https://doi.org/10.1038/s44159-024-00392-z

Review article

Check for updates

Using natural language processing to analyse text data in behavioural science

Stefan Feuerriegel[1,2] ✉, Abdurahman Maarouf[1,2], Dominik Bär[1,2], Dominique Geissler[1,2], Jonas Schweisthal[1,2], Nicolas Pröllochs[3], Claire E. Robertson[4], Steve Rathje[4], Jochen Hartmann[5], Saif M. Mohammad[6], Oded Netzer[7], Alexandra A. Siegel[8], Barbara Plank[2,9] & Jay J. Van Bavel[4,10]

Abstract

Language is a uniquely human trait at the core of human interactions. The language people use often reflects their personality, intentions and state of mind. With the integration of the Internet and social media into everyday life, much of human communication is documented as written text. These online forms of communication (for example, blogs, reviews, social media posts and emails) provide a window into human behaviour and therefore present abundant research opportunities for behavioural science. In this Review, we describe how natural language processing (NLP) can be used to analyse text data in behavioural science. First, we review applications of text data in behavioural science. Second, we describe the NLP pipeline and explain the underlying modelling approaches (for example, dictionary-based approaches and large language models). We discuss the advantages and disadvantages of these methods for behavioural science, in particular with respect to the trade-off between interpretability and accuracy. Finally, we provide actionable recommendations for using NLP to ensure rigour and reproducibility.

Sections

[1]LMU Munich School of Management, LMU Munich, Munich, Germany. [2]Munich Center for Machine Learning, Munich, Germany. [3]Department of Business and Economics JLU Giessen, Giessen, Germany. [4]Department of Psychology, New York University, New York, NY, USA. [5]TUM School of Management, Technical University of Munich, Munich, Germany. [6]National Research Council Canada, Ottawa, Ontario, Canada. [7]Columbia Business School, Columbia University, New York, NY, USA. [8]Department of Political Science, University of Colorado, Boulder, CO, USA. [9]Center for Information and Language Processing, LMU Munich, Munich, Germany. [10]Norwegian School of Economics, Bergen, Norway. ✉e-mail: feuerriegel@lmu.de

Introduction

With 5.4 billion people online today, many social interactions now occur through text[1]. Each day, the average person sends and receives approximately 80 emails[2] and 50 text messages[3]. Additionally, people scroll through nearly 300 feet of online content daily, equivalent to reading every page of the *New York Times* three times over[4,5]. Thus, people probably learn from and communicate through text more now than at any time in human history.

This abundance of text data can be leveraged by behavioural science researchers to understand psychology and capture human behaviour[6–8]. People express and reflect their personality[9], emotional state[10], intentions[11], social group membership[12,13] and psychological well-being[14] through textual communication. For example, text written by individuals can be used to predict their personality traits, such as openness, conscientiousness, extroversion, agreeableness and neuroticism[15], indicating a strong connection between language use and underlying psychological processes[16,17]. Thus, text data can provide a window into human behaviour, and researchers can use such data for both theory building and theory testing.

To analyse text data, researchers use techniques from natural language processing (NLP), a field at the intersection of computer science, artificial intelligence and linguistics[18]. NLP involves the development and use of algorithms and models that enable computers to process and generate human language in a way that is both meaningful and useful. NLP approaches can range from simple methods, such as counting word frequencies[19], to advanced techniques, such as using large language models (LLMs) to generate text[20,21]. NLP outputs can then inform downstream analyses that are grounded in theoretical concepts from behavioural science.

NLP offers several advantages over manual analyses of text historically used in behavioural science. The biggest advantage of NLP over manual text analysis is computational: automated analysis makes it possible to analyse large datasets, such as thousands of social media posts, emails or digitized books. Consequently, researchers can conduct studies with larger sample sizes than those based on manual analysis of text data, leading to more fine-grained analysis and more generalizable findings.

NLP also offers practical advantages. First, manual analysis of text data in multiple languages is typically limited by researcher expertise and costs. By contrast, NLP tools (such as machine translation) can handle multiple languages, which enables use of multilingual datasets and therefore broadens the scope of behavioural science research across different cultural contexts[21,22]. Such multilingual research is particularly important for enhancing diversity, equity and inclusion, and generating comprehensive evidence beyond WEIRD (Western, educated, industrialized, rich and democratic) countries[23–25]. Second, NLP facilitates measurement of complex constructs that are challenging to quantify manually. For example, NLP can be used to measure the novelty of ideas by comparing text against large corpora of existing literature[26–28] or to detect linguistic bias[29,30] by comparing the frequency and context of specific words or phrases across different datasets (which would be difficult to detect manually due to the sheer volume of data and the nuanced nature of such biases).

NLP methods have been used by behavioural scientists for decades, but they are growing in popularity as they become cheaper to implement and easier to use. It is therefore important for non-experts to have clear guidelines on how different NLP methods should be chosen and rigorously applied, particularly for fields in which NLP is not part of the core methodology.

In this Review, we provide a high-level overview of NLP methods and their applications in behavioural science. First, we review potential NLP applications and introduce associated methods (such as sentiment analysis). Second, we outline the NLP pipeline and explain the underlying modelling approaches, including dictionary-based approaches and LLMs. We discuss the advantages and disadvantages of these methods for behavioural science, with particular consideration of the trade-off between interpretability (the extent to which researchers can understand the decision logic underlying the inferences made) and accuracy (the extent to which NLP outputs align with the ground truth, such as human judgements or other objective measures). Finally, we provide guidelines for the rigorous use of NLP in behavioural science. Throughout our Review, we adopt an interdisciplinary perspective that aims to bridge computational and behavioural science.

Applications of NLP in behavioural science

NLP offers a powerful toolset for analysing text data from various sources (Box 1) to infer psychological constructs (Table 1). The goal of the research dictates the methods used. In this section, we discuss three main applications of text analysis in behavioural science research: exploratory content analysis; annotating text by psychological construct; and relating constructs to behavioural outcomes.

Exploratory content analysis

Exploratory content analysis using NLP aims to uncover patterns, themes or insights from large text datasets without predefined hypotheses (Fig. 1a). Exploratory content analysis is particularly useful for generating descriptive insights and understanding narratives that might not be immediately apparent (for example, discourse patterns on social media[31]). Although exploratory content analysis can be used as a stand-alone approach, it is often used to gain a preliminary understanding of text data prior to more focused analyses or modelling. As such, it can also serve as a basis for generating behavioural hypotheses and for supporting qualitative research (for example, by identifying themes or by coding interviews and open-ended survey questions[32]).

Exploratory content analysis can encompass a range of NLP methods. Common approaches include frequency-based analyses (for example, counting the most commonly used terms), co-occurrence analysis (for example, studying how often words appear together), named entity recognition (for example, identifying people, organizations or locations) and clustering approaches (for example, categorizing documents into meaningful topics). Visualization (for example, word clouds) is often used to make the extracted patterns more discernible. Researchers often use a combination of methods to explore the text data in a systematic manner.

Exploratory content analyses using NLP can also be used to track changes in language use over time. For example, representing words and documents as vectors in high-dimensional space can reveal changes in language that reflect cultural, societal or topical changes[33,34], such as shifts in public discourse[35] or public sentiment towards specific topics[36]. Thus, NLP tools can offer insights into how concepts and meanings of words evolve.

However, it is crucial to distinguish between words and concepts when tracking changes in language as well as in NLP analyses more generally[37,38]: Words function as symbols to denote concepts, whereas concepts represent collective understandings that are shaped by their use in society. This distinction is important because word usage might change over time or differ between social groups while the underlying concept they represent might remain stable or shift in different ways.

Box 1 | Text sources

Various offline and online text sources are used in behavioural science, each with distinct advantages and disadvantages. Offline text sources (such as letters[178] and essays[179]) often capture day-to-day activities and provide insights into human behaviour under real-world conditions. For example, letters enable researchers to study private interpersonal behaviour that is not shared publicly. However, a substantial challenge is that these text sources must first be obtained and then digitized. Furthermore, the size and representativeness of offline text sources are often limited, which can make it challenging to generalize findings.

Online text sources offer several advantages: they provide easy access to digitized data, enable real-time data collection through designated application programming interfaces or web scraping interfaces and facilitate longitudinal studies that capture behaviour over time. Online text sources are often large and capture a diverse sample of individuals (including under-researched populations who typically do not participate in psychological studies). Prominent examples of online text sources that have been used in behavioural science research include social media posts[12,35,54,56,145], forums[176,180,181], news articles[78,182], emails[178], web searches[183], digitized books[165,184] and online diaries[174].

However, online text sources also pose challenges for researchers owing to noise (for example, the presence of spammers, bots or other automated accounts[185]), inaccuracies in self-reported socio-demographic information and sample populations that are not representative of the broader society (for example, online text sources often reflect a young, tech-savvy population[186]). Text from some online sources (such as social media platforms and forums) also tend to reflect more extreme political or religious views than the average population because people are more likely to share extremist opinions in light of the anonymity of the Internet[187–190] and because there are specialized platforms where such extremist opinions are not prohibited[191]. Nevertheless, online spheres are important for daily life and therefore both reveal and influence behaviours related to information consumption, communication and opinion formation.

Some research examines the link between offline and online behaviour by integrating datasets from both sources. Such research can reveal how online narratives develop in response to offline events (for example, to understand collective responses to terrorist attacks[174,175] and global pandemics[166,176,192]), as well as how online social interactions influence offline behaviour (such as vaccine hesitancy[193], individuals' psychological state[145,177] and violence[36,56]). However, aligning data from different sources and making causal inferences with observational data are often challenging.

Finally, natural language processing (NLP) can be used to analyse other linguistic content (such as YouTube or TikTok videos or interviews) after using voice-to-text technology to transcribe speech into written form, and for content explicitly collected as part of a research study (such as free-text fields in survey responses[194,195]).

For example, youth might assign new meanings to existing words, or the word 'freedom' might be used across political discourses but the concept it represents could be interpreted differently depending on the community or context. Thus, researchers should not conflate words with concepts when interpreting NLP analyses.

Annotating text by psychological construct

Text can be annotated to identify psychological constructs in the data. Traditionally, annotating text by psychological constructs was done through manual labelling by humans. However, NLP can be used to automate this process (Fig. 1b). In simple keyword-based approaches, word frequencies are counted according to some keyword lists. In machine learning approaches, a small set of manually annotated texts (called labels) are used for training the model, which can then annotate a large set of texts automatically. LLMs, in particular, have been used for this purpose[21,39,40]. LLMs offer zero-shot functionality, which means that LLMs do not need task-specific training but can simply annotate a text following a prompt (for example, a prompt to classify the sentiment of a social media post as positive, neutral or negative[41]).

The choice of annotation method is subject to an interpretability–accuracy trade-off. Keyword-based approaches are often reliable and interpretable, thereby ensuring internal validity. However, keyword-based approaches can also be inaccurate if, for example, the authors of the analysed text use irony. By contrast, LLMs can consider both context and semantics, which makes LLMs highly accurate. However, it is often unclear how LLMs annotate text, rendering their decision logic non-interpretable.

Sentiment analysis is used to annotate text in terms of its positive or negative valence[42–44]. Sentiment (also referred to as valence, tone and polarity) can be captured as a categorical variable indicating whether the text is generally positive, negative or neutral, or as a numerical measure indicating how positive-leaning or negative-leaning the text is. Affective computing (algorithms that can detect and respond to human emotions[42,43]) can be used to annotate text on the basis of the perceived affective states of authors or readers (for example, machine learning can anticipate how readers will respond to a text and, thereby, predict their affective states), typically in terms of discrete emotions (such as anger, fear or sadness) or a 2 (valence) × 2 (arousal) model of emotions. Stance detection is used to annotate text based on the presented attitude towards a certain topic, entity or claim (for example, whether the author of the analysed text is in favour, against or neutral towards it)[45,46]. Opinions towards particular aspects of a product or service can be determined using aspect-based sentiment analysis[45] or opinion mining[47].

Relating constructs to outcomes

NLP-generated text features are often integrated into explanatory regression models to isolate and test the effects of specific NLP-derived features on outcomes of interest (Fig. 1c). For example, researchers can examine how specific text characteristics (such as pronoun use) relate to psychological constructs (such as personality traits[48]) or influence cognitive and affective processes[49]. Additionally, psychological constructs can be extracted from text to then study how these influence individual outcomes (for example, how emotional expressions impact attitude formation[50]).

NLP can also be used to build predictive models that can forecast future behaviour or psychological states[51] (such as emotions[52], depression[53,54], anxiety[54], well-being[52,55], mental disorders[53,54] and distress[52]) from text (Fig. 1d). Such predictive models could be used as early warning systems, for example, to predict offline violence during protests from social media posts[56] or identify critical mental states that require medical attention[53]. Predictive models based on text input can be used to inform the design of targeted interventions (for example, tailoring counter-arguments based on textual characteristics[57]).

The NLP pipeline

The process of using NLP involves several key steps: text preprocessing, text representation, modelling and analysis. These steps can vary depending on the models being used (Fig. 2).

Text preprocessing

The first step in NLP is preprocessing[58], which involves cleaning and normalizing the text data. Preprocessing is important for several reasons. First, preprocessing removes noise, such as irrelevant parts of documents and misspellings, which can otherwise lead to incorrect results. Second, text data often come from diverse sources with different formats, languages and styles. Preprocessing steps such as lowercasing and removing stop words (commonly used words in a language that do not carry meaning, such as 'a' or 'the' in English) ensure consistency across the dataset, which makes the data uniform and easier to analyse. Third, text data are typically high-dimensional, with thousands of unique words and phrases. Preprocessing simplifies the dataset by reducing the dimensionality for downstream steps.

Preprocessing should be grounded in knowledge of the data and the intended analyses. For instance, PDF files or other rendered formats typically need to be converted into a machine-readable format. When dealing with text data in a foreign language, machine translation tools (such as Google Translate or DeepL) can be used to translate the different source languages into a common language[59]. However, machine translation tools can fail to accurately capture meaning across cultural contexts[60]. Such inaccuracies are typically difficult to detect unless researchers are bilingual.

Table 1 | Examples of how natural language processing (NLP) has been used to infer psychological constructs

Construct	Description	Examples
Attitudes	Beliefs and values that guide individuals' judgements and actions	Inferring out-group animosity from social media content to study its effect on online engagement[12] Identifying political orientation from texts[141] and then comparing language use across political ideologies[142–144] Measuring how morally tinged messages about political issues are transmitted through social networks[145,146]
Cognitive complexity	The degree of abstract thinking	Measuring cognitive complexity to understand individual differences in personality development[147] Measuring cognitive complexity from press conference and debate speeches to assess differences between political candidates[148] Using language to study how humans learn[149–151]
Communication styles	Patterns in the way individuals communicate	Measuring trends of fact-speaking[152], belief speaking[152] and incivility[153] in political communication Measuring linguistic complexity in news articles to determine whether the simplicity or fluency of language increases reader attention[154] Identifying linguistic cues predictive of deception in offline versus online communication[155–160]
Creativity and novelty	How innovative and original certain ideas are	Measuring creativity of ideation processes when devising metaphors[120] Measuring how innovative new ideas are over time[28,121,122]
Cultural and social norms	Values, beliefs and behaviours considered acceptable in a specific society or group	Identifying shifts in cultural perceptions over time[161] or long-term cultural trends[162,163] Analysing the use and presence of stereotypes over time[164,165] Quantifying polarized rhetoric by humans over time[166]
Emotions	Subjectively experienced feelings usually directed towards a specific object (such as anger, disgust, fear, joy, trust, sadness and surprise)	Inferring discrete emotions in social media posts to study whether angry posts are more likely to go viral[167,168] Characterizing emotions across political ideologies[142–144] Measuring the effect of discrete emotions on the perceived helpfulness of information[169]
Motives	A person's reason for doing something	Extracting AirBnB host motives and relating them to downstream host behaviour[170]
Personality traits	Attributes that describe an individual's consistent patterns of thoughts, feelings and behaviours	Predicting the Big Five personality traits from language use[171]
Psychological well-being	Linguistic markers that are indicative of mental health states such as depression and anxiety	Identifying depression, anxiety and suicidal thoughts in blog posts[172] Determining how user demographics are related to sharing information about mental illness[173]
Sentiment	The overall positivity or negativity of text	Quantifying the sentiment of news headlines to test whether there is a negativity bias in information consumption[78] Tracking mood changes in response to offline events such as terrorist attacks[174,175] and global pandemics[176] Analysing whether exposure to positive (negative) information leads to an increased use of positive (negative) words to test whether emotions are contagious in social interactions[177]

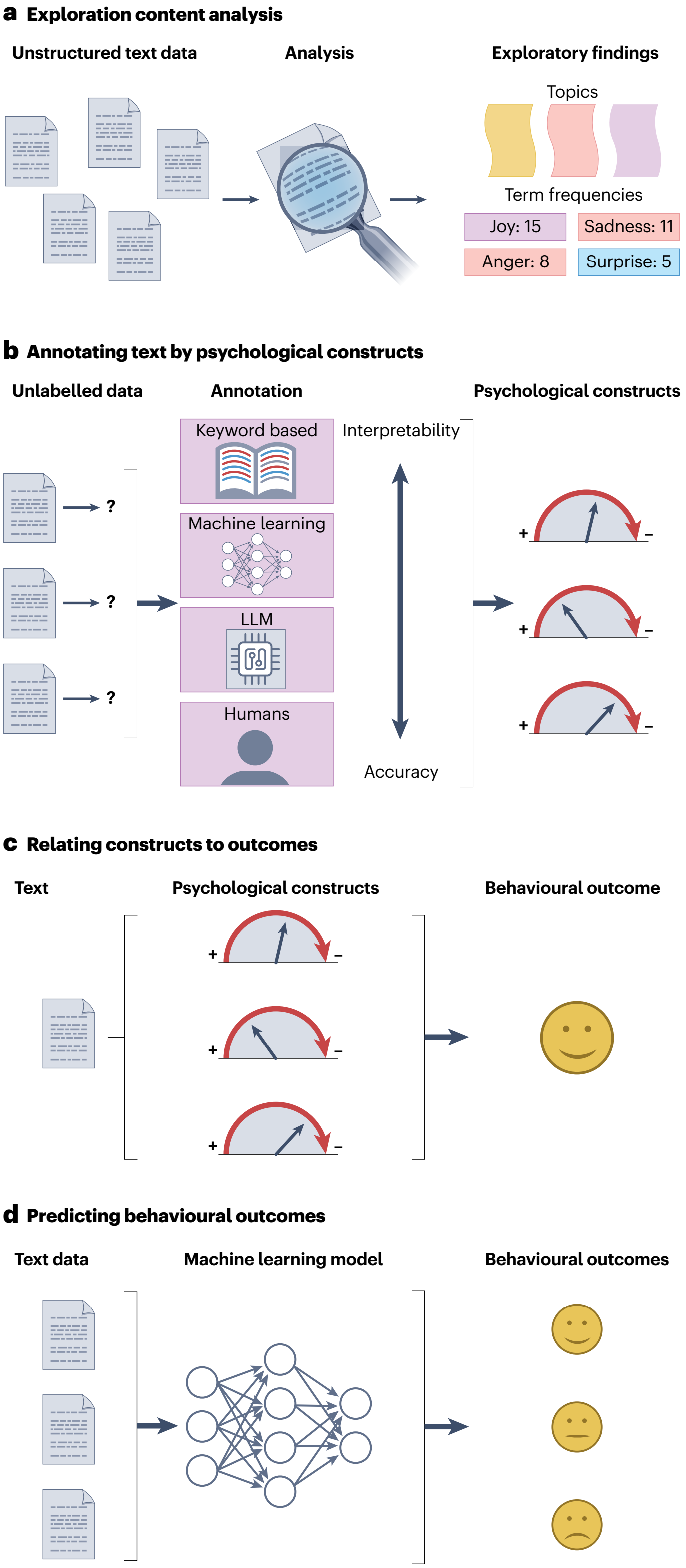

Fig. 1 | Different objectives of natural language processing (NLP) in behavioural science. **a**, In exploratory content analysis, NLP is used to uncover patterns or themes from text data without predefined hypotheses. These exploratory findings often inform and guide more focused subsequent analyses. **b**, Text can be annotated by psychological constructs through simple keyword-based approaches or machine learning approaches. **c**, Measurements of psychological constructs can be integrated into statistical models to test theories about their relationships with behavioural outcomes. **d**, Machine learning models can be trained to predict behavioural outcomes from text data, often for use as an early warning system or to trigger behavioural interventions. LLM, large language model.

There are several common preprocessing steps for normalizing text data (Box 2). These typically include converting all characters to lowercase, removing punctuation, correcting misspellings, removing URLs and replacing smileys with text-based placeholders. Although lowercasing is usually done first because it helps to standardize the text, the order of the other steps is generally of little importance (that is, preprocessing step order might lead to very minor quantitative differences but the qualitative findings will typically be consistent).

The choice of preprocessing steps depends on the specific task at hand. For example, if the purpose of the analysis is to extract common nouns (such as company names or names of people), researchers might skip lowercase conversion because capitalization can be informative for identifying these nouns in the data. Similarly, if the purpose of the analysis is linguistic style (studying how authors write instead of what they write), researchers might skip stemming (reducing words to their base or root form) and lemmatization (grouping words that have the same inflected forms), and avoid removing stop words such as pronouns because these words can be indicative of writing style. Researchers might also need to customize the dictionary of stop words to include context-specific words that appear in most text units within the corpus they are working with (for example, the term ‘review’ in online reviews).

The preprocessing steps required can also vary depending on the downstream methods that will be used. For example, feature-based approaches (where texts are carefully cleared and features are manually extracted from text) often require more extensive preprocessing compared with LLMs[44] because LLMs are designed to capture complex patterns in text in a data-driven way.

Finally, there is no single best approach to preprocessing[61]; it often involves trade-offs between accuracy and simplicity. Thus, different approaches could be tested as a robustness check. To ensure reproducibility, researchers should document and report all preprocessing steps (for example, by releasing the underlying code)[62]. Researchers should also convert data into a platform-independent format (such as UTF-8) to ensure compatibility across different operating systems.

Text representation

The next step in NLP after preprocessing is to transform the text into a numerical format that can be processed more effectively. Such numerical formats are often referred to as ‘representations’ or ‘features’. Two representations are especially common: the bag-of-words model and the paragraph vector model.

In the bag-of-words model[63], the frequency of words in a text document is counted. The output is typically a document–term matrix, which represents the text in terms of how often terms appear in each document. This approach has two main disadvantages: it loses the order of words (and therefore the context-dependent meaning of individual words or even texts as a whole) and it includes many frequent but non-informative words (such as ‘the’). The former is an inherent issue

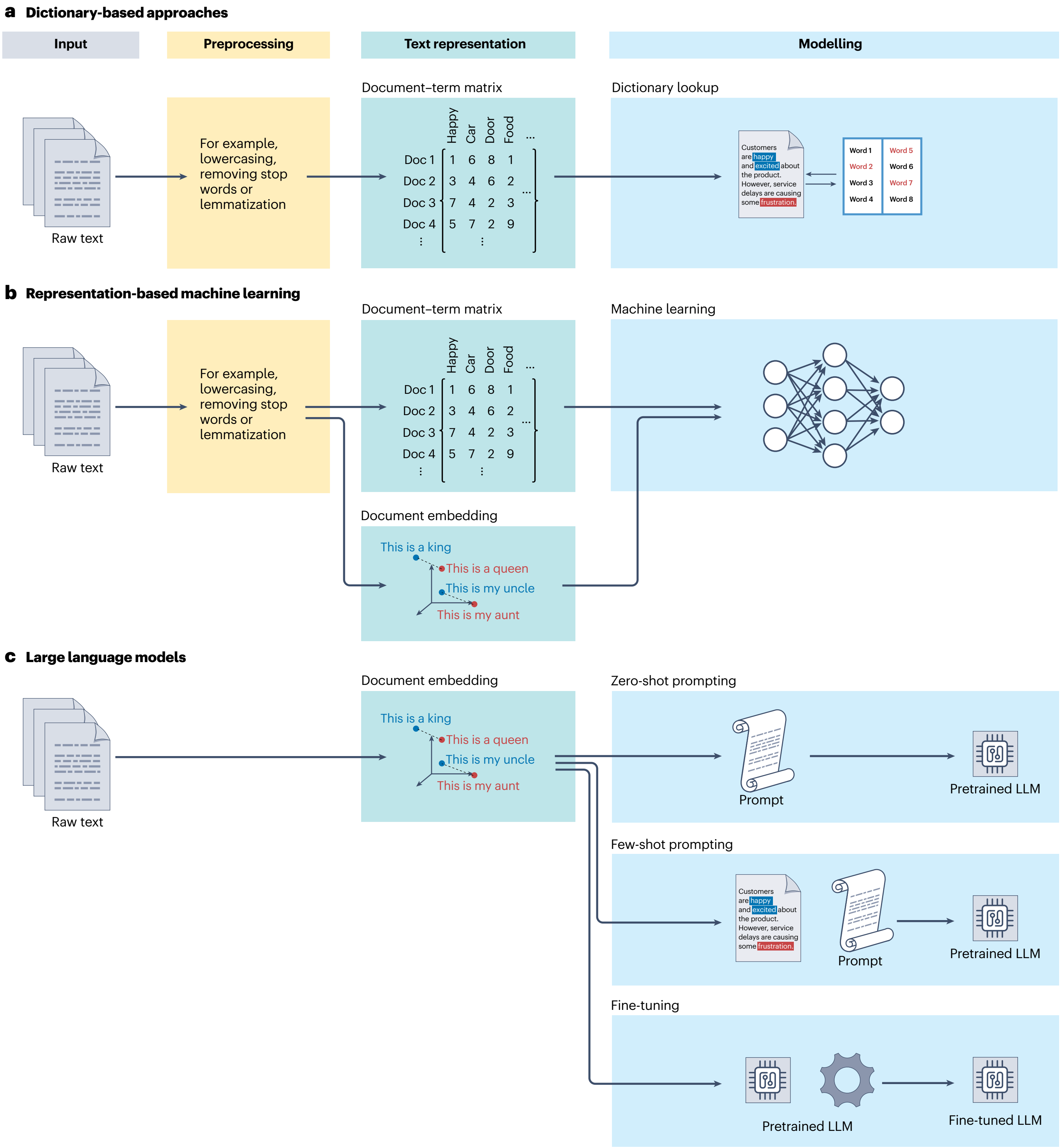

Fig. 2 | Overview of the natural language processing (NLP) pipeline. The NLP pipeline varies depending on the underlying model used. **a**, Dictionary-based approaches require preprocessing of the raw text and then a dictionary lookup is used to annotate the text according to a psychological construct. **b**, In representation-based machine learning, preprocessed text is transformed into a numerical format (document–term matrix or embedding-based representation). Machine learning models are then applied to predict annotations. **c**, Large language models (LLMs) can be prompted to output an annotation directly (zero-shot prompting), a small set of labelled data can be provided with the prompt to guide the model (few-shot prompting) or the model can be fine-tuned to a specific task. LLMs typically require minimal preprocessing steps.

with the bag-of-words model, whereas the latter can be addressed by applying a transformation that maps absolute frequencies onto relative frequencies weighted by how often words generally appear in the dataset. Feature selection (removing rare terms so that only words that appear with relatively high frequency are kept, such as the top 500 words) can be used to deal with the high-dimensional nature of the document–term matrix.

In the paragraph vector model[64], vector representations for entire paragraphs or documents – rather than just individual words[65–67] – are computed. The paragraph vector model performs exceptionally well in capturing semantic meaning in documents of varying lengths[64,68].

The output of the paragraph vector model is called word embedding or document embedding, depending on whether embeddings (mathematical representations in the form of vectors) are being computed for a set of documents or individual words. Embeddings are typically computed via large neural networks that place each word (document) vector into a high-dimensional space, so that similar words (documents) are closer together and dissimilar words (documents) are more distant. Thus, word (document) embeddings are grouped by semantic similarity[69]. For example, the embedding for 'happy' would be closer to that of 'joyful' than to that of 'depressed'. Embeddings are computationally advantageous because they are typically low-dimensional and continuous (compared with document–term matrices which are high-dimensional and sparse). Word embeddings can be created through pretrained, neural language models such as Word2Vec (ref. 66) and GloVe[67]; document embeddings are typically created using advanced transformer models (deep learning neural network models that capture relationships in sequential data) such as BERT[70]. LLMs can also be used for creating (document) embeddings.

The choice of text representation (that is, whether to use the bag-of-words model or the paragraph vector model) depends on the downstream task and the specific requirements of the analysis. For example, dictionary-based modelling approaches typically use a document–term matrix as input, whereas LLMs use embeddings. There is also a trade-off between interpretability and accuracy. Specifically, the bag-of-words model is interpretable but ignores word order, and therefore context, meaning that it might be inaccurate when sentences contain negations or irony. The bag-of-words model is therefore useful when researchers want to understand how psychological constructs are being extracted from text data. By contrast, the paragraph vector model can capture semantic meanings but lacks interpretability because the embeddings in the paragraph vector model are high-dimensional vectors where the individual dimensions have no direct interpretation. Representing text as embeddings using the paragraph vector model is therefore preferred when psychological constructs need to be measured with high accuracy.

Supervised modelling methods

The text representation is used as input in modelling, which outputs the variable of interest (such as a psychological construct). There are two common supervised NLP modelling approaches that use labelled data: dictionary-based approaches and machine learning. Machine learning can be further subdivided into representation-based methods and LLMs. Each approach has its strengths and limitations. The choice of modelling approach depends on the specific research question and is typically governed by the trade-off between interpretability and accuracy (Fig. 3).

In general, approaches that work out of the box, such as dictionary-based approaches and general-purpose LLMs, are powerful for initial prototyping, which in the context of NLP refers to quickly testing ideas or hypotheses before committing to more complex or customized methods. For example, researchers might use a simple dictionary to get a quick sense of whether the emotional tone of a set of interview transcripts correlates with some behavioural outcome of interest, whereas other machine learning models are generally needed for specialized questions (for example, predicting the risk of depression for a certain patient cohort[14]) or when dealing with multimodal data (for example, predicting receptivity to misinformation from a combination of text and network data[71]). If possible, researchers should use more than one approach to demonstrate the robustness of their findings.

Box 2 | Common preprocessing steps in natural language processing (NLP)

- Lemmatization: a technique that groups together words that have the same inflected forms. For example, lemmatization reduces 'better' to 'good'.
- Named entity recognition: identification of entities such as names of people, organizations, locations and dates. Depending on the research question, named entities are often stripped from the text, so that the downstream analysis can focus on the linguistic style.
- Negation handling: a common method of negation handling is to treat negated terms as pseudo-words (for example, transforming phrases such as 'not good' into 'not-good'), which preserves the context but introduces the inverted meaning.
- *n*-Grams: an *n*-gram is a sequence of *n* tokens or words that are used as features in the downstream analyses. Common examples are bigrams (n=2) and trigrams (n=3). *n*-Grams help to capture word order, and therefore semantics.
- One-hot encoding: a technique to convert words into binary vectors, where the values are all zero except for the one element that refers to the word and has a value of one. For example, 'cat', 'dog' and 'fish' are represented as [1, 0, 0], [0, 1, 0] and [0, 0, 1].
- Part-of-speech tagging: assigns parts of speech (nouns, verbs and adjectives) to each word in a text based on the context. Part-of-speech tagging is useful for syntactic analysis.
- Stemming: a technique similar to lemmatization used to reduce words to their base or root form. For instance, stemming reduces the words 'fishing', 'fished' and 'fisher' to 'fish'. Stemming is a common feature reduction technique.
- Stop word removal: the process of removing common words (for example, 'and', 'the' and 'is') that appear frequently in texts but offer little value in understanding the meaning. Stop word removal reduces noise.
- Term frequency–inverse document frequency: a statistical measure used to evaluate how 'characteristic' a word is to a document in a collection or corpus. This measure increases proportionally to the number of times a word appears in the document (term frequency) but is downweighted by the overall frequency of the word in the corpus (inverse document frequency), which adjusts for the fact that some words appear more frequently in general.
- Tokenization: the process of breaking down text into smaller units called tokens, which can be words, syllables, phrases or symbols.

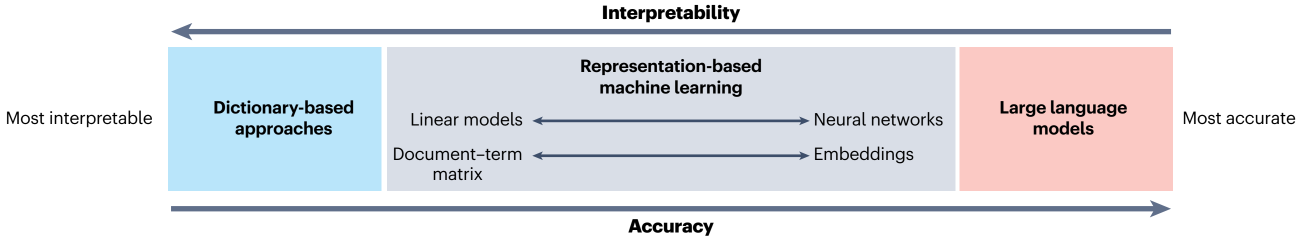

Fig. 3 | Interpretability–accuracy trade-off in supervised natural language processing (NLP) models. Dictionary-based approaches are highly interpretable; researchers can easily understand how specific words or phrases relate to psychological constructs. However, these methods might lack accuracy when dealing with complex language use, such as sarcasm or context-contingent word meanings. Representation-based methods vary in terms of interpretability and accuracy depending on the underlying machine learning model and text representation. For example, combining document–term matrices with linear models allows for interpretability in the sense that researchers can understand the exact logic of how inferences are made. By contrast, models based on neural networks and/or embeddings have typically good accuracy but their interpretability is restricted to simple post hoc explanations. Large language models are not interpretable; however, they typically have the highest accuracy of all methods owing to their consideration of context and semantics.

Dictionary-based approaches. Dictionary-based approaches (Fig. 2a) assign labels to documents on the basis of predefined lists of keywords (so-called dictionaries[72,73]). These dictionaries classify words into predefined categories, such as positive, neutral and negative[74], as well as more complex categories such as emotions[75], moral language[76], political orientation[75] and hate speech[77]. The frequency of these words within a text is counted and used to calculate a numerical score, such as the ratio of positive to negative words[78].

Dictionary-based approaches offer several benefits: they are scalable, easy to implement and highly interpretable. Researchers can manually inspect the word lists to understand precisely how inferences were made. Furthermore, many dictionaries such as Linguistic Inquiry and Word Count[79] have been extensively validated[10]. For common constructs such as sentiment, multiple dictionaries are available[80–82], enabling researchers to compare results using different dictionaries to ensure robustness. However, dictionary-based approaches might be less accurate than machine learning methods[21,44] owing to their inability to capture contextual information, such as idioms or nuanced semantics (for example, 'high' is positive in the context of revenue but negative in the context of blood pressure), or to interpret literary devices such as sarcasm. Furthermore, translating dictionaries and creating new dictionaries are labour-intensive and time-consuming.

There are several methods for constructing dictionaries. One method is to have human annotators (such as online survey participants or experts) classify words. Such manual annotation ensures that dictionary classifications align with human perception. For example, the Linguistic Inquiry and Word Count dictionary was created through expert annotation. Other methods seek to reduce the effort of manual annotation. For example, one method is to analyse co-occurrence patterns using a manually defined set of seed words, but where the number of seed words is very small and where all subsequent steps are then automated. For example, if clearly positive or negative words are used as seed words, all words that frequently appear in close proximity would also be classified as positive or negative[83]. Finally, in regression approaches a small set of documents are labelled manually according to a construct, and then the labelled data are used in a regression analysis where the (weighted) term frequencies are regressed to explain the psychological construct of interest. Again, such regression approaches require a small set of labels from human annotation but, in subsequent steps, the approach can be applied to label much larger datasets. The regression coefficients indicate the dictionary classification. Regression is often combined with regularization techniques to avoid overfitting and to address multicollinearity[84].

Dictionary-based approaches are recommended when researchers care about interpretability (that is, how psychological constructs are measured). Even if a machine learning approach is chosen, it is often beneficial to incorporate a dictionary-based approach as a robustness check that allows for interpretability. Human validation (comparing the labels assigned by NLP with those assigned by human annotators, typically for a smaller subset of the original dataset) is typically needed to ensure the accuracy of dictionary-based approaches[85], especially if researchers create their own dictionary.

Representation-based machine learning. In representation-based machine learning (Fig. 2b), text is mapped onto a suitable representation and then entered into a machine learning model to predict labels of interest (for example, the psychological construct). There are two main approaches for making predictions from text representations — feature-based representations and embedding-based representation — both of which require extensive technical expertise.

In feature-based approaches, covariates (such as term frequencies) are computed from texts and then entered into a machine learning model that is trained in a supervised manner using annotated labels that represent the psychological constructs of interest. Different machine learning models can be used. For instance, a simple approach is to input columns of the document–term matrix into a regularized linear regression model[84]. Alternatively, word order can be captured through more advanced methods such as recurrent neural networks[86]. One advantage of feature-based representations is that inferences are easy to understand when interpretable machine learning models (for example, linear regression or decision trees) are chosen; performance for these linear models can be compared with performance for non-linear models (such as neural networks) to help to interpret the importance of more accurate but less interpretable methods.

Embedding-based approaches use dense vector representations (embeddings) to capture the meaning of words or phrases within their context by positioning similar words or phrases closer together in a high-dimensional space. The embeddings can either be used as inputs for machine learning models or can be fine-tuned to a dataset to further optimize the model for a specific task without incurring high computational and monetary costs[87]. Although embedding-based approaches cannot be interpreted in a straightforward manner because they encode text into vectors in a high-dimensional space, they are very good at modelling complex linguistic patterns. For instance, embeddings from the BERT language model can differentiate meanings of the same word based on its surrounding context in the sentence[70].

Both feature-based and embedding-based approaches require a sizable number of annotated training samples (typically $n > 1{,}000$), which can be costly and time-consuming to obtain. Following best practices in machine learning, researchers must ensure rigorous splitting of training and testing datasets to avoid overfitting, hyperparameter tuning (testing different configurations of the machine learning model, such as different sizes of neural networks) should be performed to optimize model performance and the steps for hyperparameter tuning should be reported to promote reproducibility.

Large language models. LLMs can generate human-like text based on prompts (for example, using techniques such as next-word prediction[20]). Notable examples of LLMs are GPT[88] (which is the LLM behind ChatGPT) and Llama[89]. Methodologically, LLMs are machine learning models based on the transformer architecture[90] that are designed to efficiently process and generate text data. LLMs excel in various tasks, such as annotating text with psychological constructs[21,40,91,92], answering questions[93], summarizing content[94] and machine translation[95], often without specific training. This versatility can be attributed to the fact that LLMs are pretrained on large, diverse datasets.

LLMs can be used to make inferences in three ways (Fig. 2c): zero-shot prompting, few-shot prompting and fine-tuning. In zero-shot prompting, the LLM directly responds to a prompt without any additional data or training. This 'out-of-the-box' approach is user-friendly and versatile[41,96]. Different prompting strategies can be used to improve accuracy, such as chain-of-thought reasoning in which the task is broken down into smaller, incremental subtasks with clear instructions[96]. However, experimenting with different prompting strategies is crucial to ensure robustness and optimal performance because LLM outputs can be sensitive to the exact prompt wording[91]. In general, prompt effectiveness is highly dependent on the specific task, and the best prompt is often only found after extensive experimentation[97]. In few-shot prompting, a small set of labelled documents or examples (typically a few dozen are sufficient) are provided with the prompt to guide the LLM. Finally, the LLM can be fine-tuned for a specific task using a large set of labelled data. Fine-tuning can improve performance on the desired task[92] but comes with large computational and monetary costs. In general, the effectiveness of fine-tuning is highly task-specific and data-specific. For some tasks, suitable datasets to use for fine-tuning might not be available, or the costs of data annotation might be prohibitively high. In other cases, fine-tuning might not yield performance improvements because the LLM has already acquired the required knowledge during training or because the data are of poor quality[98,99].

LLMs have several advantages for behavioural science[91,92]. One important advantage is their high accuracy for measuring psychological constructs from text data[21]. For example, previous research found that LLMs can accurately predict intercorrelations between different personality scale items[100]. Moreover, LLMs are relatively easy to use in combination with prompting, which eliminates the need for manual tuning and reduces the technical knowledge required for analysis. LLMs can also automatically analyse text in different languages without having to translate the text first[21]. Although LLMs were initially limited to specialized computer systems because of their high resource demands, more recent models (such as Llama-3.1 8B) can work on traditional desktop computers while still competing with state-of-the-art LLMs[101].

However, LLMs also have notable shortcomings[91,92]. First, there is essentially no transparency regarding how LLMs make inferences[102]. For example, although LLM-based approaches could predict whether a person has a mental disorder from social media posts, LLMs could not be used to identify which linguistic cues (such as the use of pronouns) predict a mental disorder. In the latter case, dictionary-based methods and simpler machine learning models (such as linear models) are preferrable. Second, LLMs are prone to algorithmic biases, often repeating stereotypes present in their training data or other historically ingrained biases[103–107]. For example, LLMs have been found to generate racially biased outputs based on people's dialect[104]. Thus, an interpretable approach might be preferrable when processing texts from disadvantaged groups because the decision logic can be audited. Third, it can be difficult to reproduce results from LLMs because of their probabilistic nature (LLMs often rely on randomness in their internal processes, such as during model initialization and sampling when generating outputs) and fast pace of development (for example, the software packages used for certain LLMs are no longer maintained). Fourth, LLM creators might limit certain outputs such as comments on political orientation, or the use of swear words, which might be relevant to certain behavioural science research questions. Fifth, LLMs are typically less accurate for languages with limited available data for training (such as dialects or Indigenous languages) because many state-of-the-art LLMs are trained primarily using English text. Finally, although proprietary LLMs often represent the state-of-the-art, they can limit reproducibility because their underlying architectures, training data and model parameters are not publicly available. Moreover, the proprietary nature of LLMs aggravates the problem of implicit data leakage because it is unclear exactly what data the LLM have been trained on. If the LLM has already encountered a dataset during its training phase, it might 'leak' this knowledge during analysis, providing artificially high performance or biased results[108]. Both issues could undermine the reliability of research findings[102,109].

Unsupervised methods

In contrast to the supervised modelling methods described above, unsupervised methods operate without labelled data. Instead, they identify patterns within a dataset. Here, we describe two common unsupervised approaches: topic modelling and text similarity.

Topic modelling. Topic modelling uncovers the underlying themes within a large collection of text documents. Thus, the objective is primarily exploratory. That is, the topics are not known a priori but inferred in a bottom-up manner.

There are several popular methods for topic modelling. Latent semantic analysis is based on the document–term matrix and uses singular value decomposition to reduce the dimensionality of the document–term matrix and infer relevant topics[110]. It is relatively simple and intuitive. However, latent semantic analysis does not consider context, which can be problematic for synonyms or for understanding certain expressions. In addition, latent semantic analysis is not computationally scalable.

Latent Dirichlet allocation is a probabilistic approach where each document is assumed to be a mixture of various topics, with each topic distributed over words[111]. Latent Dirichlet allocation is interpretable and is typically better suited than latent semantic analysis for capturing topics in large collections of documents owing to its probabilistic nature, which enables it to better capture variability in large text datasets. However, latent Dirichlet allocation is computationally expensive and, similar to latent semantic analysis, does not capture context-specific meanings of words.

Other approaches to topic modelling involve using embeddings to create document representations, which are then clustered using algorithms such as *k*-means or DBSCAN[112]. These approaches capture the text semantics but they are not interpretable. Researchers have also developed end-to-end frameworks for embedding-based topic modelling (for example, BERTopic[113]).

Choosing the number of topics in topic modelling is inherently challenging, and in general there is no 'optimal' number of topics. Instead, the choice depends on the desired granularity and the domain of study. A common strategy is to fit several topic models with varying numbers of topics and then compare them on different metrics, such as perplexity (how well the model can predict topics for unseen data[114]). Some methods (such as HDBSCAN[115]) can suggest a recommended number of topics. Importantly, insights should not depend on a specific number of topics, and researchers should repeat the analysis with different numbers of topics to ensure that the results remain qualitatively consistent.

Given the challenges in choosing the preferred number of topics, validation is crucial to ensure that the topics are meaningful. One approach is to use visualization, where a technique for dimensionality reduction (such as *t*-SNE[116]) is applied on top of the embeddings, so that researchers can visualize the clusters to see whether they are disjoint. Another strategy is human validation[117]. For example, to validate whether content in a topic is coherent, human respondents can be asked to identify an intruding word from a set of characteristic words for each topic (word intrusion test); or to validate the assignment of documents to topics, respondents can be given a text and asked to identify the correct topic among other intruding topics (topic intrusion test).

Naming topics in topic modelling is typically done manually by considering the common words for each topic and then assigning a name that best captures the joint theme of those words. The relevance score measures how important a word is within a topic, which can be helpful for naming the underlying theme in the topic[118]. LLMs have also been used to assist in the naming process by prompting the LLM to suggest topic names[119]. Validation through manual analysis is crucial to confirm that the identified topics are meaningful.

Text similarity. Text similarity measures how different documents are from each other, which is often used as a proxy for novelty or creativity[28,120–122]. Text similarity has also been used to understand the similarity of interpersonal communication and language use between individuals, such as in employer–employee relationships[123] or in dating behaviour[124].

Classic similarity or distance measures include the Levenshtein distance, which measures distances in characters and is given by the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other. It is effective for measuring the similarity at the character level (for example, for understanding how severe spelling errors are). A common measure at the document level is the Jaccard similarity, which evaluates the number of common words within two texts. Both of these similarity metrics can also be used for linguistic style matching (assessments of whether texts from two different sources are similar in word choices, syntax and other language features[125]), but often with additional preprocessing (for example, comparing the term frequencies of function words only). However, neither Levenshtein distance nor Jaccard similarity capture the context of words, thereby potentially missing semantic nuances.

Advanced approaches to assessing text similarity use distance or similarity metrics in the embedding space. The most widely used metric is cosine similarity, which, unlike the classical Euclidean distance (the length of a line segment between two embedding vectors), adjusts for inflated distances in high-dimensional embeddings[126]. However, not all embeddings can be used together with cosine similarity, so caution is indicated[127].

In general, similarity or distance measures should be selected based on the research question (for example, whether one cares more about similarity at the character level or at the document level, or similarity in terms of topics or style). Researchers should also consider whether they are interested in similarity in terms of writing style or content (how and what somebody writes, respectively). If embedding space is used for analysing similarity, it is important to understand which embeddings are used and how they were created. Finally, researchers should consider multiple metrics and perform checks to ensure that the results are reasonable and consistent.

Analysis

Inferred psychological constructs (or other NLP modelling outputs) serve as input for downstream analysis. In explanatory analysis, inferred psychological constructs are entered into a statistical model (such as a regression model) for hypothesis testing. There are several important considerations when conducting explanatory analysis. First, the inferred measurements of psychological constructs themselves can be noisy, potentially leading to a ripple effect in the analysis, whereby inaccuracies in the initial measurement can propagate through the subsequent steps of the analysis. For example, if the NLP methods introduce noise when inferring psychological constructs, those errors might distort the results of the statistical analysis, leading to incorrect conclusions about relationships between variables or the strength of observed effects. To address this issue, a SIMEX correction (a statistical technique that accounts for measurement bias) can be applied[128]. Further, many NLP analyses involve large datasets. Thus, statistical analyses should focus on effect sizes – rather than just statistical significance – to determine the practical importance of the estimated effects. Interpreting effect sizes should be guided by theory, and researchers should assess whether effect sizes are consistent with existing knowledge. Comparing observed effect sizes with theoretical predictions is crucial for contextualizing the magnitude of effect sizes and interpreting their practical significance.

In prediction studies, the goal is to assess how accurately inferences about psychological constructs can be made for new, previously unseen data points (for example, for individuals not included in previous datasets). This evaluation process requires carefully assessing the model's ability to generalize beyond the data it was trained on. Specifically, the evaluation should focus on testing the model's out-of-sample performance (that is, its accuracy when applied to data that were not used during the model's training phase). Further, it is often helpful to use a naive baseline (a simple model used as a point of comparison). A naive baseline provides a minimal level of performance that the more complex model should surpass, which helps to contextualize the prediction performance of the complex model and ensures that the complex model has accurately learned the relationships between text data and labels. For example, to assess the contribution of text data to the overall prediction, the prediction model can be compared with a simple baseline model without text, a model with a more parsimonious structure (for example, a linear model instead of a non-linear model) or even an implausible model trained on irrelevant information (for example, how often the letter 'x' appeared in a document)[129].

It is also possible to make causal inferences from text data[130], such as the causal effect of text data (for example, the positivity or negativity of a news headline) on some behavioural outcome (for example, news consumption)[78]. In such cases, additional care is needed to ensure that the typical assumptions of causal inference hold[130]. Several tailored approaches have been developed for causal inference with text data to address this issue[131].

Recommendations

There is little consensus on how to use NLP methods in behavioural science, and we therefore offer a practical checklist in Box 3. Below, we highlight several overarching recommendations.

First, validation is crucial to ensure the accuracy of measured constructs, and therefore the validity of any downstream analysis. Validation is particularly important when the primary motive for using NLP is to efficiently replace costly human judgement and when the constructs being measured are new and underexplored. The most common validation approach is human validation, where labels assigned by the NLP method are compared with those assigned by human annotators, typically for a smaller subset of the original dataset[85]. Researchers should follow best practices for validating NLP methods against human annotations[85]: multiple annotators should be recruited and they should be given proper training to reduce individual inaccuracies and variability in perceptions; inter-rater reliability should be checked to ensure that labels were applied consistently across different annotators; annotations should be conducted in batches to protect against annotator fatigue, which can influence the quality of annotations[132]; and the validation sample should be sufficiently large and heterogeneous to capture any idiosyncrasies associated with rare labels. There are also method-specific validation approaches. For instance, dictionaries can be validated by asking humans to verify whether the assignment of each word matches their judgement.

Given the range of methodological choices (especially for machine learning models), extensive methodological validation is also crucial to assess model reliability. In particular, results from complex – and potentially more accurate – methods should be compared against those from simpler but interpretable methods as a robustness check. Consistency in these results helps to validate that NLP methods (including black-box models) measure psychological constructs as intended.

Second, transparent reporting of all aspects of modelling (for example, hyperparameters that control the specification of machine learning models, software libraries and implementation details) as well as preprocessing and analysis steps is essential. Online appendices and supplements in journals, including the possibility of providing code, make transparent reporting easier because they allow researchers to provide further implementation details that might not fit within the main article. Transparent reporting also requires that the choice of methods is carefully justified (for example, whether interpretability or accuracy was prioritized and why). Further, the psychological construct being measured must be clearly defined and interpreted to avoid ambiguity. Language can reflect aspects of the author and/or affect the intended audience[42,133,134] and researchers should clarify which they are interested in. For example, when quantifying emotional language, it should be made clear whether the measure reflects the emotions elicited in readers, the emotions of the writer or simply the frequency of emotion words embedded in language.

Third, researchers should take steps to enable reproducibility. Ideally, researchers should publicly release their code as well as their data to the extent possible given privacy and ethical concerns. For machine learning, researchers should make available their trained models and model weights after training (for example, via the model hub on Hugging Face). Proprietary tools can undergo changes without notice, which makes it impossible to replicate results if the software version changes[102,109]. Thus, to support reproducibility, open-source models should be used over proprietary software. However, researchers might have to make trade-offs between accuracy and reproducibility, especially in cases where proprietary models (such as ChatGPT) outperform open-source models.

Fourth, the ethical challenges of using NLP in behavioural science must be carefully navigated, especially because text data often capture sensitive social interactions. It is therefore important to protect the privacy of individuals when collecting and analysing text data[135] by, for example, removing or obscuring any personally identifiable information. It is also important to ensure that individuals whose data are being used have given explicit consent for their data to be analysed. When using an application programming interface or web scraping, researchers should carefully check the terms of data usage[136]. Finally, the potential effect on individuals and society should be discussed, and potential negative effects must be minimized (for example, some NLP applications such as monitoring risks of violence from social media might give rise to potentially harmful use cases such as surveillance).

Fifth, NLP methods, particularly LLMs, are known to be susceptible to algorithmic bias[29,106,107], which can also lead to inaccuracy when capturing psychological constructs. Thus, it is crucial to use diverse and representative data for calibrating NLP models. Furthermore, researchers must carefully check for algorithmic biases and apply methods for algorithmic bias mitigation[137] (such as corpus-level constraints[138]) or consider alternative modelling approaches if algorithmic bias cannot be eliminated.

Finally, the application of NLP in behavioural science should be anchored in robust theoretical frameworks. On the one hand, researchers should be guided by theory when formulating research questions and identifying which aspects of human behaviour to study using NLP methods. On the other hand, researchers should consider theory as a benchmark against which the findings have to be validated. For example, researchers can assess whether the findings align with theoretical predictions or compare model effect sizes with those of (theoretically) implausible models[129]. A strong grounding in theory can help researchers to distinguish meaningful relationships from spurious correlations and should therefore serve as a critical lens through which findings and effect sizes are interpreted.

Summary and future directions

Text data have become abundant in the Internet age. Thus, NLP offers new ways of studying human behaviour. By providing rich insights into psychological constructs, NLP can support both theory testing and theory building and inform personalized behavioural interventions. For example, predicting critical mental health issues from language use on social media could be used to offer personalized suggestions for help,

Box 3 | Reporting checklist for using natural language processing (NLP) in behavioural science

Step 1: define the research question

- Clearly describe the psychological phenomena or behaviour to be studied.
- Ensure the research question (or hypothesis) is specific and measurable.
- Define the psychological constructs to be measured and ensure the feasibility of measurement of these constructs from text data.

Step 2: collect and prepare text data

- Detail the sources of text data (for example, social media, clinical notes or surveys) and report summary statistics (for example, corpus size or average document length).
- Describe the data collection process (for example, time of data collection, software tools and/or interfaces, or inclusion or exclusion criteria).
- Ensure the data collection process is performed in an ethical and legal manner (for example, is aligned with legal frameworks, does not violate copyright laws and has ethical approval by the research institution) and discuss ethical considerations and informed consent of individuals.
- Decide on the appropriate prepossessing steps (tokenization, stemming, lemmatization, removal of stop words or negation handling) based on the research questions.
- Address potential issues due to missing data and describe how they are handled.
- Choose appropriate text representations (bag-of-words or document embeddings) and justify the choice.

Step 3: select NLP methods

- Choose suitable NLP methods (for example, dictionary-based approaches or large language models (LLMs)) based on the research question.
- Justify the choice of NLP method for the research question.
- For dictionary-based approaches, discuss and justify the choice of the dictionary.

Step 4: perform NLP analysis

- For dictionary-based approaches, state the exact equation used to compute numerical scores.
- For machine learning analysis, split data into training and testing sets (or use cross-validation) to ensure robust model evaluation and evaluate model performance using appropriate metrics (for example, accuracy or precision) and report these metrics. Report details about the implementation (for example, how hyperparameter tuning was performed).
- For LLMs, state the exact prompts and LLM used, including the version number.

Step 5: assess construct validity

- Assess the validity of the psychological construct (for example using human annotation).

Step 6: perform robustness checks

- Conduct sensitivity analyses to assess the stability of the results under different conditions or parameter settings (for example, using a different dictionary or using different machine learning models that are more interpretable or more flexible).
- Compare results using different NLP methods to ensure robustness (for example, LLMs could be compared with a dictionary-based approach that is simple but interpretable).
- Analyse the robustness of the preprocessing steps.
- If appropriate, use tools for post hoc explanation of model predictions to ensure alignment with psychological theory.
- Report and interpret the results of robustness checks.

Step 7: interpret and report results

- Clearly describe the findings (including effect sizes) and their implications for theory and practice.
- Compare the results with existing literature and discuss possible reasons for any discrepancies.
- Acknowledge the limitations of the study and the NLP methods used (for example, whether the data are representative of the population being studied).

Step 8: ensure reproducibility

- Clearly detail all steps involved in the NLP analysis in the manuscript or supplementary material.
- Make data collection, data preprocessing and analysis code publicly available.
- Release data if possible, otherwise discuss the conditions under which data will be made available or how it can be accessed (for example, social media posts often cannot be released publicly, but the corresponding identifiers can be shared so that the analysis can be repeated).

and predicting personality traits from language use could be used to personalize educational interventions and improve learning outcomes.

Our Review offers several actionable recommendations that should be followed closely to ensure the success of NLP-based analysis in behavioural science. First, the choice of the underlying NLP method and the steps involved in the analysis should be guided by both the research question and the available data. Second, NLP methods often involve trade-offs between accuracy and interpretability, which must be carefully considered. Third, researchers should adhere to best practice recommendations to ensure the reliability of the NLP methods and the reproducibility of the overall analysis.

The advent of LLMs will have a profound effect on behavioural science, and we expect to see a growing dominance of LLMs owing to their potential to provide nuanced and sophisticated analyses of text data and their ease of use. With further improvements in accuracy and reductions in resource requirements, it is likely that LLMs will become the 'go to' method for many research questions. However, LLMs might not be suitable for all research purposes because of the risk of language bias in LLMs and/or their lack of interpretability, which could undermine research aimed at theory building. Thus, researchers should be mindful of the interpretability–accuracy trade-off and should have a broad repertoire of methods at hand, including dictionary-based approaches and traditional machine learning methods, which will remain crucial to ensure robust results.

The accessibility of NLP tools has greatly improved in the past 5–10 years. However, there are several directions for the field to develop. First, the emergence of new use cases – often enabled by advances in LLMs – presents exciting research opportunities. For example, studies published in the past 2 years have used NLP to generate persuasive messages tailored to specific personality profiles[139] and to build therapeutic chatbots[140]. The ability of LLMs to interact with computer systems using natural language raises questions about how such technologies will change human language. Second, the development of standardized tools and workflows will probably improve the reliability and comparability of analyses. Third, although NLP development is largely driven by computer science research, there are numerous opportunities to customize NLP methods for behavioural science. This could include developing tailored tools and benchmarks for validation to ensure the reliability of NLP in measuring psychological constructs.

Published online: 2 January 2025

References

1. Dixon, S. J. Number of social media users worldwide from 2017 to 2028. *Statista* https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/ (2024).
2. Ceci, L. Number of sent and received e-mails per day worldwide from 2018 to 2027. *Statista* https://www.statista.com/statistics/456500/daily-number-of-e-mails-worldwide/ (2024).
3. GilPress. WhatsApp statistics, users, demographics as of 2024. *What's the Big Data* https://whatsthebigdata.com/whatsapp-statistics/ (2023).
4. Robertson, C. E., Shariff, A. & van Bavel, J. J. Morality in the anthropocene: the perversion of compassion and punishment in the online world. *PNAS Nexus* **3,** pgae193 (2024).
5. Morant, L. The truth behind 6 second ads. *Medium* https://medium.com/@Lyndon/the-tyranny-of-six-seconds-592b94160877 (2019).
6. Wilkerson, J. & Casas, A. Large-scale computerized text analysis in political science: opportunities and challenges. *Annu. Rev. Political Sci.* **20**, 529–544 (2017).
7. Kennedy, B., Ashokkumar, A., Boyd, R. L. & Dehghani, M. in *Handbook of Language Analysis in Psychology* (eds Dehghani M. & Boyd, R. L.) 3–62 (Guilford, 2022).
8. Jackson, J. C. et al. From text to thought: how analyzing language can advance psychological science. *Perspect. Psychol. Sci.* **17**, 805–826 (2022).
9. Boyd, R. L. & Pennebaker, J. W. Language-based personality: a new approach to personality in a digital world. *Curr. Opin. Behav. Sci.* **18**, 63–68 (2017).
10. Kahn, J. H., Tobin, R. M., Massey, A. E. & Anderson, J. A. Measuring emotional expression with the linguistic inquiry and word count. *Am. J. Psychol.* **120**, 263–286 (2007).
11. Rocklage, M. D., Rucker, D. D. & Nordgren, L. F. Persuasion, emotion, and language: the intent to persuade transforms language via emotionality. *Psychol. Sci.* **29**, 749–760 (2018).
12. Rathje, S., van Bavel, J. J. & van der Linden, S. Out-group animosity drives engagement on social media. *Proc. Natl Acad. Sci. USA* **118**, e2024292118 (2021).
13. Rogers, N. & Jones, J. J. Using Twitter bios to measure changes in self-identity: are Americans defining themselves more politically over time? *J. Soc. Comput.* **2**, 1–13 (2021).
14. Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H. & Eichstaedt, J. C. Detecting depression and mental illness on social media: an integrative review. *Curr. Opin. Behav. Sci.* **18**, 43–49 (2017).
15. Pennebaker, J. W. & King, L. A. Linguistic styles: language use as an individual difference. *J. Pers. Soc. Psychol.* **77**, 1296–1312 (1999).
16. Pennebaker, J. W., Chung, C. K., Frazee, J., Lavergne, G. M. & Beaver, D. I. When small words foretell academic success: the case of college admissions essays. *PLoS ONE* **9**, e115844 (2014).
17. Pennebaker, J. W. & Francis, M. E. Cognitive, emotional, and language processes in disclosure. *Cogn. Emot.* **10**, 601–626 (1996).
18. Manning, C. & Schütze, H. *Foundations of Statistical Natural Language Processing* (MIT Press, 1999).
19. Tausczik, Y. R. & Pennebaker, J. W. The psychological meaning of words: LIWC and computerized text analysis methods. *J. Lang. Soc. Psychol.* **29**, 24–54 (2009).
20. Feuerriegel, S., Hartmann, J., Janiesch, C. & Zschech, P. Generative AI. *Bus. Inf. Syst. Eng.* **66**, 111–126 (2024).
21. Rathje, S. et al. GPT is an effective tool for multilingual psychological text analysis. *Proc. Natl Acad. Sci. USA* **121**, e2308950121 (2024).
22. Steigerwald, E. et al. Overcoming language barriers in academia: machine translation tools and a vision for a multilingual future. *BioScience* **72**, 988–998 (2022).
23. Henrich, J., Heine, S. J. & Norenzayan, A. Most people are not WEIRD. *Nature* **466**, 29 (2010).
24. Ghai, S. It's time to reimagine sample diversity and retire the WEIRD dichotomy. *Nat. Hum. Behav.* **5**, 971–972 (2021).
25. Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D. & Majid, A. Over-reliance on English hinders cognitive science. *Trends Cognit. Sci.* **26**, 1153–1170 (2022).
26. Shibayama, S., Yin, D. & Matsumoto, K. Measuring novelty in science with word embedding. *PLoS ONE* **16**, e0254034 (2021).
27. Just, J., Ströhle, T., Füller, J. & Hutter, K. AI-based novelty detection in crowdsourced idea spaces. *Innovation* **6**, 359–386 (2023).
28. Toubia, O. & Netzer, O. Idea generation, creativity, and prototypicality. *Mark. Sci.* **36**, 1–20 (2017).
29. Blodgett, S. L., Barocas, S., Daumé III, H. & Wallach, H. Language (technology) is power: a critical survey of "bias" in NLP. In *Proc. Annual Meet. Assoc. Computational Linguistics* (eds. Jurafsky, D. et al.) 5454–5476 (ACL, 2020).
30. Garg, Nikhil, Schiebinger, Londa, Jurafsky, Dan & Zou, James Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc. Natl Acad. Sci. USA* **115**, E3635–E3644 (2018).
31. Page, R. *Narratives Online: Shared Stories in Social Media* (Cambridge Univ. Press, 2018).
32. Yu, C. H., Jannasch-Pennell, A. & DiGangi, S. Compatibility between text mining and qualitative research in the perspectives of grounded theory, content analysis, and reliability. *Qualitative Rep.* **16**, 730–744 (2011).
33. Hamilton, W. L., Leskovec, J. & Jurafsky, D. Diachronic word embeddings reveal statistical laws of semantic change. In *Proc. Annual Meet. Assoc. Computational Linguistics* (eds. Erk, K. & Smith, N.) 1489–1501 (ACL, 2016)
34. Kulkarni, V., Al-Rfou, R., Perozzi, B. & Skiena, S. Statistically significant detection of linguistic change. In *Proc. Int. Conf. World Wide Web* (eds. Gangemi, A. et al.) 625–635 (ACM, 2015)
35. Dunivin, Z. O., Yan, H. Y., Ince, J. & Rojas, F. Black lives matter protests shift public discourse. *Proc. Natl Acad. Sci. USA* **119**, e2117320119 (2022).
36. Jakubik, J., Vössing, M., Pröllochs, N., Bär, D. & Feuerriegel, S. Online emotions during the storming of the US capitol: evidence from the social media network Parler. In *Proc. Int. AAAI Conf. Web and Social Media* 423–434 (AAAI, 2023).
37. Murphy, G. *The Big Book of Concepts*. (MIT Press, 2004).
38. Boroditsky, L. Does language shape thought?: Mandarin and English speakers' conceptions of time. *Cognit. Psychol.* **43**, 1–22 (2001).
39. Gilardi, F., Alizadeh, M. & Kubli, M. ChatGPT outperforms crowd workers for text-annotation tasks. *Proc. Natl Acad. Sci. USA* **120**, e2305016120 (2023).
40. Ziabari, A. S. et al. Reinforced multiple instance selection for speaker attribute prediction. In *Proc. Conf. North American Chapter of the Assoc. Computational Linguistics: Human Language Technologies* (eds. Duh, K., Gomez, H. & Bethard, S.) 3307–3321 (ACL, 2024)
41. Krugmann, J. O. & Hartmann, J. Sentiment analysis in the age of generative AI. *Customer Needs Solut.* **11**, 3 (2024).
42. Mohammad, S. M. in *Emotion Measurement* (ed. Meiselman, H. L.) 201–237 (Elsevier, 2016)
43. Kratzwald, B., Ilić, S., Kraus, M., Feuerriegel, S. & Prendinger, H. Deep learning for affective computing: text-based emotion recognition in decision support. *Decis. Support. Syst.* **115**, 24–35 (2018).
44. Hartmann, J., Heitmann, M., Siebert, C. & Schamp, C. More than a feeling: accuracy and application of sentiment analysis. *Int. J. Res. Mark.* **40**, 75–87 (2023).
45. Mohammad, S. M., Kiritchenko, S., Sobhani, P., Zhu, X. & Cherry, C. SemEval-2016 Task 6: detecting stance in tweets. In *Proc. Int. Workshop on Semantic Evaluation* (eds. Bethard, S. et al.) 31–41 (ACL, 2016).
46. Mohammad, S. M., Sobhani, P. & Kiritchenko, S. Stance and sentiment in tweets. *ACM Trans. Internet Technol. Argumentati. Soc. Media* **17**, 3 (2017).
47. Liu, B. & Zhang, L. in *Mining Text Data* (eds Aggarwal, C. C. & Zhai, C.) 415–463 (Springer US, 2012).
48. Spitzley, L. A. et al. Linguistic measures of personality in group discussions. *Front. Psychol.* **13**, 887616 (2022).
49. Lutz, B., Adam, M., Feuerriegel, S., Pröllochs, N. & Neumann, D. Which linguistic cues make people fall for fake news? A comparison of cognitive and affective processing. In *Proc. ACM on Human–Computer Interaction* (eds. Nichols, Jeff) 1–22 (ACM, 2024).
50. van Kleef, G. A., van den Berg, H. & Heerdink, M. W. The persuasive power of emotions: effects of emotional expressions on attitude formation and change. *J. Appl. Psychol.* **100**, 1124–1142 (2015).
51. Schwartz, H. A. et al. Personality, gender, and age in the language of social media: the open-vocabulary approach. *PLoS ONE* **8**, e73791 (2013).
52. Vine, V., Boyd, R. L. & Pennebaker, J. W. Natural emotion vocabularies as windows on distress and well-being. *Nat. Commun.* **11**, 4525 (2020).
53. Eichstaedt, J. C. et al. Facebook language predicts depression in medical records. *Proc. Natl Acad. Sci. USA* **115**, 11203–11208 (2018).
54. Chen, S., Zhang, Z., Wu, M. & Zhu, K. Detection of multiple mental disorders from social media with two-stream psychiatric experts. In *Proc. Conf. Empirical Methods in Natural Language Processing* (eds. Bouamor, H., Pino, J. & Bali, K.) 9071–9084 (ACL, 2023).
55. Eichstaedt, J. C. et al. Psychological language on Twitter predicts county-level heart disease mortality. *Psychol. Sci.* **26**, 159–169 (2015).
56. Mooijman, M., Hoover, J., Lin, Y., Ji, H. & Dehghani, M. Moralization in social networks and the emergence of violence during protests. *Nat. Hum. Behav.* **2**, 389–396 (2018).
57. Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C. & Lee, L. Winning arguments: interaction dynamics and persuasion strategies in good-faith online discussions. In *Proc. Int. Conf. World Wide Web* (eds. Bourdeau, J. et al.) 613–624 (ACM, 2016).
58. Denny, M. J. & Spirling, A. Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it. *Political Anal.* **26**, 168–189 (2018).

59. Toetzke, M., Banholzer, N. & Feuerriegel, S. Monitoring global development aid with machine learning. *Nat. Sustain.* **5**, 533–541 (2022).
60. Tenzer, H., Feuerriegel, S. & Piekkari, R. AI machine translation tools must be taught cultural differences too. *Nature* **630**, 820 (2024).
61. Fokkens, A. et al. Offspring from reproduction problems: what replication failure teaches us. In *Proc. Annual Meet. Assoc. Computational Linguistics* (eds. Schuetze, H., Fung, P. & Poesio, M.) 1691–1701 (ACL, 2013).
62. Ulmer, D. et al. Experimental standards for deep learning in natural language processing research. In *Findings of the Association for Computational Linguistics: Empirical Methods in Natural Language Processing* (eds. Goldberg, Y., Kozareva, Z. & Zhang, Y.) 2673–2692 (ACL, 2022).
63. Salton, G. *A Theory of Indexing* (Society for Industrial and Applied Mathematics, 1975).
64. Le, Q. & Mikolov, T. Distributed representations of sentences and documents. In *Proc. Int. Conf. Machine Learning* 1188–1196 (PMLR, 2014)
65. Collobert, R. & Weston, J. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proc. Int. Conf. Machine Learning* 160–167 (ACM, 2008).
66. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* (eds. Burges, C. J. et al.) 3111–3119 (Curran Associates Inc., 2013).
67. Pennington, J., Socher, R. & Manning, C. D. GloVe: global vectors for word representation. In *Proc. Conf. Empirical Methods in Natural Language Processing* (eds. Moschitti, A., Pang, B. & Daelemans, W.) 1532–1543 (ACL, 2014).
68. Dai, A. M., Olah, C. & Le, Q. V. Document embedding with paragraph vectors. Preprint at https://doi.org/10.48550/arXiv.1507.07998 (2015).
69. Harris, Z. S. *Distributional Structure* (Word, 1954).
70. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding.In *Proc. Conf. North American Chapter of the Assoc. Computational Linguistics* (eds. Burstein, J., Doran, C. & Solorio, T.) 4171–4186 (ACL, 2019).
71. Tokita, C. K. et al. Measuring receptivity to misinformation at scale on a social media platform. *PNAS Nexus* **3**, page396 (2024).
72. Hart, R. P. & Carroll, C. *DICTION: The Text-Analysis Program* (Sage, 2011).
73. Stone, P. J., Dunphy, D. C. & Smith, M. S. *The General Inquirer: A Computer Approach to Content Analysis* (The MIT Press, 1966).
74. Rinker, T., Goodrich, B. & Kurkiewicz, D. *qdap: Bridging the Gap between Qualitative Data and Quantitative Analysis* (R Project for Statistical Computing, 2013).
75. Mohammad, S. M. & Turney, P. D. Crowdsourcing a word–emotion association lexicon. *Comput. Intell.* **29**, 436–465 (2013).
76. Graham, J., Haidt, J. & Nosek, B. A. Liberals and conservatives rely on different sets of moral foundations. *J. Pers. Soc. Psychol.* **96**, 1029–1046 (2009).
77. The Weaponized Word. Lexicons. *Weaponized Word* https://weaponizedword.org/lexicons (2024).
78. Robertson, C. E. et al. Negativity drives online news consumption. *Nat. Hum. Behav.* **7**, 812–822 (2023).
79. Boyd, R. L., Ashokkumar, A., Seraj, S. & Pennebaker, J. W. *The Development and Psychometric Properties of LIWC-22* (Univ. of Texas at Austin, 2022).
80. Thelwall, M., Buckley, K. & Paltoglou, G. Sentiment strength detection for the social web. *J. Am. Soc. Inf. Sci. Technol.* **63**, 163–173 (2011).
81. Baccianella, S., Esuli, A. & Sebastiani, F. SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Proc. Seventh International Conference on Language Resources and Evaluation (LREC'10)* (eds. Calzolari, N., et al.) http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf (European Language Resources Association, 2010).
82. Hutto, C. & Gilbert, E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. In *Proc. Int. AAAI Conf. Web and Social Media* 216–225 (AAAI, 2014).
83. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D. I. & Kappas, A. Sentiment strength detection in short informal text. *J. Am. Soc. Inf. Sci. Technol* **61**, 2544–2558 (2010).
84. Pröllochs, N., Feuerriegel, S. & Neumann, D. Statistical inferences for polarity identification in natural language. *PLoS ONE* **13**, e0209323 (2018).
85. Song, H. et al. In validations we trust? The impact of imperfect human annotations as a gold standard on the quality of validation of automated content analysis. *Political Commun.* **37**, 550–572 (2020).
86. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
87. Hussain, Z., Mata, R. & Wulff, D. U. Novel embeddings improve the prediction of risk perception. *EPJ Data Sci.* **13**, Article 38 (2024).
88. Brown, T. et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems* (Larochelle, H. et al.) 1877–1901 (Curran Associates Inc., 2020).
89. Touvron, H. et al. Llama: open and efficient foundation language models. Preprint at https://doi.org/10.48550/arXiv.2302.13971 (2023).
90. Vaswani, A. et al. Attention is all you need. In *Advances in Neural Information Processing Systems* (Guyon, I. et al.) 5998–6008 (2017).
91. Demszky, D. et al. Using large language models in psychology. *Nat. Rev. Psychol.* **2**, 688–701 (2023).
92. Abdurahman, S. et al. Perils and opportunities in using large language models in psychological research. *PNAS Nexus* **3**, 245 (2024).
93. Kamalloo, E., Dziri, N., Clarke, C. & Rafiei, D. Evaluating open-domain question answering in the era of large language models. In *Proc. Annual Meet. Assoc. Computational Linguistics* (eds. Rogers, A. et al.) 5591–5606 (ACL, 2023).
94. Zhang, T. et al. Benchmarking large language models for news summarization. *Trans. Assoc. Comput. Linguist.* **12**, 39–57 (2024).
95. Zhu, W. et al. Multilingual machine translation with large language models: empirical results and analysis. In *Findings of the ACL: North American Chapter of the Assoc. Computational Linguistics* (eds. Duh, K. et al.) 2765–2781 (ACL, 2024).
96. Lin, Z. How to write effective prompts for large language models. *Nat. Hum. Behav.* **8**, 611–615 (2024).
97. Atreja, S., Ashkinaze, J., Li, L., Mendelsohn, J. & Hemphill, L. Prompt design matters for computational social science tasks but in unpredictable ways. Preprint at https://doi.org/10.48550/arXiv.2406.11980 (2024).
98. Kuribayashi, T., Oseki, Y. & Baldwin, T. Psychometric predictive power of large language models. In *Findings of the ACL: North American Chapter of the Assoc. Computational Linguistics* (eds. Duh, K. et al.) 1983–2005 (ACL, 2024).
99. Zhang, B., Liu, Z., Cherry, C. & Firat, O. When scaling meets LLM finetuning: the effect of data, model and finetuning method. In *Proc. Int. Conf. Learn. Representations* https://doi.org/10.48550/arXiv.2402.17193 (2024).
100. Wulff, D. U. & Mata, R. Semantic embeddings reveal and address taxonomic incommensurability in psychological measurement. *Nat. Hum. Behav.* https://doi.org/10.1038/s41562-024-02089-y (2025).
101. Dubey, A. et al. The llama 3 herd of models. Preprint at https://doi.org/10.48550/arXiv.2407.21783 (2024).
102. Grimes, M., Krogh, Gvon, Feuerriegel, S., Rink, F. & Gruber, M. From scarcity to abundance: scholars and scholarship in an age of generative artificial intelligence. *Acad. Manag. J.* **66**, 1617–1624 (2023).
103. Shu, B. et al. You don't need a personality test to know these models are unreliable: assessing the reliability of large language models on psychometric instruments. In *Proc. Conf. North American Chapter of the Assoc. Computational Linguistics: Human Language Technologies* (eds. Duh, K. et al.) 5263–5281 (ACL, 2024).
104. Hofmann, V., Kalluri, P. R., Jurafsky, D. & King, S. AI generates covertly racist decisions about people based on their dialect. *Nature* **633**, 147–154 (2024).
105. Caliskan, A., Bryson, J. J. & Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science* **356**, 183–186 (2017).
106. Hartmann, J., Schwenzow, J. & Witte, M. The political ideology of conversational AI: converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. Preprint at https://doi.org/10.48550/arXiv.2301.01768 (2023).
107. Hu, T. et al. Generative language models exhibit social identity biases. Preprint at https://doi.org/10.48550/arXiv.2310.15819 (2023).
108. Balloccu, S., Schmidtová, P., Lango, M. & Dusek, O. Leak, cheat, repeat: data contamination and evaluation malpractices in closed-source LLMs. In *Proc. Conf. European Chapter of the Assoc. Computational Linguistics* (eds. Graham, Y. & Purver, M.) 67–93 (ACL, 2024).
109. Palmer, A., Smith, N. A. & Spirling, A. Using proprietary language models in academic research requires explicit justification. *Nat. Comput. Sci.* **4**, 2–3 (2024).
110. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. & Harshman, R. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* **41**, 391–407 (1990).
111. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003).
112. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. Int. Conf. Knowledge Discovery and Data Mining* (eds. Simoudis, E. et al.) 226–231 (AAAI, 1996).
113. Grootendorst, M. BERTopic: neural topic modeling with a class-based TF-IDF procedure. Preprint at https://doi.org/10.48550/arXiv.2203.05794 (2022).
114. Jelinek, F., Mercer, R. L., Bahl, L. R. & Baker, J. K. Perplexity: a measure of the difficulty of speech recognition tasks. *J. Acoust. Soc. Am.* **62**, S63 (1977).
115. Campello, R. J., Moulavi, D. & Sander, J. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia Conf. Knowledge Discovery and Data Mining* (eds. Pei, J. et al.) https://doi.org/10.1007/978-3-642-37456-2_14 (2013).
116. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
117. Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. & Blei, D. Reading tea leaves: how humans interpret topic models. In *Adv. Neural Inf. Process. Syst.* (eds. Bengio, Y. et al.) 288–296 (Curran Associates Inc., 2009).
118. Sievert, C. & Shirley, K. LDAvis: a method for visualizing and interpreting topics. In *Proc. Workshop on Interactive Language Learning, Visualization, and Interfaces* (eds. Chuang, J. et al.) 63–70 (ACL, 2014).
119. Kosar, A., Pauw, Gde & Daelemans, W. Comparative evaluation of topic detection: humans vs. LLMs. *Comput. Linguist. Neth. J.* **13**, 91–120 (2024).
120. DiStefano, P. V., Patterson, J. D. & Beaty, R. E. Automatic scoring of metaphor creativity with large language models. *Creativity Res. J.* https://doi.org/10.1080/10400419.2024.2326343 (2023).
121. Yu, Y., Chen, L., Jiang, J. & Zhao, N. Measuring patent similarity with word embedding and statistical features. *Data Anal. Knowl. Discov.* **3**, 53–59 (2019).
122. Kelly, B., Papanikolaou, D., Seru, A. & Taddy, M. Measuring technological innovation over the long run. *Am. Econ. Rev. Insights* **3**, 303–320 (2021).
123. Goldberg, A., Srivastava, S. B., Manian, V. G., Monroe, W. & Potts, C. Fitting in or standing out? The tradeoffs of structural and cultural embeddedness. *Am. Sociol. Rev.* **81**, 1190–1222 (2016).
124. Ireland, M. E. et al. Language style matching predicts relationship initiation and stability. *Psychol. Sci.* **22**, 39–44 (2011).
125. Niederhoffer, K. G. & Pennebaker, J. W. Linguistic style matching in social interaction. *J. Lang. Soc. Psychol.* **21**, 337–360 (2002).

126. Dhillon, I. S. & Modha, D. S. Concept decompositions for large sparse text data using clustering. *Mach. Learn.* **42**, 143–175 (2001).
127. Steck, H., Ekanadham, C. & Kallus, N. Is cosine-similarity of embeddings really about similarity? In *Companion Proc. ACM Web Conf.* (eds. Chua, T. et al.) 887–890 (ACM, 2024).
128. Lederer, W. & Küchenhoff, H. A short introduction to the SIMEX and MCSIMEX. *Newsl. R. Proj.* **6**, 26–31 (2006).
129. Burton, J. W., Cruz, N. & Hahn, U. Reconsidering evidence of moral contagion in online social networks. *Nat. Hum. Behav.* **5**, 1629–1635 (2021).
130. Egami, N., Fong, C. J., Grimmer, J., Roberts, M. E. & Stewart, B. M. How to make causal inferences using texts. *Sci. Adv.* **8**, eabg2652 (2022).
131. Feder, A. et al. Causal inference in natural language processing: estimation, prediction, interpretation and beyond. *Trans. Assoc. Comput. Linguist.* **10**, 1138–1158 (2022).
132. Maarouf, A., Bär, D., Geissler, D. & Feuerriegel, S. HQP: a human-annotated dataset for detecting online propaganda. In *Findings of the ACL* (eds. Ku, L. et al.) 6064–6089 (ACL, 2024).
133. Berger, J. et al. Uniting the tribes: using text for marketing insight. *J. Mark.* **84**, 1–25 (2020).
134. Mohammad, S. M. Ethics sheet for automatic emotion recognition and sentiment analysis. *Comput. Linguist.* **48**, 239–278 (2022).
135. Rivers, C. M. & Lewis, B. L. Ethical research standards in a world of big data. *F1000Research* **3**, 38 (2014).
136. Boegershausen, J., Datta, H., Borah, A. & Stephen, A. T. Fields of gold: scraping web data for marketing insights. *J. Mark.* **86**, 1–20 (2022).
137. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv.* **54**, 1–35 (2021).
138. Zhao, J., Wang, T., Yatskar, M., Ordonez, V. & Chang, K.-W. Men also like shopping: reducing gender bias amplification using corpus-level constraints. In *Proc. Conf. Empirical Methods in Natural Language Processing* (eds. Palmer, M. et al.) 2989–2989 (ACL, 2017).
139. Hackenburg, K. & Margetts, H. Evaluating the persuasive influence of political microtargeting with large language models. *Proc. Natl Acad. Sci. USA* **121**, e2403116121 (2024).
140. Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C. & Althoff, T. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nat. Mach. Intell.* **5**, 46–57 (2023).
141. Colleoni, E., Rozza, A. & Arvidsson, A. Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *J. Commun.* **64**, 317–332 (2014).
142. Wojcik, S. P., Hovasapian, A., Graham, J., Motyl, M. & Ditto, P. H. Conservatives report, but liberals display, greater happiness. *Science* **347**, 1243–1246 (2015).
143. Frimer, J. A., Brandt, M. J., Melton, Z. & Motyl, M. Extremists on the left and right use angry, negative language. *Pers. Soc. Psychol. Bull.* **45**, 1216–1231 (2019).
144. Sterling, J., Jost, J. T. & Bonneau, R. Political psycholinguistics: a comprehensive analysis of the language habits of liberal and conservative social media users. *J. Pers. Soc. Psychol.* **118**, 805–834 (2020).
145. Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A. & van Bavel, J. J. Emotion shapes the diffusion of moralized content in social networks. *Proc. Natl Acad. Sci. USA* **114**, 7313–7318 (2017).
146. Brady, W. J., Wills, J. A., Burkart, D., Jost, J. T. & van Bavel, J. J. An ideological asymmetry in the diffusion of moralized content on social media among political leaders. *J. Exp. Psychol.: Gen.* **148**, 1802–1813 (2019).
147. Lanning, K., Pauletti, R. E., King, L. A. & McAdams, D. P. Personality development through natural language. *Nat. Hum. Behav.* **2**, 327–334 (2018).
148. Slatcher, R. B., Chung, C. K., Pennebaker, J. W. & Stone, L. D. Winning words: individual differences in linguistic style among US presidential and vice presidential candidates. *J. Res. Pers.* **41**, 63–75 (2007).
149. Wiechmann, P., Lora, K., Branscum, P. & Fu, J. Identifying discriminative attributes to gain insights regarding child obesity inHispanic preschoolers using machine learning techniques. In *Proc. IEEE Int. Conf. Tools with Artificial Intelligence*, 11–15 (IEEE, 2017).
150. Teague, S. J. & Shatte, A. B. R. Exploring the transition to fatherhood: feasibility study using social media and machine learning. *JMIR Pediatrics Parent.* **1**, e12371 (2018).
151. Joel, S., Eastwick, P. W. & Finkel, E. J. Is romantic desire predictable? Machine learning applied to initial romantic attraction. *Psychol. Sci.* **28**, 1478–1489 (2017).
152. Lasser, J. et al. From alternative conceptions of honesty to alternative facts in communications by US politicians. *Nat. Hum. Behav.* **7**, 2140–2151 (2023).
153. Frimer, J. A. et al. Incivility is rising among American politicians on Twitter. *Soc. Psychol. Pers. Sci.* **14**, 259–269 (2023).
154. Shulman, H. C., Markowitz, D. M. & Rogers, T. Reading dies in complexity: online news consumers prefer simple writing. *Sci. Adv.* **10**, eadn2555 (2024).
155. Newman, M. L., Pennebaker, J. W., Berry, D. S. & Richards, J. M. Lying words: predicting deception from linguistic styles. *Pers. Soc. Psychol. Bull.* **29**, 665–675 (2003).
156. Zhou, L., Burgoon, J. K., Nunamaker, J. F. & Twitchell, D. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. *Group. Decis. Negotiation* **13**, 81–106 (2004).
157. Ho, S. M., Hancock, J. T., Booth, C. & Liu, X. Computer-mediated deception: strategies revealed by language–action cues in spontaneous communication. *J. Manag. Inf. Syst.* **33**, 393–420 (2016).
158. Siering, M., Koch, J.-A. & Deokar, A. V. Detecting fraudulent behavior on crowdfunding platforms: the role of linguistic and content-based cues in static and dynamic contexts. *J. Manag. Inf. Syst.* **33**, 421–455 (2016).
159. Zhang, D., Zhou, L., Kehoe, J. L. & Kilic, I. Y. What online reviewer behaviors really matter? Effects of verbal and nonverbal behaviors on detection of fake online reviews. *J. Manag. Inf. Syst.* **33**, 456–481 (2016).
160. Constâncio, A. S., Tsunoda, D. F., Silva, H. F. N., Da Silveira, J. M. & Carvalho, D. R. Deception detection with machine learning: a systematic review and statistical analysis. *PLoS ONE* **18**, e0281323 (2023).
161. Thompson, B., Roberts, S. G. & Lupyan, G. Cultural influences on word meanings revealed through large-scale semantic alignment. *Nat. Hum. Behav.* **4**, 1029–1038 (2020).
162. Morin, O. & Acerbi, A. Birth of the cool: a two-centuries decline in emotional expression in Anglophone fiction. *Cogn. Emot.* **31**, 1663–1675 (2017).
163. Jackson, J. C., Gelfand, M., De, S. & Fox, A. The loosening of American culture over 200 years is associated with a creativity-order trade-off. *Nat. Hum. Behav.* **3**, 244–250 (2019).
164. Charlesworth, T. E. S. & Banaji, M. R. Patterns of implicit and explicit attitudes: I. Long-term change and stability from 2007 to 2016. *Psychol. Sci.* **30**, 174–192 (2019).
165. Charlesworth, T. E. S., Caliskan, A. & Banaji, M. R. Historical representations of social groups across 200 years of word embeddings from Google Books. *Proc. Natl Acad. Sci. USA* **119**, e2121798119 (2022).
166. Simchon, A., Brady, W. J. & van Bavel, J. J. Troll and divide: the language of online polarization. *PNAS Nexus* **1**, pgac019 (2022).
167. Pröllochs, N., Bär, D. & Feuerriegel, S. Emotions explain differences in the diffusion of true vs. false social media rumors. *Sci. Rep.* **11**, 22721 (2021).
168. Pröllochs, N., Bär, D. & Feuerriegel, S. Emotions in online rumor diffusion. *EPJ Data Sci.* **10**, 51 (2021).
169. Yin, D., Bond, S. D. & Zhang, H. Anxious or angry? Effects of discrete emotions on the perceived helpfulness of online reviews. *MIS Q.* **38**, 539–560 (2014).
170. Chung, J., Johar, G. V., Li, Y., Netzer, O. & Pearson, M. Mining consumer minds: downstream consequences of host motivations for home-sharing platforms. *J. Consum. Res.* **48**, 817–838 (2022).
171. Park, G. et al. Automatic personality assessment through social media language. *J. Pers. Soc. Psychol.* **108**, 934–952 (2015).
172. O'Dea, B. et al. The relationship between linguistic expression in blog content and symptoms of depression, anxiety, and suicidal thoughts: a longitudinal study. *PLoS ONE* **16**, e0251787 (2021).
173. Preotiuc-Pietro, D. et al. The role of personality, age, and gender in tweeting about mental illness. In *Proc. 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* 21–30 (ACL, 2015).
174. Cohn, M. A., Mehl, M. R. & Pennebaker, J. W. Linguistic markers of psychological change surrounding September 11, 2001. *Psychol. Sci.* **15**, 687–693 (2004).
175. Garcia, D. & Rimé, B. Collective emotions and social resilience in the digital traces after a terrorist attack. *Psychol. Sci.* **30**, 617–628 (2019).
176. Ashokkumar, A. & Pennebaker, J. W. Social media conversations reveal large psychological shifts caused by COVID-19's onset across US cities. *Sci. Adv.* **7**, eabg7843 (2021).
177. Di Kramer, A., Guillory, J. E. & Hancock, J. T. Experimental evidence of massive-scale emotional contagion through social networks. *Proc. Natl Acad. Sci. USA* **111**, 8788–8790 (2014).
178. Kacewicz, E., Pennebaker, J. W., Davis, M., Jeon, M. & Graesser, A. C. Pronoun use reflects standings in social hierarchies. *J. Lang. Soc. Psychol.* **33**, 125–143 (2014).
179. Rude, S., Gortner, E.-M. & Pennebaker, J. Language use of depressed and depression-vulnerable college students. *Cogn. Emot.* **18**, 1121–1133 (2004).
180. Netzer, O., Feldman, R., Goldenberg, J. & Fresko, M. Mine your own business: market-structure surveillance through text mining. *Mark. Sci.* **31**, 521–543 (2012).
181. Seraj, S., Blackburn, K. G. & Pennebaker, J. W. Language left behind on social media exposes the emotional and cognitive costs of a romantic breakup. *Proc. Natl Acad. Sci. USA* **118**, e2017154118 (2021).
182. Berger, J. & Milkman, K. L. What makes online content viral? *J. Mark. Res.* **49**, 192–205 (2012).
183. Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M. & Duncan, J. W. Predicting consumer behavior with web search. *Proc. Natl Acad. Sci. USA* **107**, 17486–17490 (2010).
184. Scheffer, M., van de Leemput, I., Weinans, E. & Bollen, J. The rise and fall of rationality in language. *Proc. Natl Acad. Sci. USA* **118**, e2107848118 (2021).
185. Ferrara, E., Varol, O., Davis, C., Menczer, F. & Flammini, A. The rise of social bots. *Commun. ACM* **59**, 96–104 (2016).
186. Auxier, B. & Anderson, M. *Social Media Use in 2021* (Pew Research Center, 2021).
187. Barberá, P. & Rivero, G. Understanding the political representativeness of Twitter users. *Soc. Sci. Comput. Rev.* **33**, 712–729 (2015).
188. Schoenmueller, V., Netzer, O. & Stahl, F. The polarity of online reviews: prevalence, drivers and implications. *J. Mark. Res.* **57**, 853–877 (2020).
189. Robertson, C. E., Del Rosario, K., Rathje, S. & van Bavel, J. J. Changing the incentive structure of social media may reduce online proxy failure and proliferation of negativity. *Behav. Brain Sci.* **47**, e81 (2024).
190. Robertson, C., Del Rosario, K. & van Bavel, J. J. *Inside the Funhouse Mirror Factory: How Social Media Distorts Perceptions of Norms* (OSF, 2024).
191. Bär, D., Pröllochs, N. & Feuerriegel, S. New threats to society from free-speech social media platforms. *Commun. ACM* **66**, 37–40 (2023).
192. Zhunis, A., Lima, G., Song, H., Han, J. & Cha, M. Emotion bubbles: emotional composition of online discourse before and after the COVID-19 outbreak. In *Proc. ACM Web Conf.* (eds. Faforest, F. et al.) 2603–2613 (ACM, 2022).
193. Rathje, S., He, J. K., Roozenbeek, J., van Bavel, J. J. & van der Linden, S. Social media behavior is associated with vaccine hesitancy. *PNAS Nexus* **1**, pgac207 (2022).

194. Canché, M. S. G. Machine driven classification of open-ended responses (MDCOR): an analytic framework and no-code, free software application to classify longitudinal and cross-sectional text responses in survey and social media research. *Expert. Syst. Appl.* **215**, 119265 (2023).
195. Hartmann, J., Bergner, A. & Hildebrand, C. MindMiner: uncovering linguistic markers of mind perception as a new lens to understand consumer-smart object relationships. *J. Consum. Psychol.* **33**, 645–667 (2023).

Author contributions

S.F., A.M., D.B., D.G., J.S. and N.P. outlined the article and wrote the first draft. A.M. created the first draft of the figures. All authors contributed to subsequent iterations of the article. All authors reviewed, edited and approved the manuscript before submission.

Competing interests

The authors declare no competing interests.

Additional information

Peer review information *Nature Reviews Psychology* thanks April Bailey, Morteza Dehghani and Dirk Wulff for their contribution to the peer review of this work.

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© Springer Nature America, Inc. 2025