends on several considerations and will typically require experimentation to compare the performance of multiple models. Although the methodological trajectory of QTA provides a roadmap, specific requirements of the task, coupled with available data volume, often guide the final choice. Although some tasks require that the model be trained from scratch owing to, for instance, transparency and security requirements, it has become common to use pre-trained models that provide text representations originating from training on massive data sets. Pre-trained models can be fine-tuned for a specific task, for example, emotion classification. Training feature-based models may be optimal for smaller data sets, focusing on straightforward interpretability. By contrast, the complexities of expansive textual data often require representation learning or generative models. In QTA, achieving peak performance is a trade-off among model interpretability, computational efficiency and predictive power. As the sophistication of a model grows, hyperparameter tuning, regularization and loss function require meticulous consideration. These decisions ensure that a model is not only accurate but also customized for research-specific requirements.

Training and evaluation

During the training phase, models learn patterns from the data to predict or classify textual input. Evaluation is the assessment phase that determines how the trained model performs on unseen data. Evaluation serves multiple purposes, but first and foremost, it is used to assess how well the model performs on a specific task using metrics such as accuracy, precision and recall. For example, knowing how accurately the emotion classifier identifies emotions is crucial for any research application. Evaluation of this model also allows researchers to assess whether it is biased towards common emotions and whether it generalizes across different types of text sources. When an emotion classifier is trained on social media posts, a common practice, its effectiveness can be evaluated on different data types, such as patient journals or historical newspapers, to determine its performance across varied contexts. Evaluation enables us to compare multiple models to select the most relevant for the research problem. Additional evaluation involves hyperparameter tuning, resource allocation, benchmarking and model fairness audits.

Overfitting is often a challenge in model training, which can occur when a model is excessively tailored to the peculiarities of the training data and becomes so specialized that its generalizability is compromised. Such a model performs accurately on the specific data set but underperforms on unseen examples. Overfitting can be counteracted by dividing the data into three distinct subsets: the training set, the validation set and the test set. The training set is the primary data set from which the model learns patterns, adjusts its weights and fine-tunes itself based on the labelled examples provided. The validation set is used to monitor and assess the performance of the model during training. It acts as a checkpoint, guides hyperparameter tuning and ensures that the model is not veering off track. The test set is the final held-out set on which the performance of the model is evaluated. The test set is akin to a final examination, assessing how well the model generalizes to unseen data. If a pre-trained model is used, only the data sets used to fine-tune the model are necessary to evaluate the model.

The effectiveness of any trained model is gauged not just by how well it fits the training data but also by its performance on unseen samples. Evaluation metrics provide objective measures to assess performance on validation and test sets as well as unseen examples. The evaluation process is fundamental to QTA experiments, as demonstrated in the text classification research[10]. Several evaluation metrics are used to measure performance. The most prominent are accuracy (the proportion of all predictions that are correct), precision (the proportion of positive predictions that are actually correct) and recall (the proportion of actual positives that were correctly identified). The F1 score amalgamates precision and recall and emerges as a balanced metric, especially when class distributions are skewed. An effective evaluation typically uses various complementary metrics.

Results

In QTA, a before-and-after dynamic often emerges, encapsulating the transformation from raw data to insightful conclusions[54]. This paradigm is especially important in QTA, in which the raw textual data can be used to distil concrete answers to research questions. In the preceding section, the preliminary before phase, the process of setting up an experiment in QTA, is explored with emphasis on the importance of model training and thorough evaluation to ensure robustness. For the after phase, the focus pivots to the critical step of applying the trained model to new, unseen data, aiming to answer the research questions that guide exploration.

Research questions in QTA are often sophisticated and complex, encompassing a range of inquiries either directly related to the text being analysed or to the external phenomena the text reflects. The link between the output of QTA models and the research question is often vague and under-specified. When dealing with a complex research question, for example, the processes that govern the changing attitudes towards different migrant groups, the outcome of any one QTA model is often insufficient. Even several models might not provide a complete answer to the research question. Consequently, challenges surface during the transition from before to after, from setting up and training to applying and validating. One primary obstacle is the validation difficulty posed by the uniqueness and unseen nature of the new data.

Validating QTA models on new, unseen data introduces a layer of complexity that highlights the need for robust validation strategies, to ensure stability, generalizability and replicability of results. Although the effectiveness of a model might have been calibrated in a controlled setup, its performance can oscillate when exposed to the multifaceted layers of new real-world data. Ensuring consistent model performance is crucial to deriving meaningful conclusions aligned with the research question. This dual approach of applying the model and subsequently evaluating its performance in fresh terrains is central to the after phase of QTA. In addition to validating the models, the results that stem from the models need to be validated with respect to the research question. The results need to be representative for the data as a whole; they need to be stable such that the answer does not change if different choices are made in the before phase; and they need to provide an answer to the research question at hand.

This section provides a road map for navigating the application of QTA models to new data and a chart of methodologies for evaluating the outcomes in line with the research question (questions). The goal is to help researchers cross the bridge between the theoretical

Primer

foundations of QTA and its practical implementation, illuminating the steps that support the successful application and assessment of QTA models. The ensuing discussion covers validation strategies that cater to the challenges brought forth by new data, paving the way towards more insightful analysis.

Application to new data

After the training and evaluation phases have been completed, the next step is applying the trained model to new, unseen data (Fig. 2). The goal is to ensure that the application aligns with the research questions and aids in extracting meaningful insights. However, applying the model to new data is not without challenges.

Before application of the model, it is crucial to preprocess the new data similar to the training data. This involves routine tasks such as tokenization and lemmatization, but also demands vigilance for anomalies such as divergent text encoding formats or missing values. In such cases, additional preprocessing steps might be required and should be documented carefully to ensure reproducibility.

Another potential hurdle is the discrepancy in data distributions between the training data and new data, often referred to as domain shift. If not addressed, domain shifts may hinder the efficacy of the model. Even thematically, new data may unearth categories or motifs that were absent during training, thus challenging the interpretative effectiveness of the model. In such scenarios, transfer learning or domain adaptation techniques are invaluable tools for adjusting the model so that it aligns better with the characteristics of the new data. In transfer learning, a pre-trained model provides general language understanding and is fine-tuned with a small data set for a specific task (for example, fine-tuning a large language model such as GPT or BERT for emotion classification)[55,56]. Domain adaptation techniques similarly adjust a model from a source domain to a target domain; for example, an emotion classifier trained on customer reviews can be adapted to rate social media comments.

Given the iterative nature of QTA, applying a model is not necessarily an end point; it may simply be a precursor to additional refinement and analysis. Therefore, the adaptability of the validation strategies is paramount. As nuances in the new data are uncovered, validation strategies may need refinement or re-adaptation to ensure the predictions of the model remain accurate and insightful, ensuring that the answers to the research questions are precise and meaningful. Through careful application and handling of the new data, coupled with adaptable validation strategies, researchers can significantly enhance the value of their analysis in answering the research question.

Evaluation metrics

QTA models are often initially developed and validated on well-defined data sets, ensuring their reliability in controlled settings. This controlled environment allows researchers to set aside a held-out test set to gauge the performance of a model, simulating how it will fare on new data. The real world, however, is considerably more complex than any single data set can capture. The challenge is how to transition from a controlled setting to novel data sets.

One primary challenge is the mismatch between the test set and real-world texts. Even with the most comprehensive test sets, capturing the linguistic variation, topic nuance and contextual subtlety present in new data sets is not a trivial task, and researchers should not be overconfident regarding the universal applicability of a model[57]. The situation does not become less complicated when relying on pre-trained or off-the-shelf models. The original training data and its characteristics might not be transparent or known with such models. Without appropriate documentation, predicting the behaviour of a model on new data may become a speculative endeavour[58].

The following sections summarize strategies for evaluating models on new data.

Model confidence scores. In QTA, models often generate confidence or probability scores alongside predictions, indicating the confidence of the model in its accuracy. However, high scores do not guarantee correctness and can be misleading. Calibrating the model refines these scores to align better with true label likelihoods[59]. This is especially crucial in high-stakes QTA applications such as legal or financial text analysis[60]. Calibration techniques adjust the original probability estimates, enhancing model reliability and the trustworthiness of predictions, thereby addressing potential discrepancies between the expressed confidence of the model and its actual performance.

Precision at _k_. Precision at _k_ (P@_k_) is useful for tasks with rankable predictions, such as determining document relevance. P@_k_ measures the proportion of relevant items among the top-_k_ ranked items, providing a tractable way to gauge the performance of a model on unseen data by focusing on a manageable subset, especially when manual evaluation of the entire data set is infeasible. Although primarily used in information retrieval and recommender system, its principles apply
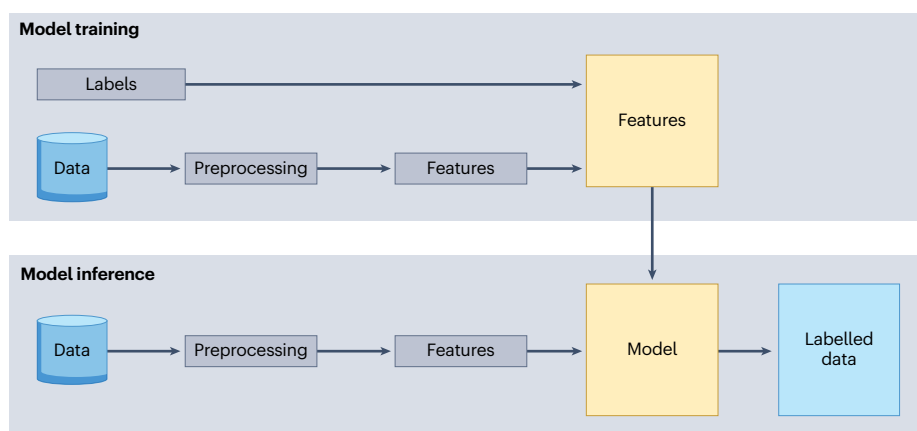

Fig. 2 | Comparative workflows of model training (upper section) and model application or inference (lower section) for a text classification task in the context of the quantitative text analysis. Although the illustration demonstrates a feature-based modelling approach, the fundamental principle remains consistent across different methodologies, be it feature-based, representation learning or generative. A critical consideration is ensuring the consistency in content and preprocessing between the training data and any new data subjected to inference.

Primer

Glossary

Application programming interface
A set of rules, protocols and tools for building software and applications, which programs can query to obtain data.

Bag-of-words model
A model that represents text as a numerical vector based on word frequency or presence. Each text corresponds to a predefined vocabulary dictionary, with the vector.

Computer linguistics
Intersection of linguistics, computer science and artificial intelligence that is concerned with computational aspects of human language. It involves the development of algorithms and models that enable computers to understand, interpret and generate human language.

Corpus linguistics
The branch of linguistics that studies language as expressed in corpora (samples of real-world text) and uses computational methods to analyse large collections of textual data.

Data augmentation
A technique used to increase the size and diversity of language data sets to train machine-learning models.

Data science
The application of statistical, analytical and computational techniques to extract insights and knowledge from data.

Fleiss' kappa
(κ). A statistical measure used to assess the reliability of agreement between multiple raters when assigning categorical ratings to a number of items.

Frequency bias
A phenomenon in which elements that are over-represented in a data set receive disproportionate attention or influence in the analysis.

Information retrieval
A field of study focused on the science of searching for information within documents and retrieving relevant documents from large databases.

Lemmatization
A text normalization technique used in natural language processing in which words are reduced to their base or dictionary form.

Machine learning
In quantitative text analysis, machine learning refers to the application of algorithms and statistical models to enable computers to identify patterns, trends and relationships in textual data without being explicitly programmed. It involves training these models on large data sets to learn and infer from the structure and nuances of language.

Natural language processing
A field of artificial intelligence using computational methods for analysing and generating natural language and speech.

Recommender system
A type of information filtering system that seeks to predict user preferences and recommend items (such as books, movies and products) that are likely to be of interest to the user.

Representation learning
A set of techniques in machine learning in which the system learns to automatically identify and extract useful features or representations from raw data.

Stemming
A text normalization technique used in natural language processing, in which words are reduced to their base or root form.

Supervised learning
A machine-learning approach in which models are trained on labelled data, such that each training text is paired with an output label. The model learns to predict the output from the input data, with the aim of generalizing the training set to unseen data.

Transformer
A deep learning model that handles sequential data, such as text, using mechanisms called attention and self-attention, allowing it to weigh the importance of different parts of the input data. In the quantitative text analysis, transformers are used for tasks such as sentiment analysis, text classification and language translation, offering superior performance in understanding context and nuances in large data sets.

Unsupervised learning
A type of machine learning in which models are trained on data without output labels. The goal is to discover underlying patterns, groupings or structures within the data, often through clustering or dimensionality reduction techniques.

---

to QTA, in which assessing the effectiveness of a model in retrieving or categorizing relevant texts is crucial.

External feedback mechanisms. Soliciting feedback from domain experts is invaluable in evaluating models on unseen data. Domain experts can provide qualitative insights into the output of the model, identifying strengths and potential missteps. For example, in topic modelling, domain experts can assess the coherence and relevance of the generated topics. This iterative feedback helps refine the model, ensuring its robustness and relevance when applied to new, unseen data, thereby bridging the gap between model development and practical application.

Software and tools
When analysing and evaluating QTA models on unseen data, researchers often turn to specialized tools designed to increase model transparency and explain model predictions. Among these tools, LIME (Local Interpretable Model-agnostic Explanations)[61] and SHAP (SHapley Additive exPlanations)[62] have gained traction for their ability to provide insights into model behaviour per instance, which is crucial when transitioning to new data domains.

LIME focuses on the predictions of machine-learning models by creating locally faithful explanations. It operates by perturbing the input data and observing how the predictions change, making it a useful tool to understand model behaviour on unseen data. Using LIME, researchers can approximate complex models with simpler, interpretable models locally around the prediction point. By doing so, they can gain insight into how different input features contribute to the prediction of the model, which can be instrumental in understanding how a model might generalize to new, unseen data.

SHAP, by contrast, provides a unified measure of feature importance across different data types, including text. It uses game theoretic principles to attribute the output of machine-learning models to their input features. This method allows for a more precise understanding of how different words or phrases in text data influence the output of the model, thereby offering a clearer picture of the behaviour of the model

Primer

on new data domains. The SHAP library provides examples of how to explain predictions from text analysis models applied to various NLP tasks including sentiment analysis, text generation and translation.

Both LIME and SHAP offer visual tools to help researchers interpret the predictions of the model, making it easier to identify potential issues when transitioning to unseen data domains. For instance, visualizations allow researchers to identify words or phrases that heavily influence the decisions of the model, which can be invaluable in understanding and adjusting the model for new text data.

Interpretation

Interpretability is paramount in QTA as it facilitates the translation of complex model outcomes into actionable insights relevant to the research questions. The nature and complexity of the research question can significantly mould the interpretation process by requiring various information signals to be extracted from the text, see, for example, ref. 63. For example, in predicting election outcomes based on sentiments expressed in social media[64], it is essential to account for both endorsements of parties as expressed in the text and a count of individuals (that is, statistical signals) to avoid the results being skewed because some individuals make a high number of posts. It is also important to note whether voters of some political parties are under-represented in the data.

The complexity amplifies when delving into understanding why people vote (or do not vote) for particular parties and what arguments sway their decisions. Such research questions demand a more comprehensive analysis, often necessitating the amalgamation of insights from multiple models, for example, argument mining, aspect-based sentiment analysis and topic models. There is a discernible gap between the numerical or categorical outputs of QTA models – such as classification values, proportions of different stances or vectors representing individual words – and the nuanced understanding required to fully address the research question. This understanding is achieved either using qualitative human analysis or applying additional QTA methods and extracts a diverse set of important arguments in support of different stances, or provides qualitative summaries of a large set of different comments. Because it is not only a matter of 'what' results are found using QTA, but the value that can be attributed to those results.

When interpreting the results of a computational model applied to textual data for a specific research question, it is important to consider the completeness of the answer (assess whether the output of the model sufficiently addresses the research question or whether there are aspects left unexplored), the necessity of additional models (determine whether the insights from more models are needed to fully answer the research question), the independence or co-dependence of results (in cases in which multiple models are used, ascertain whether their results are independent or co-dependent and adjust for any overlap in insights accordingly), clarify how the results are used to support an answer (such as the required occurrence of a phenomenon in the text to accept a concept, or how well a derived topic is understood and represented) and the effect of methodology (evaluate the impact of the chosen method or preprocessing on the results, ensuring the reproducibility and robustness of the findings against changes in preprocessing or methods).

Using these considerations alongside techniques such as LIME and SHAP enhances the evaluation of the application of the model. For instance, in a scenario in which a QTA model is used to analyse customer reviews, LIME and SHAP could provide nuanced insights on a peer-review basis and across all reviews, respectively. Such insights are pivotal in assessing the alignment of the model with the domain-relevant information necessary to address the research questions and in making any adjustments needed to enhance its relevance and performance. Moreover, these techniques and considerations catalyse a dialogue between model and domain experts, enabling a more nuanced evaluation that extends beyond mere quantitative metrics towards a qualitative understanding of the application of the model.

Applications

The applicability of QTA can be found in its ability to address research questions across various disciplines. Although these questions are varied and tasks exist that do not fit naturally into categories, they can be grouped into four primary tasks: extracting, categorizing, predicting and generating. Each task is important in advancing understanding of large textual data sets, either by examining phenomena specific to a text or by using texts as a proxy for phenomena outside the text.

Extracting information

In the context of QTA, information extraction goes beyond mere data retrieval; it also involves identifying and assessing patterns, structures and entities within extensive textual data sets. At its core are techniques such as frequency analysis, in which words or sets of words are counted and their occurrences plotted over periods to reveal trends or shifts in usage and syntactical analysis, which targets specific structures such as nouns, verbs and intricate patterns such as passive voice constructions. Named entity recognition pinpoints entities such as persons, organizations and locations using syntactic information and lexicons of entities.

These methodologies have proven useful in various academic domains. For example, humanities scholars have applied QTA to track the evolution of literary themes[65]. Word embedding has been used to shed light on broader sociocultural shifts such as the conceptual change of 'racism', or detecting moments of linguistic change in American foreign relations[40,66]. In a historical context, researchers have used diachronic word embeddings to scrutinize the role of abolitionist newspapers in influencing public opinion about the abolition of slavery, revealing pathways of lexical semantic influence, distinguishing leaders from followers and identifying others who stood out based on the semantic changes that swept through this period[67]. Topic modelling and topic linkage (the extent to which two topics tend to co-appear) have been applied to user comments and submissions from the 'subreddit' group r/TheRedPill to study how people interact with ideology[68]. In the medical domain[69], QTA tools have been used to study narrative structures in personal birth stories. The authors utilized a topic model based on latent Dirichlet allocation (LDA) to not only represent the sequence of events in every story but also detect outlier stories using the probability of transitioning between topics.

Historically, the focus was predominantly on feature-based models that relied on manual feature engineering. Such methods were transparent but rigid, constraining the richness of the textual data. Put differently, given the labour-intensive selection of features and the need to keep them interpretable, the complexity of a text was reduced to a limited set of features. However, the advent of representation learning has catalysed a significant paradigm shift. It enables more nuanced extraction, considers contextual variations and allows for sophisticated trend analysis. Studies using these advanced techniques have been successful in, for example, analysing how gender stereotypes and attitudes towards ethnic minorities in the USA evolved during the twentieth and twenty-first centuries[38] and tracking the emergence of

Primer

Box 1

Using text mining to model prescient ideas

Vicinanza et al.[70] focused on the predictive power of linguistic markers within the domains of politics, law and business, positing that certain shifts in language can serve as early indicators of deeper cognitive changes. They identified two primary attributes of prescient ideas: their capacity to challenge existing contextual assumptions, and their ability to foreshadow the future evolution of a domain. To quantify this, they utilized Bidirectional Encoder Representations from Transformers, a type 2 language model, to calculate a metric termed contextual novelty to gauge the predictability of an utterance within the prevailing discourse.

Their study presents compelling evidence that prescient ideas are more likely to emerge from the periphery of a domain than from its core. This suggests that prescience is not solely an individual trait but also significantly influenced by contextual factors. Thus, the researchers extended the notion of prescience to include the environments in which innovative ideas are nurtured, adding another layer to our understanding of how novel concepts evolve and gain acceptance.

ideas in the domains of politics, law and business through contextual embeddings combined with statistical modelling[70] (Box 1).

Categorizing content

It remains an indispensable task in QTA to categorize content, especially when dealing with large data sets. The challenge is not only logistical but also methodological, demanding sophisticated techniques to ensure precision and utility. Text classification algorithms, supervised or unsupervised, continue to have a central role in labelling and organizing content. They serve crucial functions beyond academic settings; for instance, digital libraries use these algorithms to manage and make accessible their expansive article collections. These classification systems also contribute significantly to the systematic review of the literature, enabling more focused and effective investigations of, for example, medical systematic reviews[71]. In addition, unsupervised techniques such as topic modelling have proven invaluable in uncovering latent subject matter within data sets[72] (Box 2). This utility extends to multiple scenarios, from reducing redundancies in large document sets to facilitating the analysis of open-ended survey responses[73,74].

Earlier approaches to categorization relied heavily on feature-based models that used manually crafted features for organization. This traditional paradigm has been disrupted by advances in representation learning, deep neural networks and word embeddings, which has introduced a new age of dynamic unsupervised and semi-supervised techniques for content categorization. GPT models represent another leap forward in text classification tasks, outpacing existing benchmarks across various applications. From the sentiment analysis to text labelling and psychological construct detection, generative models have demonstrated a superior capability for context understanding, including the ability to parse complex linguistic cues such as sarcasm and mixed emotions[75–77]. Although the validity of these models is a matter of debate, they offer explanations for their reasoning, which adds a layer of interpretability.

Predicting outcomes

QTA is not limited to understanding or classifying text but extends its reach into predictive analytics, which is an invaluable tool across many disciplines and industries. In the financial realm, sentiment analysis tools are applied to news articles and social media data to anticipate stock market fluctuations[78]. Similarly, political analysts use sentiment analysis techniques to make election forecasts, using diverse data sources ranging from Twitter (now X) feeds to party manifestos[79]. Authorship attribution offers another intriguing facet, in which predictive abilities of the QTA are harnessed to identify potential authors of anonymous or pseudonymous works[80]. A notable instance was the unmasking of J.K. Rowling as the author behind the pseudonym Robert Galbraith[81]. Health care has also tapped into predictive strengths of the QTA: machine-learning models that integrate natural language and binary features from patient records have been shown to have potential as early warning systems to prevent unnecessary mechanical restraint of psychiatric inpatients[82] (Box 3).

In the era of feature-based models, predictions often hinged on linear or tree-based structures using manually engineered features. Representation learning introduced embeddings and sequential models that improved prediction capabilities. These learned representations enrich predictive tasks, enhancing accuracy and reliability while decreasing interpretability.

Generating content

Although the initial QTA methodologies were not centred on content generation, the rise of generative models has been transformative. Models such as GPT-4 and Llama2 (ref. 20) have brought forth previously unimagined capabilities, expanding the potential of QTA to create content, including coherent and contextually accurate paragraphs to complete articles. Writers and content creators are now using tools based on models such as GPT-4 to augment their writing processes by offering suggestions or even drafting entire sections of texts. In education, such models aid in developing customized content for students,

Box 2

Exploring molecular data with topic modelling

Schneider et al.[72] introduced a novel application of topic modelling to the field of medicinal chemistry. The authors adopt a probabilistic topic modelling approach to organize large molecular data sets into chemical topics, enabling the investigation of relationships between these topics. They demonstrate the effectiveness of the quantitative text analysis method in identifying and retrieving chemical series from molecular sets. The authors are able to reproduce concepts assigned by humans in the identification and retrieval of chemical series from sets of molecules. Using topic modelling, the authors are able to show chemical topics intuitively with data visualization and efficiently extend the method to a large data set (ChEMBL22) containing 1.6 million molecules.

Primer

ensuring adaptive learning[83]. The capacity to create synthetic data also heralds new possibilities. Consider the domain of historical research, in which generative models can simulate textual content, offering speculative yet data-driven accounts of alternate histories or events that might have been; for example, relying on generative models to create computational software agents that simulate human behaviour[84]. However, the risks associated with text-generating models are exemplified by a study in which GPT-3 was used for storytelling. The generated stories were found to exhibit many known gender stereotypes, even when prompts did not contain explicit gender cues or stereotype-related content[85].

Reproducibility and data deposition

Given the rapidly evolving nature of the models, methods and practices in QTA, reproducibility is essential for validating the results and creating a foundation upon which other researchers can build. Sharing code and trained models in well-documented repositories are important to enable reproducible experiments. However, sharing and depositing raw data can be challenging, owing to the inherent limitations of unstructured data and regulations related to proprietary and sensitive data.

Code and model sharing

In QTA research, using open source code has become the norm and the need to share models and code to foster innovation and collaboration has been widely accepted. QTA is interdisciplinary by nature, and by making code and models public, the field has avoided unnecessary silos and enabled collaboration between otherwise disparate disciplines. A further benefit of open source software is the flexibility and transparency that comes from freely accessing and modifying software to meet specific research needs. Accessibility enables an iterative feedback loop, as researchers can validate, critique and build on the existing work. Software libraries, such as scikit-learn, that have been drivers for adopting machine learning in QTA are testimony to the importance of open source software[21].

Sharing models is not without challenges. QTA is evolving rapidly, and models may use specific versions of software and hardware configurations that no longer work or that yield different results with other versions or configurations. This variability can complicate the accessibility and reproducibility of research results. The breakthroughs of generative AI in particular have introduced new proprietary challenges to model sharing as data owners and sources raise objections to the use of models that have been trained on their data. This challenge is complicated, but fundamentally it mirrors the disputes about intellectual property rights and proprietary code in software engineering. Although QTA as a field benefits from open source software, individual research institutions may have commercial interests or intellectual property rights related to their software.

On the software side, there is currently a preference for scripting languages, especially Python, that enable rapid development, provide access to a wide selection of software libraries and have a large user community. QTA is converging towards code and model sharing through open source platforms such as GitHub and GitLab with an appropriate open source software license such as the MIT license. Models often come with additional disclaimers or use-based restrictions to promote responsible use of AI, such as in the RAIL licenses. Pre-trained models are also regularly shared on dedicated machine-learning platforms such as Hugging Face[22] to enable efficient fine-tuning and deployment. It is important to emphasize that although these platforms support open science, these services are provided by companies with commercial interests. Open science platforms such as Zenodo and OSF can also be used to share code and models for the purpose of reproducibility.

Popular containerization software has been widely adopted in the machine-learning community and has spread to QTA. Containerization, that is, packaging all parts of a QTA application — including code and other dependencies — into a single standalone unit ensures that model and code run consistently across various computing environments. It offers a powerful solution to challenges such as reproducibility, specifically variability in software and hardware configurations.

Data management and storage

Advances in QTA in recent years are mainly because of the availability of vast amounts of text data and the rise of deep learning techniques. However, the dependency on large unstructured data sets, many of which are proprietary or sensitive, poses unique data management challenges. Pre-trained models irrespective of their use (for example, representation learning or generative) require extensive training on large data sets. When these data sets are proprietary or sensitive, they cannot be readily available, which limits the ability of researchers to reproduce results and develop competitive models. Furthermore, models trained on proprietary data sets often lack transparency regarding their collection and curation processes, which can hide potential biases in the data. Finally, there can be data privacy issues related to training or using models that are trained on sensitive data. Individuals whose data are included may not have given their explicit consent for their information to be used in research, which can pose ethical and legal challenges.

Box 3

Predicting mechanical restraint: assessing the contribution of textual data

Danielsen et al.[82] set out to assess the potential of electronic health text data to predict incidents of mechanical restraint of psychiatric patients. Mechanical restraint is used during inpatient treatments to avert potential self-harm or harm to others. The research team used feature-based supervised machine learning to train a predictive model on clinical notes and health records from the Central Denmark Region, specifically focusing on the first hour of admission data. Of 5,050 patients and 8,869 admissions, 100 patients were subjected to mechanical restraint between 1h and 3 days after admission. Impressively, a random forest algorithm could predict mechanical restraint with considerable precision, showing an area under the curve of 0.87. Nine of the ten most influential predictors stemmed directly from clinical notes, that is, unstructured textual data. The results show the potential of textual data for the creation of an early detection system that could pave the way for interventions that minimize the use of mechanical restraint. It is important to emphasize that the model was limited by a narrow scope of data from the Central Denmark Region, and by the fact that only initial mechanical restraint episodes were considered (in other words, recurrent incidents were not included in the study).

It is a widely adopted practice in QTA to share data and metadata with an appropriate license whenever possible. Data can be deposited in open science platforms such as Zenodo, but specialized machine-learning platforms are also used for this purpose. However, it should be noted that QTA data are rarely unique, unlike experimental data collected through random controlled trials. In many cases, access to appropriate metadata and documentation would enable the data to be reconstructed. In almost all cases, it is therefore strongly recommended that researchers share metadata and documentation for data, as well as code and models, using a standardized document or framework, a so-called datasheet. Although QTA is not committed to one set of principles for (meta)data management, European research institutions are increasingly adopting the FAIR principles[50].

Documentation

Although good documentation is vital in all fields of software development and research, the reliance of QTA on code, models and large data sets makes documentation particularly crucial for reproducibility. Popular resources for structuring projects include project templating tools and documentation generators such as Cookiecutter and Sphinx. Models are often documented with model cards that provide a detailed overview of the development, capabilities and biases of the model to promote transparency and accountability[86]. Similarly, datasheets or data cards can be used to promote transparency for data used in QTA[87]. Finally, it is considered good practice to provide logs for models that document parameters, metrics and events for QTA experiments, especially during training and fine-tuning. Although not strictly required, logs are also important for documenting the iterative process of model refinement. There are several platforms that support the creation and visualization of training logs (Weights & Biases and MLflow).

Limitations and optimizations

The application of QTA requires scrutiny of its inherent limitations and potentials. This section discusses these aspects and elucidates the challenges and opportunities for further refinement.

Limitations in QTA

Defining research questions. In QTA, the framing of research questions is often determined by the capabilities and limitations of the available text analysis tools, rather than by intellectual inquiry or scientific curiosity. This leads to task-driven limitations, in which inquiry is confined to areas where the tools are most effective. For example, relying solely on bag-of-words models might skew research towards easily quantifiable aspects, distorting the intellectual landscape. Operationalizing broad and nuanced research questions into specific tasks may strip them of their depth, forcing them to conform to the constraints of existing analytical models[88].

Challenges in interpretation. The representation of language of underlying phenomena is often ambiguous or indirect, requiring careful interpretation. Misinterpretations can arise, leading to challenges related to historical, social and cultural context of a text, in which nuanced meanings that change across time, class and cultures are misunderstood[89]. Overlooking other modalities such as visual or auditory information can lead to a partial understanding of the subject matter and limit the full scope of insights. This can to some extent be remedied by the use of grounded models (such as GPT-4), but it remains a challenge for the community to solve long term.

Determining reliability and validation. The reliability and stability of the conclusions drawn from the QTA require rigorous validation, which is often neglected in practice. Multiple models, possibly on different types of data, should be compared to ensure that conclusions are not artefacts of a particular method or of a different use of the method. Furthermore, cultural phenomena should be evolved to avoid misguided insights. Building a robust framework that allows testing and comparison enhances the integrity and applicability of QTA in various contexts[90].

Connecting analysis to cultural insights. Connecting text analysis to larger cultural claims necessitates foundational theoretical frameworks, including recognizing linguistic patterns, sociolinguistic variables and theories of cultural evolution that may explain changes. Translating textual patterns into meaningful cultural observations requires understanding how much (or how little) culture is expressed in text so that findings can be generalized beyond isolated observations. A theoretical foundation is vital to translate textual patterns into culturally relevant insights, making QTA a more effective tool for broader cultural analysis.

Balancing factors in machine learning

Balancing factors is critical in aligning machine-learning techniques with research objectives. This includes the trade-off between quality and control. Quality refers to rigorous, robust and valid findings, and control refers to the ability to manage specific variables for clear insights. It is also vital to ensure a balance between quantity and quality in data source to lead to more reliable conclusions. Balance is also needed between correctness and accuracy, in which the former ensures consistent application of rules, and the latter captures the true nature of the text.

From features-based to generative models

QTA has undergone a profound evolution, transitioning from feature-based approaches to representation learning and finally to generative models. This progression demonstrates growing complexity in our understanding of language, reflecting the maturity in the field of QTA. Each stage has its characteristics, strengths and limitations.

In the early stages, feature-based models were both promising and limiting. The simplicity of their design, relying on explicit feature engineering, allowed for the targeted analysis. However, this simplicity limited their ability to grasp complex, high-level patterns in language. For example, the use of bag-of-words models in the sentiment analysis showcased direct applicability, but also revealed limitations in understanding contextual nuances. The task-driven limitations of these models sometimes overshadowed genuine intellectual inquiry. Using a fixed (often modern) list of words with corresponding emotional valences may limit our ability to fully comprehend the complexity of emotional stances in, for example, historical literature. Despite these drawbacks, the ability to customize features provided researchers with a direct and specific understanding of language phenomena that could be informed by specialized domain knowledge[91].

With the emergence of representation learning, a shift occurred within the field of QTA. These models offered the ability to capture higher-level abstractions, forging a richer understanding of language. Their scalability to handle large data sets and uncover complex relationships became a significant strength. However, this complexity introduced new challenges, such as a loss of specificity in analysis and difficulties in translating broad research questions into specific tasks.

Primer

Techniques such as Word2Vec enabled the capture of semantic relationships but made it difficult to pinpoint specific linguistic features. Contextualized models, in turn, allow for more specificity, but are typically pre-trained on huge data sets (not available for scrutiny) and then applied to a research question without any discussion of how well the model fits the data at hand. In addition, these contextualized models inundate with information. Instead of providing one representation for a word (similar to Word2Vec does), they provide one representation for each occurrence of the word. Each of these representations is one order of magnitude larger than vectors typical for Word2Vec (768–1,600 dimensions compared with 50–200) and comes in several varieties, one for each of the layers of the mode, typically 12.

The introduction of generative models represents the latest stage of this evolution, providing even greater complexity and potential. Innovative in their design, generative models provide opportunities to address more complex and open-ended research questions. They fuel the generation of new ideas and offer avenues for novel approaches. However, these models are not without their challenges. Their high complexity can make interpretation and validation demanding, and if not properly managed, biases and ethical dilemmas will emerge. The use of generative models in creating synthetic text must be handled with care to avoid reinforcing stereotypes or generating misleading information. In addition, if the enormous amounts of synthetically generated text are used to further train the models, this will lead to a spiral of decaying quality as eventually a majority of the training data will have been generated by machines (the models often fail to distinguish synthetic text from genuine human-created text)[92]. However, it will also allow researchers to draw insights from a machine that is learning on data it has generated itself.

The evolution from feature-based to representation learning to generative models reflects increasing maturity in the field of QTA. As models become more complex, the need for careful consideration, ethical oversight and methodological innovation intensifies. The challenge now lies in ensuring that these methodologies align with intellectual and scientific goals, rather than being constrained by their inherent limitations. This growing complexity mirrors the increasing demands of this information-driven society, requiring interdisciplinary collaboration and responsible innovation. Generative models require a nuanced understanding of the complex interplay between language, culture, time and society, and a clear recognition of constraints of the QTA. Researchers must align their tools with intellectual goals and embrace active efforts to address the challenges through optimization strategies. The evolution in QTA emphasizes not only technological advances but also the necessity of aligning the ever-changing landscape of computational methodologies with research questions. By focusing on these areas and embracing the accompanying challenges, the field can build robust, reliable conclusions and move towards more nuanced applications of the text analysis. This progress marks a significant step towards an enriched exploration of textual data, widening the scope for understanding multifaceted relationships. The road ahead calls for a further integration of theory and practice. It is essential that evolution of QTA ensures that technological advancement serves both intellectual curiosity and ethical responsibility, resonating with the multifaceted dynamics of language, culture, time and society[93].

Outlook

Balancing size and quality

In QTA, the relationship between data quantity and data quality is often misconceived. Although large data sets serve as the basis for training expansive language models, they are not always required when seeking answers to nuanced research questions. The wide-ranging scope of large data sets can offer comprehensive insights into broad trends and general phenomena. However, this often comes at the cost of a detailed understanding of context-specific occurrences. An issue such as frequency bias exemplifies this drawback. Using diverse sampling strategies, such as stratified sampling to ensure representation across different social groups and bootstrapping methods to correct for selection bias, can offer a more balanced, contextualized viewpoint. Also, relying on methods such as burst or change-point detection can help to pinpoint moments of interest in data sets with a temporal dimension. Triangulating these methods across multiple smaller data sets can enhance reliability and depth of the analysis.

The design of machine-learning models should account for both the frequency and the significance of individual data points. In other words, the models should be capable of learning not just from repetitive occurrences but also from singular, yet critical, events. This enables the machine to understand rare but important phenomena such as revolutions, seminal publications or watershed individual actions, which would typically be overlooked in a conventional data-driven approach. The capacity to learn from such anomalies can enhance the interpretative depth of the model, enabling them to offer more nuanced insights.

Although textual data have been the mainstay for computational analyses, it is not the only type of data that matters, especially when the research questions involve cultural and societal nuances. Diverse data types including images, audio recordings and even physical artefacts should be integrated into the research to provide a more rounded analysis. Additionally, sourcing data from varied geographical and sociocultural contexts can bring multiple perspectives into the frame, thus offering a multifaceted understanding that textual data from English sources alone cannot capture.

Ethical, practical and efficient models

The evolving landscape of machine learning, specifically with respect to model design and utility, reflects a growing emphasis on efficiency and interpretive value. One notable shift is towards smaller, more energy-efficient models. This transition is motivated by both environmental sustainability and economic pragmatism. With computational costs soaring and the environmental toll becoming untenable, the demand for smaller models that maintain or even exceed the quality of larger models is escalating[94].

Addressing the data sources used to train models is equally critical, particularly when considering models that will serve research or policy purposes. The provenance and context of data dictate its interpretive value, requiring models to be designed with a hierarchical evaluation of data sources. Such an approach could improve the understanding of a model of the importance of each data type given a specific context, thereby improving the quality and reliability of its analysis. Additionally, it is important to acknowledge the potential ethical and legal challenges within this process, including the exploitation of workers during the data collection and model development.

Transparency remains another pressing issue as these models become integral to research processes. Future iterations should feature a declaration of content that enumerates not only the origin of the data but also its sociocultural and temporal context, preprocessing steps, any known biases, along with the analytical limitations of the model. This becomes especially important for generative models, which may produce misleading or even harmful content if the original

data sources are not properly disclosed and understood. Important steps have already been taken with the construction of model cards and data sheets[95].

Finally, an emergent concern is the risk of feedback loops compromising the quality of machine-learning models. If a model is trained on its own output, errors and biases risk being amplified over time. This necessitates constant vigilance as it poses a threat to the long-term reliability and integrity of AI models. The creation of a gold-standard version of the Internet, not polluted by AI-generated data, is also important[96].

Refining the methodology and ethos

The rapid advances in QTA, particularly the rise of generative models, have opened up a discourse that transcends mere technological prowess. Although earlier feature-based models require domain expertise and extensive human input before they could be used, generative models can already generate convincing output based on relatively short prompts. This shift raises crucial questions about the interplay between machine capability and human expertise. The notion that advanced algorithms might eventually replace researchers is a common misplaced apprehension. These algorithms and models should be conceived as tools to enhance human scholarship by automating mundane tasks, spawning new research questions and even offering novel pathways for data analysis that might be too complex or time-consuming for human cognition.

This paradigm shift towards augmentative technologies introduces a nuanced problem-solving framework that accommodates the complexities intrinsic to studying human culture and behaviour. The approach of problem decomposition, a cornerstone in computer science, also proves invaluable here, converting overarching research queries into discrete, operationalizable components. These elements can then be addressed individually through specialized algorithms or models, whose results can subsequently be synthesized into a comprehensive answer. As we integrate increasingly advanced tuning methods into generative models — such as prompt engineering, retrieval augmented generation and parameter-efficient fine-tuning — it is important to remember that these models are tools, not replacements. They are most effective when employed as part of a broader research toolkit, in which their strengths can complement traditional scholarly methods.

Consequently, model selection becomes pivotal and should be intricately aligned with the nature of the research inquiry. Unsupervised learning algorithms such as clustering are well suited to exploratory research aimed at pattern identification. Conversely, confirmatory questions, which seek to validate theories or test hypotheses, are better addressed through supervised learning models such as regression.

The importance of a well-crafted interpretation stage cannot be overstated. This is where the separate analytical threads are woven into a comprehensive narrative that explains how the individual findings conjoin to form a cohesive answer to the original research query. However, the lack of standardization across methodologies is a persistent challenge. This absence hinders the reliable comparison of research outcomes across various studies. To remedy this, a shift towards establishing guidelines or best practices is advocated. These need not be rigid frameworks but could be adapted to fit specific research contexts, thereby ensuring methodological rigor alongside innovative freedom.

Reflecting on the capabilities and limitations of current generative models in QTA research is crucial. Beyond recognizing their utility, the blind spots — questions they cannot answer and challenges they have yet to overcome — need to be addressed[97,98]. There is a growing need to tailor these models to account for nuances such as frequency bias and to include various perspectives, possibly through more diverse data sets or a polyvocal approach.

In summary, a multipronged approach that synergizes transparent and informed data selection, ethical and critical perspectives on model building and selection, and an explicit and reproducible result interpretation offers a robust framework for tackling intricate research questions. By adopting such a nuanced strategy, we make strides not just in technological capability but also in the rigor, validity and credibility of QTA as a research tool.

References

1. Miner, G. *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications* (Academic Press, 2012).
2. Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. From data mining to knowledge discovery in databases. *AI Mag.* **17**, 37 (1996).
3. Hand, D. J. Data mining: statistics and more? *Am. Stat.* **52**, 112–116 (1998).
4. McEnery, T. & Wilson, A. *Corpus Linguistics: An Introduction* (Edinburgh University Press, 2001).
5. Manning, C. D. & Schütze, H. *Foundations of Statistical Natural Language Processing* 1st edn (The MIT Press, 1999).
6. Manning, C., Raghavan, P. & Schütze, H. *Introduction to Information Retrieval* 1st edn (Cambridge University Press, 2008).
7. Wankhade, M., Rao, A. C. S. & Kulkarni, C. A survey on sentiment analysis methods, applications, and challenges. *Artif. Intell. Rev.* **55**, 5731–5780 (2022).
8. Jehangir, B., Radhakrishnan, S. & Agarwal, R. A survey on named entity recognition — datasets, tools, and methodologies. *Nat. Lang. Process. J.* **3**, 100017 (2023).
9. Fu, S. et al. Clinical concept extraction: a methodology review. *J. Biomed. Inform.* **109**, 103526 (2020).
10. Sebastiani, F. Machine learning in automated text categorization. *ACM Comput. Surv.* **34**, 1–47 (2002).
11. Talley, E. M. et al. Database of NIH grants using machine-learned categories and graphical clustering. *Nat. Meth.* **8**, 443–444 (2011).
12. Bommasani, R. et al. On the opportunities and risks of foundation models. Preprint at https://arxiv.org/abs/2108.07258 (2022).
13. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003).
14. Angelov, D. Top2Vec: distributed representations of topics. Preprint at https://arxiv.org/abs/2008.09470 (2020).
15. Barron, A. T. J., Huang, J., Spang, R. L. & DeDeo, S. Individuals, institutions, and innovation in the debates of the French Revolution. *Proc. Natl Acad. Sci. USA* **115**, 4607–4612 (2018).
16. Mitchell, T. M. *Machine Learning* 1st edn (McGraw-Hill, 1997).
17. Vaswani, A. et al. Attention is all you need. in *Advances in Neural Information Processing Systems* (eds Guyon, I. et al.) Vol. 30 (Curran Associates, Inc., 2017).
18. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805 (2018).
19. Brown, T. et al. Language models are few-shot learners. in *Advances in Neural Information Processing Systems* Vol. 33 (eds Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. & Lin, H.) 1877–1901 (Curran Associates, Inc., 2020).
20. Touvron, H. et al. Llama 2: open foundation and fine-tuned chat models. Preprint at https://arxiv.org/abs/2307.09288 (2023).
21. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
22. Wolf, T. et al. Transformers: state-of-the-art natural language processing. in *Proc. 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* 38–45 (Association for Computational Linguistics, Online, 2020).
23. Demartini, G., Siersdorfer, S., Chelaru, S. & Nejdl, W. Analyzing political trends in the blogosphere. in *Proceedings of the International AAAI Conference on Web and Social Media* vol. 5 466–469 (AAAI, 2011).
24. Goldstone, A. & Underwood, T. The quiet transformations of literary studies: what thirteen thousand scholars could tell us. *New Lit. Hist.* **45**, 359–384 (2014).
25. Tangherlini, T. R. & Leonard, P. Trawling in the sea of the great unread: sub-corpus topic modeling and humanities research. *Poetics* **41**, 725–749 (2013).
26. Mei, Q. & Zhai, C. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. in *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 198–207 (Association for Computing Machinery, 2005).
27. Frermann, L. & Lapata, M. A Bayesian model of diachronic meaning change. *Trans. Assoc. Comput. Linguist.* **4**, 31–45 (2016).

Primer

28. Koplenig, A. *Analyzing Lexical Change in Diachronic Corpora*. PhD thesis, Mannheim https://nbn-resolving.org/urn:nbn:de:bsz:mh39-48905 (2016).
29. Dubossarsky, H., Weinshall, D. & Grossman, E. Outta control: laws of semantic change and inherent biases in word representation models. in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* 1136–1145 (Association for Computational Linguistics, 2017).
30. Dubossarsky, H., Hengchen, S., Tahmasebi, N. & Schlechtweg, D. Time-out: temporal referencing for robust modeling of lexical semantic change. in *Proc. 57th Annual Meeting of the Association for Computational Linguistics* 457–470 (Association for Computational Linguistics, 2019).
31. Koplenig, A. Why the quantitative analysis of diachronic corpora that does not consider the temporal aspect of time-series can lead to wrong conclusions. *Digit. Scholarsh. Humanit.* **32**, 159–168 (2017).
32. Tahmasebi, N., Borin, L. & Jatowt, A. Survey of computational approaches to lexical semantic change detection. *Zenodo* https://doi.org/10.5281/zenodo.5040302 (2021).
33. Bizzoni, Y., Degaetano-Ortttlieb, S., Fankhauser, P. & Teich, E. Linguistic variation and change in 250 years of English scientific writing: a data-driven approach. *Front. Artif. Intell.* **3**, 73 (2020).
34. Haider, T. & Eger, S. Semantic change and emerging tropes in a large corpus of New High German poetry. in *Proc. 1st International Workshop on Computational Approaches to Historical Language Change* 216–222 (Association for Computational Linguistics, 2019).
35. Vylomova, E., Murphy, S. & Haslam, N. Evaluation of semantic change of harm-related concepts in psychology. in *Proc. 1st International Workshop on Computational Approaches to Historical Language Change* 29–34 (Association for Computational Linguistics, 2019).
36. Marjanen, J., Pivovarova, L., Zosa, E. & Kurunmäki, J. Clustering ideological terms in historical newspaper data with diachronic word embeddings. in *5th International Workshop on Computational History, HistoInformatics 2019* (CEUR-WS, 2019).
37. Tripodi, R., Warglien, M., Levis Sullam, S. & Paci, D. Tracing antisemitic language through diachronic embedding projections: France 1789–1914. in *Proc. 1st International Workshop on Computational Approaches to Historical Language Change* 115–125 (Association for Computational Linguistics, 2019).
38. Garg, N., Schiebinger, L., Jurafsky, D. & Zou, J. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc. Natl. Acad. Sci. USA* **115**, E3635–E3644 (2018).
39. Wevers, M. Using word embeddings to examine gender bias in Dutch newspapers, 1950–1990. in *Proc. 1st International Workshop on Computational Approaches to Historical Language Change* 92–97 (Association for Computational Linguistics, 2019).
40. Sommerauer, P. & Fokkens, A. Conceptual change and distributional semantic models: an exploratory study on pitfalls and possibilities. in *Proc. 1st International Workshop on Computational Approaches to Historical Language Change* 223–233 (Association for Computational Linguistics, 2019).
    This article examines the effects of known pitfalls on digital humanities studies, using embedding models, and proposes guidelines for conducting such studies while acknowledging the need for further research to differentiate between artefacts and actual conceptual changes.
41. Doermann, D. & Tombre, K. (eds) *Handbook of Document Image Processing and Recognition* 2014th edn (Springer, 2014).
42. Yu, D. & Deng, L. *Automatic Speech Recognition: A Deep Learning Approach* 2015th edn (Springer, 2014).
43. Dasu, T. & Johnson, T. *Exploratory Data Mining and Data Cleaning* (John Wiley & Sons, Inc., 2003).
44. Prabhavalkar, R., Hori, T., Sainath, T. N., Schlüter, R. & Watanabe, S. End-to-end speech recognition: a survey https://arxiv.org/abs/2303.03329 (2023).
45. Pustejovsky, J. & Stubbs, A. *Natural Language Annotation for Machine Learning: A Guide to Corpus-Building for Applications* 1st edn (O'Reilly Media, 2012).
    A hands-on guide to data-intensive humanities research, including the quantitative text analysis, using the Python programming language.
46. Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174 (1977).
47. Gurav, V., Parkar, M. & Kharwar, P. Accessible and ethical data annotation with the application of gamification. in *Data Science and Analytics* (eds Batra, U., Roy, N. R. & Panda, B.) 68–78 (Springer Singapore, 2020).
48. Paolacci, G., Chandler, J. & Ipeirotis, P. G. Running experiments on Amazon Mechanical Turk. *Judgm. Decis. Mak.* **5**, 411–419 (2010).
49. Bergvall-Kåreborn, B. & Howcroft, D. Amazon mechanical turk and the commodification of labour. *New Technol. Work Employ.* **29**, 213–223 (2014).
50. Wilkinson, M. D. et al. The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
51. Klymenko, O., Meisenbacher, S. & Matthes, F. Differential privacy in natural language processing the story so far. in *Proc. Fourth Workshop on Privacy in Natural Language Processing* 1–11 (Association for Computational Linguistics, 2022).
52. Lassen, I. M. S., Almasi, M., Enevoldsen, K. & Kristensen-McLachlan, R. D. Detecting intersectionality in NER models: a data-driven approach. in *Proc. 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* 116–127 (Association for Computational Linguistics, 2023).
53. *DaCy: A Unified Framework for Danish NLP* Vol. 2989, 206–216 (CEUR Workshop Proceedings, 2021).
54. Karsdorp, F., Kestemont, M. & Riddell, A. *Humanities Data Analysis: Case Studies with Python* (Princeton Univ. Press, 2021).
55. Ruder, S., Peters, M. E., Swayamdipta, S. & Wolf, T. Transfer learning in natural language processing. in *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials* 15–18 (Association for Computational Linguistics, 2019).
    The paper presents an overview of modern transfer learning methods in natural language processing, highlighting their emergence, effectiveness in improving the state of the art across various tasks and potential to become a standard tool in natural language processing.
56. Malte, A. & Ratadiya, P. Evolution of transfer learning in natural language processing. Preprint at https://arxiv.org/abs/1910.07370 (2019).
57. Groh, M. Identifying the context shift between test benchmarks and production data. Preprint at https://arxiv.org/abs/2207.01059 (2022).
58. Wang, H., Li, J., Wu, H., Hovy, E. & Sun, Y. Pre-trained language models and their applications. *Engineering* **25**, 51–65 (2023).
    This article provides a comprehensive review of the recent progress and research on pre-trained language models in natural language processing, including their development, impact, challenges and future directions in the field.
59. Wilks, D. S. On the combination of forecast probabilities for consecutive precipitation periods. *Weather Forecast.* **5**, 640–650 (1990).
60. Loughran, T. & McDonald, B. Textual analysis in accounting and finance: a survey. *J. Account. Res.* **54**, 1187–1230 (2016).
61. Ribeiro, M. T., Singh, S. & Guestrin, C. 'Why should I trust you?': explaining the predictions of any classifier. Preprint at https://arxiv.org/abs/1602.04938 (2016).
62. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. in *Advances in Neural Information Processing Systems* Vol. 30 (eds Guyon, I. et al.) 4765–4774 (Curran Associates, Inc., 2017).
63. Tahmasebi, N. & Hengchen, S. The strengths and pitfalls of large-scale text mining for literary studies. *Samlaren* **140**, 198–227 (2019).
64. Jaidka, K., Ahmed, S., Skoric, M. & Hilbert, M. Predicting elections from social media: a three-country, three-method comparative study. *Asian J. Commun.* **29**, 252–273 (2019).
65. Underwood, T. *Distant Horizons: Digital Evidence and Literary Change* (Univ. Chicago Press, 2019).
66. Jo, E. S. & Algee-Hewitt, M. The long arc of history: neural network approaches to diachronic linguistic change. *J. Jpn Assoc. Digit. Humanit.* **3**, 1–32 (2018).
67. Soni, S., Klein, L. F. & Eisenstein, J. Abolitionist networks: modeling language change in nineteenth-century activist newspapers. *J. Cultural Anal.* **6**, 1–43 (2021).
68. Perry, C. & Dedeo, S. The cognitive science of extremist ideologies online. Preprint at https://arxiv.org/abs/2110.00626 (2021).
69. Antoniak, M., Mimno, D. & Levy, K. Narrative paths and negotiation of power in birth stories. *Proc. ACM Hum. Comput. Interact.* **3**, 1–27 (2019).
70. Vicinanza, P., Goldberg, A. & Srivastava, S. B. A deep-learning model of prescient ideas demonstrates that they emerge from the periphery. *PNAS Nexus* **2**, pgac275 (2023).
    Using deep learning on text data, the study identifies markers of prescient ideas, revealing that groundbreaking thoughts often emerge from the periphery of domains rather than their core.
71. Adeva, J. G., Atxa, J. P., Carrillo, M. U. & Zengotitabengoa, E. A. Automatic text classification to support systematic reviews in medicine. *Exp. Syst. Appl.* **41**, 1498–1508 (2014).
72. Schneider, N., Fechner, N., Landrum, G. A. & Stiefl, N. Chemical topic modeling: exploring molecular data sets using a common text-mining approach. *J. Chem. Inf. Model.* **57**, 1816–1831 (2017).
73. Kayi, E. S., Yadav, K. & Choi, H.-A. Topic modeling based classification of clinical reports. in *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop* 67–73 (Association for Computational Linguistics, 2013).
74. Roberts, M. E. et al. Structural topic models for open-ended survey responses. *Am. J. Political Sci.* **58**, 1064–1082 (2014).
75. Kheiri, K. & Karimi, H. SentimentGPT: exploiting GPT for advanced sentiment analysis and its departure from current machine learning. Preprint at https://arxiv.org/abs/2307.10234 (2023).
76. Pelaez, S., Verma, G., Ribeiro, B. & Shapira, P. Large-scale text analysis using generative language models: a case study in discovering public value expressions in AI patents. Preprint at https://arxiv.org/abs/2305.10383 (2023).
77. Rathje, S. et al. GPT is an effective tool for multilingual psychological text analysis. Preprint at https://psyarxiv.com/sekf5/ (2023).
78. Bollen, J., Mao, H. & Zeng, X. Twitter mood predicts the stock market. *J. Comput. Sci.* **2**, 1–8 (2011).
    Analysing large-scale Twitter feeds, the study finds that certain collective mood states can predict daily changes in the Dow Jones Industrial Average with 86.7% accuracy.
79. Tumasjan, A., Sprenger, T. O., Sandner, P. G. & Welpe, I. M. Election forecasts with twitter: how 140 characters reflect the political landscape. *Soc. Sci. Comput. Rev.* **29**, 402–418 (2011).
80. Koppel, M., Schler, J. & Argamon, S. Computational methods in authorship attribution. *J. Am. Soc. Inf. Sci. Tech.* **60**, 9–26 (2009).
81. Juola, P. The Rowling case: a proposed standard analytic protocol for authorship questions. *Digit. Scholarsh. Humanit.* **30**, i100–i113 (2015).

Primer

82. Danielsen, A. A., Fenger, M. H. J., Østergaard, S. D., Nielbo, K. L. & Mors, O. Predicting mechanical restraint of psychiatric inpatients by applying machine learning on electronic health data. *Acta Psychiatr. Scand.* **140**, 147–157 (2019).
The study used machine learning from electronic health data to predict mechanical restraint incidents within 3 days of psychiatric patient admission, achieving an accuracy of 0.87 area under the curve, with most predictive factors coming from clinical text notes.

83. Rudolph, J., Tan, S. & Tan, S. ChatGPT: bullshit spewer or the end of traditional assessments in higher education? *J. Appl. Learn. Teach.* **6**, 342–363 (2023).

84. Park, J. S. et al. Generative agents: interactive Simulacra of human behavior. in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)* 1–22 (Association for Computing Machinery, 2023).

85. Lucy, L. & Bamman, D. Gender and representation bias in GPT-3 generated stories. in *Proc. Third Workshop on Narrative Understanding* 48–55 (Association for Computational Linguistics, Virtual, 2021).
The paper shows how GPT-3-generated stories exhibit gender stereotypes, associating feminine characters with family and appearance, and showing them as less powerful than masculine characters, prompting concerns about social biases in language models for storytelling.

86. Mitchell, M. et al. Model cards for model reporting. in *Proc. Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2019).
The paper introduces model cards for documentation of machine-learning models, detailing their performance characteristics across diverse conditions and contexts to promote transparency and responsible use.

87. Gebru, T. et al. Datasheets for datasets. *Commun. ACM* **64**, 86–92 (2021).

88. Bailer-Jones, D. M. When scientific models represent. *Int. Stud. Philos. Sci.* **17**, 59–74 (2010).

89. Guldi, J. *The Dangerous Art of Text Mining: A Methodology for Digital History* 1st edn (Cambridge Univ. Press, (2023).

90. Da, N. Z. The computational case against computational literary studies. *Crit. Inquiry* **45**, 601–639 (2019).

91. Mäntylä, M. V., Graziotin, D. & Kuutila, M. The evolution of sentiment analysis — a review of research topics, venues, and top cited papers. *Comp. Sci. Rev.* **27**, 16–32 (2018).

92. Alemohammad, S. et al. Self-consuming generative models go mad. Preprint at https://arxiv.org/abs/2307.01850 (2023).

93. Bockting, C. L., van Dis, E. A., van Rooij, R., Zuidema, W. & Bollen, J. Living guidelines for generative AI — why scientists must oversee its use. *Nature* **622**, 693–696 (2023).

94. Wu, C.-J. et al. Sustainable AI: environmental implications, challenges and opportunities. in *Proceedings of Machine Learning and Systems 4 (MLSys 2022)* vol. 4, 795–813 (2022).

95. Pushkarna, M., Zaldivar, A. & Kjartansson, O. Data cards: purposeful and transparent dataset documentation for responsible AI. in *2022 ACM Conference on Fairness, Accountability, and Transparency* 1776–1826 (Association for Computing Machinery, 2022).

96. Shumailov, I. et al. The curse of recursion: training on generated data makes models forget. Preprint at https://arxiv.org/abs/2305.17493 (2023).

97. Mitchell, M. How do we know how smart AI systems are? *Science* https://doi.org/10.1126/science.adj5957 (2023).

98. Wu, Z. et al. Reasoning or reciting? Exploring the capabilities and limitations of language models through counterfactual tasks. Preprint at https://arxiv.org/abs/2307.02477 (2023).

99. Birjali, M., Kasri, M. & Beni-Hssane, A. A comprehensive survey on sentiment analysis: approaches, challenges and trends. *Knowl. Based Syst.* **226**, 107134 (2021).

100. Acheampong, F. A., Wenyu, C. & Nunoo Mensah, H. Text based emotion detection: advances, challenges, and opportunities. *Eng. Rep.* **2**, e12189 (2020).

101. Pauca, V. P., Shahnaz, F., Berry, M. W. & Plemmons, R. J. Text mining using non-negative matrix factorizations. in *Proc. 2004 SIAM International Conference on Data Mining* 452–456 (Society for Industrial and Applied Mathematics, 2004).

102. Sharma, A., Amrita, Chakraborty, S. & Kumar, S. Named entity recognition in natural language processing: a systematic review. in *Proc. Second Doctoral Symposium on Computational Intelligence* (eds Gupta, D., Khanna, A., Kansal, V., Fortino, G. & Hassanien, A. E.) 817–828 (Springer Singapore, 2022).

103. Nasar, Z., Jaffry, S. W. & Malik, M. K. Named entity recognition and relation extraction: state-of-the-art. *ACM Comput. Surv.* **54**, 1–39 (2021).

104. Sedighi, M. Application of word co-occurrence analysis method in mapping of the scientific fields (case study: the field of informetrics). *Library Rev.* **65**, 52–64 (2016).

105. El-Kassas, W. S., Salama, C. R., Rafea, A. A. & Mohamed, H. K. Automatic text summarization: a comprehensive survey. *Exp. Syst. Appl.* **165**, 113679 (2021).

Competing interests

The authors declare no competing interests.

Additional information

Peer review information *Nature Reviews Methods Primers* thanks F. Jannidis, L. Nelson, T. Tangherlini and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.