sociological science

Using Machine Learning to Uncover the Semantics of Concepts: How Well Do Typicality Measures Extracted from a BERT Text Classifier Match Human Judgments of Genre Typicality?

Gaël Le Mens,[a,b,c] Balázs Kovács,[d] Michael T. Hannan,[e] Guillem Pros[a]

a) Universitat Pompeu Fabra (UPF); b) Barcelona School of Economics; c) UPF Barcelona School of Management;

d) Yale University; e) Stanford University

Abstract: Social scientists have long been interested in understanding the extent to which the typicalities of an object in concepts relate to its valuations by social actors. Answering this question has proven to be challenging because precise measurement requires a feature-based description of objects. Yet, such descriptions are frequently unavailable. In this article, we introduce a method to measure typicality based on text data. Our approach involves training a deep-learning text classifier based on the BERT language representation and defining the typicality of an object in a concept in terms of the categorization probability produced by the trained classifier. Model training allows for the construction of a feature space adapted to the categorization task and of a mapping between feature combination and typicality that gives more weight to feature dimensions that matter more for categorization. We validate the approach by comparing the BERT-based typicality measure of book descriptions in literary genres with average human typicality ratings. The obtained correlation is higher than 0.85. Comparisons with other typicality measures used in prior research show that our BERT-based measure better reflects human typicality judgments.

Keywords: categories; concepts; deep learning; typicality; BERT; transformer models

Citation: Le Mens, Gaël, Balázs Kovács, Michael T. Hannan, and Guillem Pros. 2023. "Using Machine Learning to Uncover the Semantics of Concepts: How Well Do Typicality Measures Extracted from a BERT Text Classifier Match Human Judgments of Genre Typicality?" Sociological Science 10: 82-117.

Received: September 28, 2022

Accepted: November 9, 2022

Published: March 3, 2023

Editor(s): Ari Adut, Filiz Garip

DOI: 10.15195/v10.a3

THIS article addresses issues of measurement in sociological research that builds on cognitive processes. We propose a way to use state-of-the-art natural language processing to measure aspects of categorization. Categorization decisions concern distinguishing what is and what is not an instance of some mental representation, such as a concept or schema. These issues arise routinely in sociological work on culture and economic organization.

Cognitive anthropology, initially a founding discipline of cognitive science, provided an early template for making explicit links between culture and cognition in work done primarily in the 1950s and 60s (D'Andrade 1995; Bender, Hitchins, and Medin 2010). Interest in these issues has waned in anthropology (Beller, Bender, and Medin 2012); sociologists, however, following the lead of Paul DiMaggio (1997), have taken up the challenge. Recent years have seen a flurry of activity seeking to exploit notions from cognitive science (principally cognitive psychology) in cultural analysis (for reviews, see Cerulo, Leschziner, and Sheperd [2021] and Vaisey [2021]).

A similar development has taken place in the study of organizations and markets. In this case, the focus was on how agents acting as audience members judge the offers of producers. A crucial part of the evaluation process entails categoriz-

ing the producers/products (Porac et al. 1995; Zuckerman 1999; Hannan, Pólos, and Carroll 2007; Hannan 2010). In other words, concepts such as industry and genre serve as the basis for audience expectations, and categorizations tell which producers/products deserve attention.

Unlike efforts like *Measuring Culture* (Mohr et al. 2020) that attempt to deal with measurement of the realm of culture as a whole, our aims are narrower. We limit the scope of methods to modern natural language processing. A second limit on scope is theoretical. The sociological research of interest generally uncovers issues of *typicality*, a measure of the *degree* to which an object/agent/situation exemplifies the focal concept. The possible advantage of such narrowing is that it makes it feasible to provide specific advice about measurement. This choice of focus also facilitates making explicit connections with cognitive science/psychology, because since the work of Rosch in the early 1970s, exploring typicality has been a strong focus in those disciplines.

A number of natural language processing (NLP) techniques have proven useful to analyze sociological processes. Latent semantic analysis has been used to study the structure of the healthcare sector (Ruef 2000), topic modeling has been used to study newspaper coverage of U.S. government arts funding (DiMaggio, Nag, and Blei 2013), and word embeddings have been used to study how the markers of social class have shifted over time (Kozlowski, Taddy, and Evans 2019). In recent years, the performance of NLP techniques has experienced a qualitative jump with the advent of the "transformer models" class of language representation (Vaswani et al. 2017).

NLP based on deep learning and transformer models far outperforms prior approaches such as content analysis, bag-of-words representations, topic modeling, or word embeddings. A dramatic breakthrough occurred in 2018 with the public release of the BERT (Bidirectional Encoder Representations from Transformers) language representation (Devlin et al. 2018). At the time of this writing, this model is used to interpret Google search queries in more than 70 languages (see announcement on Twitter: https://twitter.com/searchliaison/status/1204152378292867074?s=20), and it is approaching human-level performance in a number of natural language understanding tasks (Nangia and Bowman 2019). Moreover, virtually all subsequent state-of-the-art language models have been based on BERT (see https://gluebenchmark.com/leaderboard).[1]

Despite the impressive performance of models based on BERT (and related language representations) in solving language understanding tasks and reasoning problems, we lack direct evidence that these techniques can be used to produce typicality measures that parallel human judgements. In a tour-de-force analysis, Bhatia and Richie (2022) demonstrated that BERT can reproduce human judgment patterns obtained in a wide variety of previous studies of semantic structures. Generally, these take the form of patterns of agreement/disagreement of "is-a" statements relating subconcepts to concepts, for example, "a penguin is a bird." This research gives further confidence that sociologists can profitably employ BERT in analyzing culture and markets. However, we still do not have direct evidence that the *typicality* of *objects* produced by a BERT-based model (e.g., a particular artwork)

resembles human typicality judgments. Addressing this shortcoming of current knowledge is the main focus of this article.

Our focus on *objects* contrasts with the focus of recent work that used word embeddings to measure semantic associations (Garg et al. 2018; Kozlowski et al. 2019; Lewis and Lupyan 2020). This work measured the association between concepts (an occupation [e.g., teacher] and a gender [e.g., female]—Garg et al. [2018]) or associations between dimensions of concepts (e.g., affluence and education—Kozlowski et al. [2019]), but not the typicality of a particular object in a concept (e.g., the typicality of a worker in the teacher concept).

What sets apart our approach from earlier approaches to typicality measurement concerns the nature of the data used to construct typicality measures. First, we construct typicality from textual descriptions of objects. Prior work generally did not analyze feature values of objects, only categorizations (Hsu 2006; Hsu, Hannan, and Koçak 2009; Pontikes and Hannan 2014; Kovács and Hannan 2015; Pontikes 2022) (but see Kovács and Johnson 2014). This is a severe limitation because, due to lack of better information, this work assumes zero probability of categorization in all unassigned concepts, and thus minimal typicality in these concepts. For example, Kovács and Hannan (2010) studied categorizations of restaurants, and they assumed that a restaurant that was classified as French and Japanese would have zero (minimal) typicality in all other cuisines such as Mexican or Californian. Second, our approach is also applicable in empirical settings in which objects can have at most one label. Typicality measures that do not rely on features but only on categorization would produce only two levels of typicality in such settings, rendering typicality measures discrete, just like categorizations. This is inconsistent with the definition of typicality as a *graded* construct (a measure of the *degree* to which an object/agent/situation exemplifies the focal concept). Relying on deep-learning NLP allows for fine-grained measurement of typicality because the text data it uses are less coarse than the categorical assignments used in prior research.

Prior work has measured similarity between objects, represented as vectors of feature values, using a similarity function such as cosine similarity or Jaccard similarity. A simple way to construct a typicality measure based on such an approach starts by defining the position of (the center of) a concept in feature space as the average position of objects categorized as instances of this concept and then defines the typicality of an object in the concept in terms of the similarity between the object and the center of the concept (e.g., Smith 2011; Pontikes and Hannan 2014; Durand and Kremp 2016).[2] A related approach computes the similarity between the object and known instances of the concept and then takes the average as the typicality measure.[3] These two approaches implicitly give the same weight to all feature dimensions. By contrast, our approach gives more weight to features that matter more for categorization and typicality judgments. This is because our approach constructs a feature space optimized for categorization performance. In comparisons of the fit of competing typicality measures with human typicality ratings, we will see that this characteristic of our approach is key to its superior ability to reflect human typicality ratings.

Our work also differs from articles that used BERT classifiers to label large quantities of text data (more than humanly possible; see Bonikowski, Luo, and Stuhler

[2022] and Schöll, Gallego, and Le Mens [2023] for recent examples). Whereas this work has used discrete predictions by machine learning classifiers, we use the continuous predictions of such models (i.e., the predicted categorization probabilities) to construct a *graded* measure of the extent to which an object exemplifies the focal concept.

The article is organized as follows. In the section Concepts, Categories, and Typicality, we sketch the theoretical background needed to motivate our approach. The core idea is that categorization, the act of applying a concept to an object, can be seen as probabilistic inference. An agent observes the features of an object and uses them (along with prior beliefs) to infer the probability that the object is an instance of the concept. We define the typicality of an object in a concept in terms of such categorization probability. We will say that an object is typical of a concept if, given its features, it is likely to be an instance of this concept. The empirical challenge to measuring typicality then becomes a challenge to measuring categorization probabilities.

In Using a Probabilistic Classifier to Measure the Typicality of Objects in a Concept, we explain how a standard class of machine-learning models, probabilistic classifiers, can contribute to solving this challenge when a researcher has access to feature-based descriptions of objects and their categorizations. This approach is only applicable when a feature space is available to represent objects and concepts. This is not the case when objects consist of text documents. We address this challenge in the following section.

Then, in Measuring the Typicality of Text Documents with a BERT Probabilistic Classifier, we explain how deep-learning NLP can be used to construct a feature-based description of text documents and produce categorization probabilities that are, in turn, used to construct the typicality of each document in the focal concept. Our text-categorization model uses the BERT language representation to construct a feature space adapted to categorization in the focal concept. We call the resulting measure *BERT typicality*.

In Validation of the *BERT Typicality* Measure, we apply our approach to a particular empirical setting: we measure the typicality of books with respect to literary genres from analysis of book descriptions from Goodreads.com. We show that the *BERT typicality* is highly correlated with human typicality ratings, providing validation that this measure could be used as a substitute for human typicality ratings when these are difficult or impossible to obtain directly.

Finally, in Benchmarking: Comparing BERT Typicalities with Typicalities Obtained with Other Probabilistic Classifiers or with Label Assignments, we compare the *BERT typicality* with other model-based typicality measures that rely on other language representations such as GloVe (Global Vectors for Word Representation) word embeddings or bag-of-words representations and typicality measures produced by techniques that do not rely on training a probabilistic classifier (e.g., cosine similarity in pre-trained embedding space) and approaches that rely on sets of labels given to objects. We will see that the *BERT typicality* reflects human typicality ratings better than other approaches. We attribute this performance to a combination of two factors: the construction of a feature space adapted to categorization in the focal concept, and the definition of typicality in terms of categorization probabilities

produced by a probabilistic classifier that gives more weight to the features that matter more for categorization. Arguably, the method we advance in this article allows the construction of a feature space and a typicality function that jointly provide a mathematical representation of concepts that reflects humans' mental representation of concepts. This is why we claim in the title that our approach "uncovers the semantics of concepts."

Concepts, Categories, and Typicality

The contemporary view of concepts on which we build sees them as mental representations with no clear boundaries (Anderson 1991; Ashby and Alfonso-Reese 1995; Feldman, Griffiths, and Morgan 2009; Hannan 2010; Sanborn, Griffiths, and Shiffrin 2010). Consequently, there is often vagueness in judgments about which concepts, if any, apply to an object. Psychologists characterize the extent to which an object fits a concept as the typicality of the object in the concept. Nearly 50 years of research, initiated by Rosch (1973), have shown that concepts are structured by typicality. For instance, apple is a highly typical fruit, grape is moderately typical, and tomato is highly atypical. Recent research has shown that typicality affects valuation; people generally place more value on more typical objects (e.g., Vogel et al. 2018).

In this section, we provide a formal definition of typicality that we will use to construct the empirical measure of typicality we propose in the next two sections. We begin with definitions of concepts and categorization probabilities.

Concepts

Following the modern approach to concepts in cognitive psychology, we model concepts as probability distributions over a *feature space*—a space of feature values in which the meaning of a concept is expressed. Its dimensions are the features that a focal person uses in forming a mental representation of the concept. When we turn to thinking about the categorization of objects, then this is also the space for the mental representations of the objects. Each object is represented as a position in this feature space—a particular combination of values of the relevant features. We denote the focal agent's feature space by $\mathbb{G}$.

Concepts specify which positions in feature space are more likely than others for objects that "belong" to a concept. The key formal notion is concept likelihood, $\pi_{\mathbb{G}}(x|c)$, which gives the subjective probability (or belief) that an object known to be an instance of the concept c has some particular combination of values of relevant features (is at position x in the feature space $\mathbb{G}$).

Categorization Probabilities

We define categorization as the act of applying a concept to an object. We model categorization in a probabilistic way. We denote the probability that an agent who perceives an object to be at position x in feature space categorizes it as a c by $P(c \mid x)$.

The Bayesian approach to categorization (on which we build) holds that the categorization probability is a function of the concept likelihood, the prior belief on

position, and the prior belief that the object is a c as follows:

$$P(c \mid x) = \pi_{\mathbb{G}}(x|c)\frac{P(c)}{P_{\mathbb{G}}(x)}, \tag{1}$$

where $P(c)$ denotes the subjective probability that an object is a c based on background information about the categorization context, but without any information about the position of the object in feature space, and $P_{\mathbb{G}}(x)$ denotes the subjective probability that an object is at position x in feature space $\mathbb{G}$ if its category is not known.

Typicality

As we mentioned in the introductory section, Rosch (1973) proposed that concepts have an internal structure that can be represented in terms of typicality as goodness of representation of a concept. Despite its importance, typicality has been treated largely as a primitive notion, and researchers have generally measured typicality by asking people to tell how "typical" of some concept are each of a set of subconcepts (apples and fruit, for instance). Here we focus on the typicality of individual *objects* (a particular apple) (Vogel et al. 2018; Hannan et al. 2019), rather than of subconcepts.

Suppose that an object has a set of features x relevant for categorization in the focal concept. In a setting where objects are text descriptions, x could be a sequence of words. In a setting where objects are images, x could be the red, green, and blue luminance values for all the pixels that form an image. In a choice between customer products, x could be a feature of technical specifications. Hannan et al. (2019) employ the intuition that a position is highly typical for a concept if the concept likelihood is high. Here we deploy a slightly different intuition, that an object is highly typical of a concept if its features make it a very likely member of this concept—if $P(c \mid x)$ is high. In particular, we expect the feature combination x to be all the more typical of the concept if it increases the probability of c significantly above the baseline value, $P(c)$, that is, if $P(c \mid x)$ is greater than $P(c)$. In this article, we build on this intuition and define the typicality of an object with features x as follows:[4]

$$\tau_c(x) \equiv \log \frac{P(c \mid x)}{P(c)}, \tag{2}$$

where $P(c)$ is the prior on membership in the concept c, the subjective probability that an object taken at random in the domain will be an instance of c. As is common in Bayesian models of categorization, we assume that the prior is given by the empirical proportion of objects in the domain that are cs.

We think that the proposed formulation in Equation (2) provides a more intuitive rendering of typicality than defining typicality as the concept likelihood or its logarithm (as in Hannan et al. [2019]). The following stylized example provides an illustration. The context is that of restaurants in Germany. The feature space contains just one binary-valued dimension such that $x = \textit{vegetarian}$ if the focal restaurant is entirely vegetarian, offering no meat item on the menu, and $x = \textit{non-vegetarian}$ otherwise. The focal concept is Indian restaurant. A small proportion of all restaurants are Indian such that $P(\text{Indian}) = 0.05$. Most Indian

restaurants have some meat items. Yet 30 percent of them are entirely vegetarian such that $P(\textit{vegetarian} \mid \textsf{Indian}) = 0.3$. Now consider all restaurants in Germany: a small proportion of them are *vegetarian*, whereas most offer some meat item on their menus, that is, $P(\textit{vegetarian}) = 0.1$. With these numbers, Bayes' rule implies $P(\textsf{Indian} \mid \textit{vegetarian}) = 0.15$ and $P(\textsf{Indian} \mid \textit{non-vegetarian}) = 0.04$. Even though most *vegetarian* restaurants are not members of the Indian restaurant concept, knowing that a restaurant is *vegetarian* makes it more than three times more likely to be an Indian restaurant, and knowing that a restaurant is *non-vegetarian* makes it slightly less likely to be an Indian restaurant. Consistent with this pattern, the typicality of the *vegetarian* position in the Indian restaurant concept ($\log(0.15/0.05) = 1.1$) is higher than the typicality of the $x = \textit{non-vegetarian}$ position ($\log(0.04/0.05) = -0.25$).

By contrast, the definition of typicality as the concept likelihood (used in Hannan et al. [2019]) implies the opposite ranking of typicality values. Because $P(\textit{vegetarian} \mid \textsf{Indian}) = 0.3$ and $P(\textit{non-vegetarian} \mid \textsf{Indian}) = 0.7$, the typicality of the *vegetarian* position in the Indian concept would be lower than the typicality of the *non-vegetarian* position. This ranking seems to clash with intuition.

With the definition of typicality in terms of categorization probability, the main empirical challenge in measuring the typicality of an object in a concept pertains to estimating the categorization probability of this object in the focal concept. In research on cognitive psychology, categorization probabilities are treated as latent psychological variables that depend on agents' concepts and their perceptions of objects' positions in the relevant feature space. Unless the agents are directly asked to provide categorization probability judgments, these quantities are not observable, as is the case with archival data. Of course, sociologists could follow psychologists in asking agents to provide typicality judgments about the objects of interest, eliminating the need for estimating categorization probabilities. This is unfeasible for the analysis of archival data and or very large data sets.

Using a Probabilistic Classifier to Measure the Typicality of Objects in a Concept

We consider a setting in which a researcher wants to measure the typicality of an object o in a concept c. The researcher has access to categorization data $\mathcal{D}$ of N objects in concept c. The feature space used to represent objects has H dimensions. Each observation in the categorization data consists of the vector of feature values of the object $x = (x_1, \ldots, x_H)$ and a dummy variable that takes a value of 1 if the object has been categorized as a c or a value of 0 otherwise.

A probabilistic classifier is a function f_c that, given a vector of feature values x, returns the probability that an object represented by vector x is an instance of concept c:

$$f_c(x) = P(c \mid x). \tag{3}$$

The central proposition of this article is that the researcher can produce typicality measures that reflect human typicality judgments from the categorization probabilities produced by a machine-learning "probabilistic classifier" constructed from the data.[5]

According to this conjecture, an analyst who has access to a probabilistic classifier can measure the typicality of objects in concepts by applying Equation (2). We call such a typicality measure PC typicality. The challenge of typicality measurement thus becomes a challenge of constructing a probabilistic classifier from the available categorization data.

PC typicalities will reflect human typicality judgments if the probabilistic classifier on which they are built is sensitive to the same feature combination as humans who judge typicality. And if the feature combinations that best explain categorization in the input data are the same as those that explain human typicality judgments, the goal becomes one of identifying the feature combinations that capture categorization in the data. The field of machine learning has developed a robust methodology for this purpose. The procedure proceeds in several stages and relies on three disjoint data sets: training set, validation set, and prediction set. The training set and validation set are subsets of the available categorization data $\mathcal{D}$.

The first stage consists of specifying the probabilistic classifier f_c as a function whose outputs depends on the input (the vector x of feature values) and a set of model parameters. A simple example likely familiar to most readers consists of a logistic regression model that returns the logistic transformation of a linear combination of the feature values. The model parameters here are the regression weights.

The second stage consists of using the categorization data to find the best fitting parameters. In the field of machine learning, this is called "model training" and is achieved by minimizing a *loss function* using numerical optimization routines. A frequently used loss function for training classifiers is the opposite of the log-likelihood of the data (called "categorical cross-entropy"). In this case, the loss associated to an observation at position x can take two values depending on the ground truth. If the object is an instance of the concept, the loss is equal to $-\log P(c|x)$. If the object is not an instance of the concept, the loss is equal to $-\log(1 - P(c|x))$. With the categorical cross-entropy loss function, model training is the same as what is called maximum-likelihood estimation in econometrics and statistics. Model training thus requires some input data in the form of a table of feature values (X_{train}: a table of N_{train} rows and H columns) and some ground truth categorization data that indicate whether each observation belongs to the focal concept c (Y_{train}: a vector of N_{train} rows and populated with 0s and 1s).

The numerical optimization routines used to minimize the loss function on the training data frequently have some so-called training parameters that have to be set manually by the researcher (e.g., learning rates, step size, stopping criterion, ...). Moreover, the classifier might also have some other parameters that are set manually (e.g., "number of hidden nodes") before launching the loss minimization routine. Training the model also generally encompasses finding the best combination of such manually set model parameters. Because machine-learning models frequently have many parameters (several millions in the case of BERT classifiers), there is always a risk of overfitting the model to the training data, meaning that the model will capture some pattern in the training data that does not exist in the prediction data. Overfitting hurts generalization performance and thus the quality of model predictions on data not included in the training set.

The machine-learning approach for dealing with this issue consists of evaluating the performance of trained models on the validation set. This is the third stage. It requires that the validation set has the same structure as the training set: an input table that contains the feature values for each observation in the validation data (X_{val}: a table of N_{val} rows and H columns) and the ground truth categorization data (Y_{val}: a vector of N_{val} rows populated with 0s and 1s). A standard approach in constructing the training set and the validation set consists of randomly splitting the input data $\mathcal{D}$ into these two sets (e.g., 95 percent of the data go to the training set and five percent to the validation set).[6]

The objective of training is to produce a model that achieves maximal performance when applied to the validation set (i.e., minimizing the validation loss). This is achieved by looping over the training and validation stages until validation performance cannot be further improved. The validation loss generally goes down with the amount of model training and at some point starts to go up while the training loss keeps going down. This is a signal that at this point the model starts to overfit the data: it identifies patterns in the training data that are not present in the validation data. This harms the model's generalization performance. Therefore, we stop training at the point at which the validation loss begins to increase.

Finally, the trained model is applied on the prediction set. The prediction data must include vectors of feature values for each observation (X_{pred}: a table of N_{pred} rows and H columns), but it need not contain ground truth categorization data. This is one of the advantages of the approach set forth in this article, as compared with the frequently used label-based approach described in Comparison with Label-Based Approaches to Measuring Typicality. For each vector of feature values $x \in X_{\text{pred}}$, the trained model returns a categorization probability in the focal concept c: $P(c \mid x)$. This categorization probability is then used to construct the PC typicality using Equation (2).

Measuring the Typicality of Text Documents with a BERT Probabilistic Classifier

In the previous section, we assumed that a feature space was available and that objects were represented as vectors in this space. Text documents consist of sequences of words, and their representation in computer code does not generally correspond to a vector of feature values. In this section, we explain how deep learning not only can transform text documents into vectors of feature values but also automatically constructs a feature space optimized for classification performance.

A distinctive characteristic of the deep-learning approach we advocate is that it constructs a feature space especially adapted to the categorization of text documents in the focal concept c through training of a probabilistic classifier based on the input data $\mathcal{D}$ specified in the previous section. This classifier is made of two distinct but interacting components:

1. A *representation* component that takes text documents and represents them as points in a feature space $\mathbb{H} = \mathbb{R}^H$ (where $\mathbb{R}$ denotes real numbers). This component is an artificial neural network that consists of a set of functions

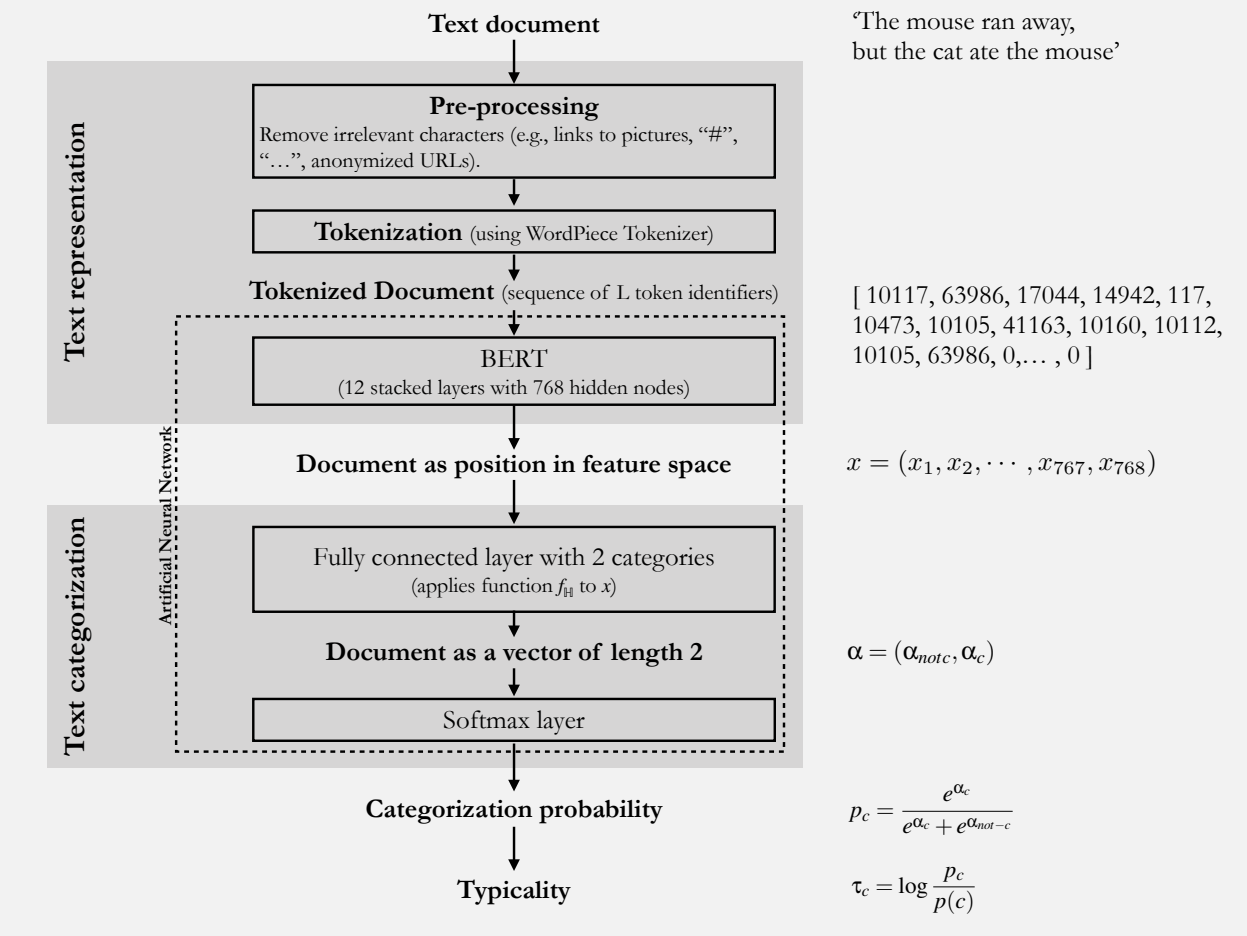

Figure 1: Structure of the BERT probabilistic classifier.

that operate in sequence on the inputs and are often represented in terms of a vertical stack of linear functions (layers) with some nonlinear intermediary steps (activation functions).

2. A *categorization* component that takes positions in the feature space as inputs and produces a vector of categorization probabilities as outputs. This component can be as simple as a logistic regression model.

Figure 1 summarizes our probabilistic classifier. The representation component involves the BERT model (Devlin et al. 2018). Thus, we call this classifier a "BERT probabilistic classifier." It takes as input a text document and returns the probability of categorization in the focal concept p_c, which is then transformed in a typicality measure: the *BERT typicality*. Next, we describe the *representation* and *categorization* components of the model.

Representation Component: BERT

The representation component of the BERT probabilistic classifier is made of two sub-components: text preparation and the BERT model itself.

Text preparation. Text documents need to be represented in a numerical format to be used as inputs to the BERT model. Figure 1 shows the standard processing operations used in our empirical applications. There is an optional pre-processing stage that removes parts of the document deemed irrelevant by the analyst. The indispensable component of the text preparation consists of tokenization, as described in Text Tokenization in Appendix: Methodological Details. Similar operations are frequently used to prepare inputs to other machine-learning techniques that take text as input. The output of the text preparation stage applied to a text document is called a tokenized document. This consists of an L-long sequence of indices where L is a parameter that characterizes the maximal length of text documents that can be processed by the model (in terms of number of tokens). Documents that contain more than L tokens are truncated.[7]

BERT model. BERT consists of an artificial neural network with many layers (it is a deep neural network) that takes a sequence of tokens as input and outputs a 768-dimension vector of real values that represents the position of a text document in feature space $\mathbb{H} = \mathbb{R}^{768}$: $x = (x_1, x_2, \ldots, x_{767}, x_{768})$. This number of dimensions (768) was not chosen by the authors of the present article but instead is a characteristic of the pre-trained model we used (BERT-base-cased). The BERT model is made of a stack of 12 transformer layers (Vaswani et al. 2017). Discussion of the internal structure of the BERT block goes beyond the scope of this article, and we refer interested readers to the original paper for formal details (Devlin et al. 2018).

A distinctive advantage of applying BERT to a text is that this produces a representation sensitive to the sequence of words in the entire text. This goes beyond models that rely on a bag-of-words approach, for example, the naive Bayes classifier (Maron 1961). Even though the sensitivity to word sequences is noteworthy, it is not unique to BERT but is shared with previous models such as deep-learning categorization models based on a long short-term memory (LSTM) layer (Hochreiter and Schmidhuber 1997).[8] The crucial innovation is that BERT constructs word representations that are contextual: the mathematical representation of a word depends on the words that come before and after the focal word. The model is thus sensitive to the fact that the meaning of a word depends on the words that come before and after it (possibly long before and longer after). It is widely accepted that this ability to capture bidirectional dependency in word meaning is one of the factors that make BERT perform so well.[9]

Categorization Component: Probabilistic Text Classification

This is implemented in the neural network by means of two layers: a fully connected layer and a softmax layer.

Fully connected layer. This layer takes the position of the text document in the feature space $\mathbb{H}$ and outputs a pair of real values, $\alpha = (\alpha_{not\text{-}c}, \alpha_c)$, where $\alpha_{not\text{-}c}$ and α_c are linear combinations of the inputs $(x_1, \ldots, x_{768})$. We denote by $f^c_{\mathbb{H}}(x)$ the function that returns α_c. It has $768 + 1$ parameters: one "bias" parameter

(constant term) and one coefficient for each of 768 dimensions in the input. This layer characterizes the similarity of a text document, represented as a position in space $\mathbb{H}$, to concept c.

Softmax layer. This applies the following softmax function to the pair of category scores $(\alpha_{not\text{-}c}, \alpha_c)$ and outputs a vector of categorization probabilities. Specifically,

$$p_c = \frac{e^{\alpha_c}}{e^{\alpha_c} + e^{\alpha_{not\text{-}c}}}, \quad p_{not\text{-}c} = 1 - p_c.$$

The combination of the fully connected layer and the softmax layer specifies a logit model. To make explicit the dependence of categorization probabilities on positions in the feature space $x \in \mathbb{H}$, we rewrite the categorization probabilities as follows:

$$p_c(x) = \frac{1}{1 + e^{-(f^c_{\mathbb{H}}(x) - f^{not\text{-}c}_{\mathbb{H}}(x))}}.$$

BERT Typicality

The *BERT typicality* of a text document *text* in concept c is computed from Equation (2) by inserting BERT's estimate of $p(c \mid x)$:

$$\text{typ}_c(text) = \tau_c(x) = \log \frac{p_c(x)}{p(c)}, \tag{4}$$

where $p(c)$ is the proportion of text documents from the training data in c.

Adjusting Model Parameters Using Data: Model Training

The model is trained on categorization data following the procedure described in Using a Probabilistic Classifier to Measure the Typicality of Objects in a Concept. A categorical cross-entropy loss function is used. It takes the average of the log of the predicted probability for the ground truth label. Model training aims to optimize model parameters to minimize the loss value computed over the validation set.

The process of training the model is called deep learning. "Deep" in this context refers to the fact that the artificial neural network that makes the representation component has many layers; "learning" means that the many weights of the linear functions (often several millions) of this model component are learned from the data in the training stage. When training the model, not only the parameters of the categorization component but also the parameters of the representation component are adjusted so as to minimize the loss (and thus maximize classification performance). Therefore the representation of text documents is adapted to maximize classification performance. If the task were different (e.g., the set of candidate categories changes), then the trained representation component would be different, and text documents would be represented by different vectors of feature values. The training process essentially constructs features that are useful for categorization and ensures that the predicted category is sensitive to these features. Model training occurs via sophisticated numerical optimization algorithms that have been designed to efficiently process extremely large quantities of data (e.g., millions of text documents, images, or voice recordings).

A distinctive advantage of using BERT to represent text documents is that several BERT models that have been pre-trained on vast amounts of text (hundreds of gigabytes) to learn a generic language representation are publicly available and free for researchers to use. Such pre-trained language representation can then be fine-tuned for specific tasks like categorization, question answering, text generation, or translation. Pre-training does not involve categorization in the focal concept c but another task that has been chosen by the creators of the BERT model because it allows the model to learn language regularities that are useful for a variety of tasks, such as categorization, but also question answering, translation, entity recognition, et cetera.[10] The BERT model we use (BERT-base-cased) has been trained on a large corpus of English texts.[11]

Most prior text-categorization models based on machine-learning techniques are trained from scratch on a particular data set. This is the case for bag-of-words–based approaches, such as naive Bayesian categorization models (Maron 1961). This is also the case for more sophisticated deep-learning models sensitive to word sequences. The basic approach to training such categorization models uses only information from the data at hand to learn all the model parameters. If the training data set is of limited size, performance will suffer.

The process that consists of *fine-tuning* a pre-trained model allows for high performance on specific tasks even if the training data set is of limited size. Fine-tuning consists of updating the parameters of the language representation component of the model (BERT) as well as the parameters of the categorization component using the data for the task at hand (categorization data). The combination of pre-training and fine-tuning allows the model to learn general language regularities from vast amounts of data while at the same time adapting the language representation to a specific application using the data of the particular study. This aspect of the approach is particularly germane to research in the social sciences because it frequently focuses on settings with domain-specific, idiosyncratic languages. Later in the article, we compare the ability of fine-tuned and non–fine-tuned language representations to reflect human typicality judgments.

Next, we apply the approach presented in this section to the computation of the *BERT typicality* of book descriptions in certain literary genres (Mystery, Romance), based on training data that consist of binary categorizations.

Validation of the *BERT Typicality* Measure

In this section, we compare the *BERT typicality* of book descriptions in two literary genres (Mystery and Romance) with typicality ratings provided by human participants.

Data: Labels and Book Descriptions from Goodreads.com

We obtained book labeling data from Goodreads.com, the largest user-contributed book review website, covering more than a million books. We used the 2018 version of the Goodreads database, made publicly available by the University of California San Diego (https://sites.google.com/eng.ucsd.edu/ucsdbookgraph/home; Wan and McAuley 2018; Wan et al. 2019). We analyze all English-language books

that have a short description in English of up to 300 words and that have been labeled by readers as one of 36 genre labels. Our sample contains 738,451 books.[12]

The text descriptions were generally taken from the cover-jacket text. Originally, when Goodreads.com started in 2007, the description texts were uploaded by the authors/publishers themselves. But since 2013, when Amazon bought Goodreads, short descriptions of books have been pulled from Amazon.com's description. See Appendix: Goodreads.com Data for an example of a book's short description.

The book labeling at Goodreads.com is outsourced to Goodreads users (i.e., book readers). Readers can tag books with any labels and put books on (virtual) named shelves.[13] Although there is no predefined set of labels (e.g., no drop-down menu or autocomplete), readers tend to use genre labels for shelving, although not exclusively. Shelves such as "to-read" or "read-in-school" are also common. In this article, we focus on the genre labels. Specifically, we use the 36 main genre labels as listed on the Goodreads search page (`https://www.goodreads.com/genres?ref=nav_brws_genres`). These include labels such as Sports, Fantasy, Suspense, Travel, Humor and Comedy, Mystery, and Memoir (see Appendix: Goodreads.com Data for the full list of labels, along with their frequency). Importantly, this set of labels defines a cohort of concepts as defined above: there is no hierarchical embedding among labels. Together these labels cover most English-language books in the Goodreads data (93 percent).

The available data provide for each book the distribution of genre assignments—a vector of proportions. Most of our analysis assesses the extent to which typicality measures based on the predictions by a BERT-based probabilistic classifier trained on binary categorization data match human typicality judgments. For this part of the analysis, we collapse the data to create a binary distinction that associates each book with its most commonly assigned genre. In the section Comparison with Label-Based Approaches to Measuring Typicality we use the full set of proportions of assignments. See Table A in Appendix: Goodreads.com Data for the proportion of each genre.

Because collecting human typicality judgments on 36 genres would require a very large number of participants, we decided to focus on two genres: Mystery and Romance. We chose them because they are two of the most popular genres in our data, and we expected that people who read books would be familiar enough with them to be able to provide typicality judgments of books in these genres. For both genres, we created training, validation, and prediction data sets. In each case, the prediction sets consist of 500 book descriptions of Mystery books and 500 descriptions of other books. These were randomly selected from the sample, stratifying by length of book description. The remaining observations were split into a validation set ($N_{\text{val}} \sim 50,000$) and a training set ($N_{\text{train}} \sim 680,000$).

Training the BERT Classifier

We trained separate BERT classifiers for Mystery and Romance using the approach presented in the sections Using a Probabilistic Classifier to Measure the Typicality of Objects in a Concept and Measuring the Typicality of Text Documents with a BERT Probabilistic Classifier. Model training does not aim to maximize categorization

Table 1: Confusion matrices for the trained BERT classifiers on the validation sets

		Predicted Mystery		
		No	Yes	Sum
Mystery (ground truth)	No	0.93	0.01	0.94
	Yes	0.02	0.04	0.06
	Sum	0.94	0.06	1.00

		Predicted Romance		
		No	Yes	Sum
Romance (ground truth)	No	0.84	0.04	0.88
	Yes	0.03	0.09	0.12
	Sum	0.87	0.13	1.00

Notes: Left: for Mystery ($N = 47,699$). *Right:* for Romance ($N = 47,697$).

Table 2: Categorization performance of the BERT classifiers on the validation sets

	BERT	
	Mystery	Romance
Accuracy	0.97	0.93
Precision	0.74	0.70
Recall	0.68	0.77
F1 score	0.71	0.73
Mean loss per observation	0.09	0.16

Notes: "Precision" is the percentage, out of all the objects predicted to be in the focal category, that actually are in this category. "Recall" is the percentage, out of all the objects that are in the focal category, that are predicted to be in this category. F1 score is the harmonic mean of precision and recall. Mean loss per observation is the average per observation loss, computed using the categorical cross-entropy loss function.

accuracy but instead minimizes the cross-entropy categorization loss. The difference between these two criteria is that the cross-entropy categorization loss gives a large penalty to big mistakes (e.g., the model gives a low probability of being a Mystery book to a book that is actually a Mystery book). Yet, to get an intuitive sense of the categorization performance of the trained model, it is useful to examine categorization accuracy. The trained classifiers were very accurate, reaching classification accuracy on the validation sets of 0.97 (Mystery) and 0.93 (Romance).[14]

Note that, because the proportion of books in any focal genre is relatively small, a simplistic strategy that would categorize all books as not belonging to the focal genre would achieve a high accuracy. To account for this, we use other metrics, such as recall and precision, which can be computed based on the confusion matrices (see Tables 1 and 2). Overall, the categorization performance of the model is excellent for both genres. We invite interested readers to download the "compute_typicality" folder from the project's Open Science Framework page and experiment with model training and predictions.[15]

Constructing BERT Typicality Measures on the Prediction Set

We computed the *BERT typicality* in the Mystery genre for each book description in the prediction set by using the formula in Equation (4) applied to the categorization probability in the Mystery genre produced by the trained BERT classifier. We did the same for the book descriptions in the prediction set used for the Romance genre.

Typicality Ratings by Human Participants

For typicality ratings in the Mystery genre, we split the prediction set of 1,000 books (500 Mystery books and 500 other books) in 50 subsets of 20 books (each with 10 Mystery books). Four hundred ninety-seven Prolific participants rated the typicality of a subset of 20 book descriptions in the Mystery genre. For each book description, they responded to the question "How typical is this book to the mystery genre?" using a 0 to 100 slider (centered at 50 when the page appears on the screen). Each book excerpt received about 10 typicality ratings (with a minimum of seven and maximum of 12).[16] For each book, we computed an average of the typicality ratings across the participants who rated it. We call this quantity the *human typicality*.

We followed the same procedure for the 1,000 books of the prediction set used for Romance. Five hundred two Prolific participants provided typicality ratings for 20 book descriptions in the Romance genre.

Validation: Comparing BERT Typicality with Human Typicality Ratings

The *BERT typicality* is highly correlated with the *human typicality*, at 0.87 for the Mystery genre and 0.86 for the Romance genre. We see this performance as excellent and in any case good enough to use *BERT typicality* measures as substitutes for human typicality ratings in empirical studies that require measuring the typicality of objects into concepts.

A skeptical reader might wonder if the high correlation between *BERT typicality* and *human typicality* is directly implied by the excellent categorization performance of the BERT classifier, or if the *BERT typicality* reflects something more than correct classification. If the correlation was mostly driven by the model's making binary or quasi-binary predictions (either very high or very low typicalities, with few in-between predictions), then the model could perform well according to this metric but would fail to reflect graded differences in typicality. This is not the case, as shown by the scatter plots of Figure 2. In particular, the upper-right panel reveals that, among Mystery books (as determined by our ground truth), there exists a strong positive association between *human typicality* and *BERT typicality* ($\rho = 0.63$). The positive association also holds among Non-Mystery books ($\rho = 0.67$). A similar pattern holds for the Romance genre, as revealed by the lower-right panel (Romance books: $\rho = 0.54$; Non-Romance books: $\rho = 0.72$). The model therefore reflects between-book differences in *human typicality* beyond differences in category membership.

Typicality Measures Based on Other BERT Models

To develop an intuition for what explains the very good fit of the *BERT typicality* with *human typicality*, we proceed to a set of comparisons with three other BERT-based typicality measures. As explained in sections Using a Probabilistic Classifier to Measure the Typicality of Objects in a Concept and Measuring the Typicality of Text Documents with a BERT Probabilistic Classifier, the BERT classifier used to construct the *BERT typicality* involves two distinct components (language

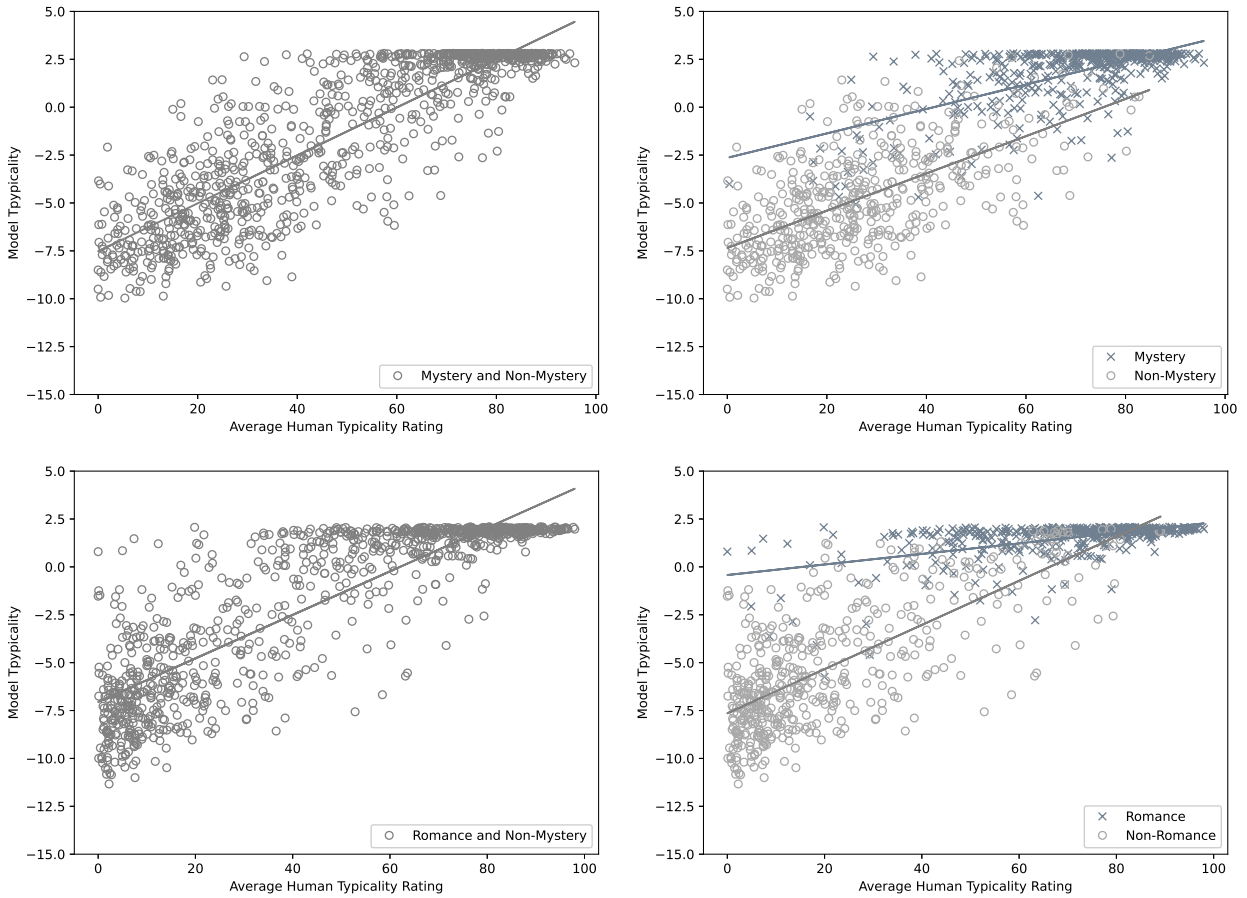

Figure 2: There exists a strong positive association between *BERT typicality* and *human typicality*. *Upper left:* All books in the prediction data for the Mystery genre. *Upper right:* The positive association holds for Mystery and Non-Mystery books. *Lower left:* All books in the prediction data for the Romance genre. *Lower right:* The positive association holds for Romance and Non-Romance books.

representation—the mapping of text documents into vector of coordinates in feature space—and categorization) that each have a set of free parameters that are adjusted during model training. Here, we aim to unpack the contributions of these two components. The very good fit could come from the fact that model training allowed the construction of a language representation adapted to the judgment task (providing a typicality rating in the focal concept). It could also come from the definition of the typicality measure in terms of the categorization probabilities produced by the trained categorization component. This is because the categorization component includes many free parameters (several hundreds) that are adjusted during model training to give more weight to the feature dimensions that matter more for categorization. If the features that matter more for categorization (in the ground truth data) are also the features that matter more for typicality judgment

(by humans), defining the typicality based on the categorization probability likely contributes to high correlation with human typicality.

The three additional BERT-based typicality measures we consider in this analysis take one or both of these characteristics away. More specifically, we consider the following four BERT-based typicality measures:

1. Baseline *BERT typicality*: Fine-tuned BERT representation; typicality measure is based on categorization probabilities produced by the trained categorization component.
2. Fine-tuned BERT representation; typicality measure is constructed with no free parameter.
3. Pre-trained BERT representation; typicality measure is based on categorization probabilities produced by the trained categorization component.
4. Pre-trained BERT representation; typicality measure is constructed with no free parameter.

We obtain version 2 by using the BERT language representation fine-tuned in the construction of the baseline *BERT typicality*, but with a different formula for typicality. Instead of defining the typicality of an object in the focal concept as a transformation of the categorization probability of the object in the concept, we use the cosine similarity (i.e., correlation) between the position of object in the feature space and position of the concept prototype (the average position of objects that are instances of the concept in the training data; see Kozlowski et al. [2019] for a similar formulation). This definition gives the same weight to all 768 feature dimensions of the fine-tuned language representation.[17]

Version 3 uses the same definition of typicality in terms of categorization probabilities as in the baseline *BERT typicality* (Eq. [4]) but differs in how the BERT classifier is trained. The BERT language representation is not fine-tuned: the parameters of the BERT representation are "frozen" at their initial (pre-trained) values. Therefore the language representation is not adapted to the specific categorization task in the focal genre. Only the parameters of the categorization component are adjusted during training. Model training adjusts these parameters in a way that gives more weight to the features that matter more for classification (as in a logistic regression).

Finally, we obtain version 4 by calculating the cosine similarity between object and prototype, using the positions in the feature space produced by the pre-trained BERT language representation. This typicality measure does not involve the adjustment of any free parameter. It gives the same weight to all feature dimensions of the pre-trained language representation.

The results obtained with the four approaches are reported in the first four rows of Tables 3 and 4. For both genres, the performance of the baseline *BERT typicality* (version 1) is the best, in terms of both overall correlation with *human typicality* and correlation within category. The performance of the typicality measure that does not involve any free parameter (version 4) is poor and in any case much lower than that of the three other versions. The two versions that involve either a

Table 3: Comparing the respective fit of model typicality measures with *human typicality*. Mystery genre

	Correlations between model-based typicalities and human typicality ratings			Model components			Model training	
Typicality measure	Mystery and Non-Mystery books	Mystery books	Non-Mystery books	Language representation	Categorization	Similarity between text document and concept	Fine-tuning the language representation	Training a probabilistic classifier
BERT fine-tuned / categorization probability (baseline BERT typicality)	0.87	0.67	0.63	BERT	Probabilistic binary	Typicality (Eq. [4])	Yes	Yes
BERT fine-tuned / correlation with prototype	0.80	0.58	0.61	BERT	Probabilistic binary	Cosine	Yes	Yes
BERT pre-trained / categorization probability	0.64	0.33	0.49	BERT	Probabilistic binary	Typicality (Eq. [4])	No	Yes
BERT pre-trained / correlation with prototype	0.07	0.03	0.07	BERT	None	Cosine	No	No
BERT fine-tuned / categorization probability 36	0.85	0.61	0.59	BERT	Probabilistic 36 categories	Typicality (Eq. [4])	Yes	Yes
GloVe fine-tuned / categorization probability	0.81	0.59	0.50	GloVe	Probabilistic binary	Typicality (Eq. [4])	Yes	Yes
GloVe fine-tuned / correlation with prototype	0.41	0.19	0.36	GloVe	Probabilistic binary	Cosine	Yes	Yes
GloVe pre-trained / categorization probability	0.79	0.58	0.48	GloVe	Probabilistic binary	Typicality (Eq. [4])	No	Yes
GloVe pre-trained / correlation with prototype	0.22	0.29	0.17	GloVe	None	Cosine	No	No
Word frequencies / categorization probability	0.64	0.43	0.47	BoW term frequencies	Probabilistic binary	Typicality (Eq. [4])	No	Yes
TF-IDF / categorization probability	0.81	0.60	0.58	BoW TF-IDF	Probabilistic binary	Typicality (Eq. [4])	No	Yes
TF-IDF / correlation with prototype	0.15	0.13	0.18	BoW TF-IDF	None	Cosine	No	No
Label proportion	0.76	0.20	0.33	None	None	Label proportion	—	—
BERT fine-tuned / categorization probability proportion	0.89	0.69	0.73	BERT	Probabilistic binary proportions	Typicality (Eq. [4])	Yes	Yes

Table 4: Comparing the respective fits of model typicality measures with *human typicality*. Romance genre

	Correlations between model-based typicalities and human typicality ratings			Model components			Model training	
Typicality measure	Romance and Non-Romance books	Romance books	Non-Romance books	Language representation	Categorization	Similarity between text document and concept	Fine-tuning the language representation	Training a probabilistic classifier
BERT fine-tuned / categorization probability (baseline BERT typicality)	0.86	0.54	0.72	BERT	Probabilistic binary	Typicality (Eq. [4])	Yes	Yes
BERT fine-tuned / correlation with prototype	0.85	0.56	0.70	BERT	Probabilistic binary	Cosine	Yes	Yes
BERT pre-trained / categorization probability	0.70	0.40	0.55	BERT	Probabilistic binary	Typicality (Eq. [4])	No	Yes
BERT pre-trained / correlation with prototype	0.07	0.11	0.04	BERT	None	Cosine	No	No
BERT fine-tuned / categorization probability 36	0.85	0.51	0.67	BERT	Probabilistic 36 categories	Typicality (Eq. [4])	Yes	Yes
GloVe fine-tuned / categorization probability	0.77	0.53	0.54	GloVe	Probabilistic binary	Typicality (Eq. [4])	Yes	Yes
GloVe fine-tuned / correlation with prototype	0.59	0.27	0.47	GloVe	Probabilistic binary	Cosine	Yes	Yes
GloVe pre-trained / categorization probability	0.76	0.45	0.56	GloVe	Probabilistic binary	Typicality (Eq. [4])	No	Yes
GloVe pre-trained / correlation with prototype	0.30	0.30	0.22	GloVe	None	Cosine	No	No
Word frequencies / categorization probability	0.62	0.34	0.42	BoW term frequencies	Probabilistic binary	Typicality (Eq. [4])	No	Yes
TF-IDF / categorization probability	0.76	0.48	0.54	BoW TF-IDF	Probabilistic binary	Typicality (Eq. [4])	No	Yes
TF-IDF / correlation with prototype	0.18	0.14	0.20	BoW TF-IDF	None	Cosine	No	No
Label proportion	0.76	0.06	0.50	None	None	Label proportion	—	—
BERT fine-tuned / categorization probability proportion	0.84	0.56	0.72	BERT	Probabilistic binary proportions	Typicality (Eq. [4])	Yes	Yes

fine-tuned representation (with typicality defined in terms of the cosine similarity with the prototype) or the pre-trained representation with typicality defined in terms of the categorization probabilities achieve a fairly high performance. Both versions involve the training of a probabilistic classifier to fine-tune the language representation (even version 2, which defines typicality in terms of cosine similarity in feature space rather than in terms of categorization probabilities).

We conclude from the comparison of these four BERT-based typicality measures that the crucial ingredient necessary for BERT-based typicalities to reflect human typicality judgments lies in training a probabilistic classifier. Model training allows for the construction of a language representation adapted to the task at hand, or the identification of features that matter most for categorization, or both. Achieving one of these two goals seems sufficient to obtain a typicality measure that reflects human judgments well, but achieving both allows for even better performance. The resulting combination of a feature space adapted to the categorization task and typicality function provides a mathematical representation of concepts that reflects humans' mental representation of concepts. In other words, this approach uncovers the semantics of concepts.

Typicality Based on Training a Multiclass BERT Classifier

In the sections Using a Probabilistic Classifier to Measure the Typicality of Objects in a Concept and Measuring the Typicality of Text Documents with a BERT Probabilistic Classifier, we proposed that the text classifier be trained on data that include binary labels: ground truth categorization data that indicate whether each observation belongs to the focal concept c (Y_{train}: a vector of N_{train} rows and populated with 0s and 1s). In a setting like Goodreads.com, there are many candidate genres (36 of them). Our proposed approach implies that the nature of the genres that are not the focal genre (e.g., those other than Mystery) is ignored. All observations that are not instances of the focal genre are labeled "other" in the training and validation data.

An alternative approach would train a classifier that predicts the genre of a book description among the 36 candidates. Such a classifier would output a vector of 36 categorization probabilities for each book description. We trained the model with the same training sets used for the construction of the baseline *BERT typicality*, except that Y_{train} was now a matrix of N_{train} rows and 36 columns, with each row indicating the genre of the corresponding book description. The results obtained using this multiclass classifier are almost the same as those obtained with the baseline *BERT typicality* (see row "BERT fine-tuned / categorization probability 36" in Tables 3 and 4).

This analysis suggests that jointly training the model to output categorization probabilities for many candidate genres does not substantially hurt the correspondence of the BERT-based typicality measure with the *human typicality*, as compared with what is obtained with a narrower focus on the focal genre. From a practical standpoint, this is good news. This suggests that, to measure the typicality of objects in many concepts, the analyst does not have to train one probabilistic classifier per concept but can do just as well by training one model. Given that each model training can take several hours (if the data have several hundred thousand observations,

as the Goodreads.com data used in this section), this implies considerable savings in computing time.

Benchmarking: Comparing BERT Typicalities with Typicalities Obtained with Other Probabilistic Classifiers or with Label Assignments

Typicality Based on GloVe Embeddings

We replicated the comparison of the four BERT-based typicality measures using word embeddings as a language representation instead of the BERT language representation. More precisely, we used a word-embedding layer with pre-trained weights in our classifiers instead of the BERT language representation. (See Appendix: Methodological Details for further details). We employed GloVe word embeddings (Pennington, Socher, and Manning 2014) to transform text documents into vectors. GloVe is a *word*-embedding model, not a text-embedding model. Accordingly, we needed to combine word positions in the embedding space to create a unique position for text documents (book descriptions from Goodreads.com). We used the average position of the words in the book description as the position of the book description in feature space.

The results obtained with the four GloVe-based typicality measures are reported in Tables 3 and 4. The overall performance of the GloVe-based typicality measures is very good, although not as high as that obtained with the BERT language representation. Because the BERT classifier is sensitive to bidirectional dependencies between words but the GloVe classifier is not, this unsurprisingly suggests that typicality judgments are also sensitive to such dependencies.

Comparison of the performance of the four GloVe-based typicality measures leads to the same conclusion as that obtained from comparing the four BERT-based typicality measures: what is crucial in achieving a good performance is that measure construction involves training a probabilistic classifier.

Typicality Based on Bag-of-Words Representations of Text Documents

We also used a standard machine-learning text classifier based on a bag-of-words (BoW) representation of text documents: the naive Bayes classifier (Maron 1961). This machine-learning classifier produces categorization probabilities based on word co-occurrences. It is computationally undemanding, but its representation of text documents is not sensitive to the order of words in sentences. Also the representation of words does not depend on their semantic similarity. We call the typicality measure constructed by applying Equation (2) to the resulting categorization probabilities the *word frequency categorization probability typicality*. Because of the simpler nature of the language representation used in the classifier, we expected that this typicality measure would provide a lower fit to human typicality judgments than the *BERT typicality*.

Results reported in Tables 3 and 4 confirm this prediction. It is noteworthy that despite the simplicity of the classifier used here, the resulting typicality measure is fairly highly correlated with *human typicality*.

In additional analyses, we used a version of the BoW representation that does not rely on simple word frequencies but weighs them by diminishing the importance of words that occur in many text documents. This approach is known as term frequency–inverse document frequency, or TF-IDF (Jones 1972). We constructed two typicality measures based on this representation. The first one is based on the categorization probabilities produced by a naive Bayes classifier that uses the TF-IDF representation (instead of simple word frequencies). We call it the *TF-IDF categorization probability typicality*. The second measure uses the cosine similarity between vectors of weighted frequencies that correspond to text documents and the prototype (just as we did with BERT and GloVe embedding representations). We call it the *TF-IDF correlation with prototype typicality*.

Results reported in Tables 3 and 4 show that the performance of the *TF-IDF categorization probability typicality* is better than that of the *word frequency categorization probability typicality*, but not as high as that of the *BERT typicality*. The performance of the *TF-IDF correlation with prototype typicality* measure is poor. This is not surprising, because this typicality measure relies on a generic language representation, and the transformation of positions in the feature space into typicality does not have free parameters that are adjusted via model training.

Comparison with Label-Based Approaches to Measuring Typicality

As explained in the introductory section, a central motivation for the development of typicality measures based on the predictions of machine-learning probabilistic classifiers is that these classifiers can produce typicality measures in settings in which the data source includes only binary information about concept membership (e.g., a book is either a Mystery or not). In such settings, the widely used approach to measuring typicality using label proportions (Pontikes 2008; Kovács and Hannan 2015) cannot be used. Therefore, the *BERT typicality* measure we proposed in earlier sections offers a clear benefit.

Yet, the reader might wonder if the benefit of our approach exclusively resides in the possibility of constructing typicality measures in settings where prior methods would not allow this to be done, or if the approach we advocate in this article also presents benefits in settings where prior methods are also applicable—when label proportions are available.

The original data source we used for our empirical illustrations (the Goodreads data set) includes multiple label assignments.[18] Next, we use these to construct measures based on label proportions and assess their fit with human typicality ratings. We compare performance with the baseline *BERT typicality* measure and another version of the *BERT typicality* that uses label proportions as inputs.

In the first and largely implicit step in devising a measure, the analyst assumes that objects with only one categorical assignment generally fit better to the concept than those assigned two concepts. The reasoning then makes a similar assertion about dual categorization versus triple categorization, and so forth. This reason-

ing leads to the expectation that the typicality in any assigned concept decreases monotonically with the number of concepts assigned subject to the condition that it remain non-negative. Prior research has used the following functional form:

$$t(c,o) = \frac{p_c(o)}{\sum_{c' \in \kappa} p_{c'}(o)}. \quad (5)$$

For example, if reviewers apply the concept c_1 to an object eight times and apply the concepts c_2 and c_3 each one time, then $t_{c_1}(o) = 0.8$ and $t_{c_2}(o) = 0.1 = t_{c_3}(o)$.

It seems a priori unfair to compare the performance of *BERT typicality* with that of these measures based on label proportions, because the *BERT typicality* is based on the predictions of a BERT classifier trained on binary categorization. To make the comparison more meaningful, we also trained the BERT classifier on training data that included label proportions. We followed the same approach as that exposed in Using a Probabilistic Classifier to Measure the Typicality of Objects in a Concept and Measuring the Typicality of Text Documents with a BERT Probabilistic Classifier except for a change in the nature of the training data (we kept the categorical cross-entropy loss function). The training data consist of the proportions of assignments of each book to the focal genre. In other words, Y_{train} is a vector of N_{train} rows and populated with real values in [0,1], the proportions of labels that correspond to the focal concept. In Tables 3 and 4 "BERT fined-tuned / categorization probability proportion" refers to the BERT-based typicality measure obtained with this different training procedure, and "Label proportion typicality" refers to the typicality measure defined by Equation (5).

The results are very clear: the *label proportion typicality* reflects *human typicality* less well than BERT-based typicalities, be it the baseline version trained with binary labeled data or the version trained with label proportions. This is the case in terms of overall correlation but most crucially in terms of within-category correlations. The finding that BERT typicalities obtained from coarse categorizations (binary labeling) are a better fit to human typicality ratings than typicality measures based on label proportions suggests that the language representation constructed by a BERT classifier more than compensates for the coarseness of the training data. Even more so, the quasi absence of difference in performance between the two versions of the BERT typicalities suggests that there might be little potential gain associated to more fine-grained training data in the form of label proportions.

In summary, our findings provide evidence that the typicalities based on the categorization probabilities produced by a BERT classifier trained on data that consist of coarse categorizations (binary labeling) allow us to achieve the objective we stated in the introductory section to produce fine-grained typicality measurements that closely match human typicality ratings. More research is clearly needed to assess the extent to which similar findings would hold in other domains, but the evidence reported here is an important proof of concept.

Discussion

In this article, we investigate how deep learning can contribute to the measurement of typicality of objects in concepts. Although the question of "what belongs" (and what difference this makes) has interested many thinkers at least since Aristotle, these thinkers had to rely on anecdotes, literary analyses, smaller scale observational studies, interviews, surveys, and lab experiments. With the new revolution in data availability and big-data methods, we can finally embark on systematic exploration of meanings of concepts, their fuzziness, and how people's reaction to entities depends on their typicality.

This article provides a methodological contribution. We illustrate how large-scale text data can be analyzed for sociological analysis with a deep-learning text-categorization model. Deep learning is not just a powerful method in machine learning; its ability to learn high-dimensional feature spaces from natural language and to produce categorization probabilities for positions in the space directly mirrors recent theorizing in cognitive psychology that relies on probabilistic representations of concepts and categories (see Hannan et al. [2019] for a review). This effectively ties this powerful theoretical approach to the types of issues and data of interest in sociological analysis.

We obtained a feature space by fine-tuning a BERT language representation for categorization, taking book descriptions as input, and training the model to predict the genre of a text. Our model also produces a mapping between positions in the feature space and categorization probabilities. We then use these categorization probabilities to measure typicality by direct application of the equation outlined in the theoretical part of the article (Eq. [2]). The excellent fit of the resulting typicality measure with human typicality judgments indicates that the joint construction of the feature space and the mapping between positions and categorization probabilities results in a mathematical representation of concepts that reflects humans' mental representation of the concept (at least with respect to typicality judgments).

Besides providing a general framework that illustrates how machine learning could be used to construct typicality measures, our article advocates the use of one specific language representation, BERT, to do so. The initial motivation for this analysis was that models based on the BERT language representation have proven to have exceptional performance in solving language tasks. It seemed intuitive that this class of models would also do a good job in capturing how humans make typicality judgements. To the best of our knowledge, this article is the first to provide evidence that this is the case.

We have addressed this question in the context of a sociological problem of judging the typicalities of books (specifically, their descriptions) with respect to a genre (agreed-upon concept). We judge performance as the strength of the correlation of typicalities calculated from BERT with average human typicalities of the same book descriptions. Our main analysis picks a pair of genres (Mystery, Romance) and trains BERT separately on each. The correlation of typicalities derived from our trained model with human judgements is 0.87 for Mystery and 0.86 for Romance.

We judge these correlations to be sufficiently high to warrant a positive answer to the question posed in the article's subtitle: How well do typicalities extracted from a BERT classifier match human judgments of genre typicalities?[19]

We find it interesting and important that the use of BERT gives high performance that goes beyond categorization—whether a book is an instance of a genre. It also gives useful *graded* answers. Within subsets of books that the majority of those making genre assignments regard as an instance of a genre, the procedure we recommend also does well at matching humans in judging typicality.

This impressive pattern of performance also holds when we vary the categorization task and train BERT on the full set of 36 genres and then calculate typicalities in focal genres for comparison with human judgements. This result suggests that researchers can begin by training BERT on multiple-concept tasks and, if desired, later narrow the focus.

None of the other options we tried (variations on typicalities based on word vectorization, bag-of-words representations, or label-based approaches) perform as well as BERT especially on capturing between-text-document differences in human typicality judgements within a focal category. Nonetheless some variations we tried worked nearly as well as BERT. Our analysis of the performance differences suggests that using a language model or a trained machine-learning classifier is necessary to generate even moderately good performance. But using both leads to even better performance.

Although most social scientists have traditionally focused on hypothesis testing and cared less about model fit, we do think that even theorists should embrace novel methodologies with much improved model fit. This is because if the prediction power of a key theoretical concept such as typicality increases from 64 to 87 percent (see Table 3). as is the case with the typicalities based on the naive Bayes (bag-of-words) model, versus BERT typicalities, as found in this article, then the empirical tests become much more reliable. Therefore, empirical tests will be less likely to lead down dead-end theoretical paths.

Finally, we think that the empirical application of computing the genre typicality of books based on their text descriptions is just one of the many potential applications of the use of BERT classifiers for sociological analysis. One could conduct similar analyses for films to measure typicalities of scripts, or in the case of music, lyrics. One could study the menus of restaurants to see how novel dishes spread, or the abstracts and texts of academic articles to check whether articles that are more typical of a journal receive more citations. By locating objects in a conceptual space, one could study the extent to which producers change, for example, by calculating the distance between the texts of an author's new and prior books (Kovács, Hsu, and Sharkey 2021). Analyzing the text of patents, one can measure the extent to which a patent is groundbreaking (Kelly et al. 2021) or ascertain the extent to which a firm changed the direction of its innovation. Finally, one could use the same approach to compute the political orientations of tweets posted by politicians and, in turn, the political orientation of their online discourse (Le Mens et al. 2020; Konovalova, Le Mens, and Schöll 2022).

We believe that social science in general, and the study of categorization specifically, is on the brink of a revolution rendered possible by the application of deep-learning methods to big data.

Appendix: Methodological Details

The script used for training the BERT classifier is available at https://osf.io/ta273/.

Text Tokenization

Text tokenization consists in identifying the relevant semantic units in the pre-processed text. Specifically, the implementation of the BERT model we used relies on the *WordPiece* tokenizer, which creates a token dictionary with the most frequent tokens in a text (a token being either a full word or a subset of the characters of a word). The dictionary is a file that consists of two columns. Elements of first column are token strings, and elements of the second column are unique token identifiers. The dictionary for the BERT-base-cased model we used contains 30,000 distinct tokens and was constructed on the pre-training data set made of BookCorpus, a data set consisting of 11,038 unpublished books and English Wikipedia (excluding lists, tables, and headers).[20] Applying the WordPiece tokenizer to each text document produces a sequence of tokens. Consider an example sentence: "The mouse ran away, but the cat ate the mouse." The corresponding sequence of tokens is ["The", "mouse", "ran", "away", ",", "but", "the", "cat", "at", "##e", "the", "mouse"]. Each such sequence is then matched with BERT's token vocabulary.[21] For each text document, a fixed-length sequence of matched token identifiers is finally returned, denoted as the *tokenized document*. The length of each sequence was set to $L = 512$ tokens, where longer or shorter sequences were respectively trimmed or zero-padded (zeros were added to the right of the last token identifiers until the sequence reached a length of L elements).[22] The tokenized document for the example sentence is $[10117, 63986, 17044, 14942, 117, 10473, 10105, 41163, 10160, 10112, 10105, 63986, 0, \ldots, 0]$, where the zeros are added to make the length of the sentence equal to 100 elements.

Parameters of the Optimizer Used for Fine-Tuning the BERT Classifier

We implemented our BERT categorization model using the TensorFlow machine-learning library and its higher-level wrapper Keras. We used the Adam optimizer to minimize the cross-entropy loss function for fine-tuning the model. This amounts to trying to find the parameters that maximize the likelihood of the true categories in the training data. We used the following (standard) parameter values:

- Batch size: 64
- Maximum number of tokens: 512 (This is the maximum possible with pre-trained BERT.)
- Optimizer: Adam
- Loss function: categorical_crossentropy
- Learning rate: 2e-5

- Epochs: 2

For the frozen BERT we used the same parameter values, except for a faster learning rate (5e-4) and more epochs (50).

All the models were specified by attaching an average layer on top of the embedding layer (which computes the mean position of the tokens in the text), a dense layer with number of nodes equal to the number of categories with a softmax activation.

GloVe

From the training data we got the top 20,000 most frequent words in the book descriptions; then we used the GloVe embedding model "glove.6B.300d" (trained on Wikipedia) to transform each of the 20,000 words into vector positions (all but 651 words appeared in both our top 20,000 most frequent words and the GloVe embedding). With this embedding we created a basic deep-learning model consisting of an embedding layer, a pooling layer, and a dense layer (with softmax activation).

We trained the model for five epochs with a 2e-3 learning rate, with the task of classification (using the categorical cross-entropy loss function). The model parameters were selected after some exploration, maximizing model performance in terms of loss minimization in the validation set.

Bag-of-Words Models

From the training data, we selected a subset of 50,000 book descriptions and got the 3,000 most frequent words (we chose 3,000 words due to RAM limitation on the platform we used to run the computations). We assigned to each word an ID and transformed all the book descriptions using this ID dictionary. Finally, we fit a multinomial naive Bayes classifier on all the book descriptions in the training data.

For the construction of the dictionary, we used the sklearn package "CountVectorizer," and to fit the model we used the sklearn package "MultinomialNB."

Appendix: Goodreads.com Data

The following is an example of a Goodreads entry (https://www.goodreads.com/book/show/15806231-calculated-in-death). The counts of categorizations for this book are Mystery (631), Romance (248 users), Mystery and Crime (152), Fiction (152), Romance and Romantic Suspense (128), Futuristic (118), Suspense (109), Science Fiction (83), Thriller (65).

Calculated in Death

J.D. Robb

4.28 [average rating] · 23,205 ratings · 1,469 reviews

> On Manhattan's Upper East Side a woman lies dead at the bottom of the stairs, stripped of all her valuables. Most cops might call it a mugging gone wrong, but Lieutenant Eve Dallas knows better.
>
> A well-off accountant and a beloved wife and mother, Marta Dickenson doesn't seem the type to be on anyone's hit list. But when Eve and her partner, Peabody, find blood inside the building, the lieutenant knows Marta's murder was the work of a killer who's trained, but not professional or smart enough to remove all the evidence.
>
> But when someone steals the files out of Marta's office, Eve must immerse herself in her billionaire husband Roarke's world of big business to figure out who's cruel and callous enough to hire a hit on an innocent woman. And as the killer's violent streak begins to escalate, Eve knows she has to draw him out, even if it means using herself as bait...

Published February 26, 2013 by Putnam Adult

Table A: Relative frequencies of assignment by genre for the whole sample

Genre	Percent	Genre	Percent	Genre	Percent
Romance	12.11	Comics	2.41	Self Help	1.31
Children's	8.88	Humor and Comedy	2.12	Thriller	1.27
History	7.68	Horror	2.07	Philosophy	1.20
Fantasy	6.95	Religion	1.88	Travel	1.18
Mystery	6.01	Science	1.82	Chick Lit	1.06
Young Adult	5.05	Memoir	1.80	Psychology	0.91
Science Fiction	4.69	Christian	1.73	Suspense	0.9
Contemporary	3.92	Cookbooks	1.70	Sports	0.89
Historical Fiction	3.32	Manga	1.55	Spirituality	0.87
Paranormal	3.06	Classics	1.51	Crime	0.82
Poetry	2.53	Business	1.49	Music	0.82
Biography	2.49	Graphic Novels	1.47	Gay and Lesbian	0.52

Notes

1 For another sociological application of BERT, see Vincianza, Goldberg, and Srivastava (2020).

2 Work that has examined the similarity of concepts in word-embedding spaces employs this kind of averaging (e.g., Kozlowski et al. 2019); so does the work on beauty-in-averageness (e.g., Vogel et al. 2018).

3 Garg et al. (2018) and Lewis and Lupyan (2020) use this kind of averaging, although they do not focus on objects but on the typicality of particular words in a concept.

4 Hannan et al. (2019) postulate that $\pi_{\pi_G}(x\,|\,c) = P(x\,|\,c)$. Bayes' theorem implies that the new formulation of typicality is increasing in $P(x\,|\,c)$:

$$\tau_c(x) \equiv \log \frac{P(c\,|\,x)}{P(c)} = \log \frac{P(x\,|\,c)}{P(x)} = \log \frac{\pi_G(x\,|\,c)}{P(x)}, \tag{6}$$

where the last equality holds under the assumption made by Hannan et al. (2019). So both renderings of intuitions about typicality are increasing in $P(x\,|\,c)$. The new definition offers a clear practical advantage, however, in that the machine-learning classifiers we use to construct empirical measures of typicality output the categorization probabilities associated to a position (x), $P(c\,|\,x)$, but do not provide access to empirical measures of the concept likelihood at position x, $\pi_G(x\,|\,c)$.

5 This proposition is similar to that of Kozlowski et al. (2019) who proposed that similarity relations in embedding spaces "reflect widely shared cultural associations" (P. 918).

6 When observations in the available categorization data $\mathcal{D}$ are not independent, and the dependence structure is known, it is advisable to put all dependent observations in one of the two sets, ensuring independence between the training and validation sets.

7 The maximal L value for use with pre-trained BERT models is 512, which corresponds to approximately 300 English words.

8 State-of-the-art text-categorization models were based on the LSTM architecture until the advent of BERT.

9 One of the most widely accepted benchmarks for natural language understanding models is the GLUE (General Language Understanding Evaluation) benchmark (https://gluebenchmark.com/leaderboard). Examination of this ranking reveals that BERT performs much better than previously introduced models, in particular bag-of-words–based models (CBOW). Virtually all the models that perform better than the original BERT are direct extensions of this model.

10 The task used for pre-training is "word-masking": a small proportion of words in the input text are masked to the model, and the model has to predict which word has been masked on the basis of the non-masked words that come before and after the masked word.

11 BERT-base-cased was trained on Wikipedia and BookCorpus. Other available models have been pre-trained in other languages (e.g., French, Spanish) or on several languages at once (BERT-base-multilingual-cased). See HuggingFace.co for a large library of pre-trained models.

12 For computational reasons, we excluded books whose descriptions exceed 300 words. These are about 4.6 percent of the data. The original sample contained 768,249 books.

13 We do not have data on the individual shelving events. Our data on categorizations do not contain information on individual categorization events but provide only the

aggregated distribution of categorizations (e.g., X people tagged the book as Mystery, Y as Comedy).

14 For the sake of discussion of classification accuracy, we assume that a book is predicted to be a Mystery book if the probability of categorization in Mystery is higher than in Non-Mystery.

15 This project is available on the Open Science Framework (OSF) at https://osf.io/ta273/. The "compute_typicality" folder contains a Python notebook that can be used to train the model and compute typicalities using dedicated hardware freely available via the Google Colab service (https://colab.research.google.com), the data set used to compute the typicalities of books in the Mystery genre, and a Readme file proving instructions about how to use the notebook.

16 We had 50 experimental conditions each corresponding to a set of 20 book descriptions. The random allocation of participants in conditions implied some variability in the final number of participants in each condition.

17 In ancillary analyses, we defined the typicality of an object in a concept by taking the average of the cosine similarities between the object and all instances of the concept in the training data. The results obtained with this "exemplar approach" are almost the same as those obtained with the "prototype approach."

18 For the sake of illustration of our method, we "impoverished" the training data by discretizing them: we considered a book as a Mystery book or not.

19 Our claim is not that BERT is the best model ever; indeed, very recent work has already extended BERT to provide representations that improve on BERT (e.g., RoBERTa, CPT2, CPT3). Our point is that BERT typicalities are already quite close to human typicality judgments.

20 The tokenizer function we used is a component of the publicly available transformers library of language models: https://huggingface.co/transformers/main_classes/tokenizer.html.

21 Tokens with the same letters but different font cases (e.g., "The" and "the") are treated as distinct tokens.

22 $L = 512$ is the largest number of tokens that can be used with pre-trained BERT. Given the empirical distribution of the number of tokens for the book descriptions in our data, 0.06 percent of the book descriptions were trimmed. We chose the parameter value of $L = 512$ to maximize prediction accuracy at the cost of longer computing time (as compared with what would be obtained with a smaller L).

References

Anderson, John R. 1991. "The Adaptive Nature of Human Categorization." *Psychological Review* 98(3):409–29. https://doi.org/10.1037/0033-295X.98.3.409.

Ashby, F. Gregory, and Leola A. Alfonso-Reese. 1995. "Categorization as Probability Density Estimation." *Journal of Mathematical Psychology* 39(2):216–33. https://doi.org/10.1006/jmps.1995.1021.

Beller, Sieghard, Andrea Bender, and Douglas L. Medin. 2012. "Should Anthropology Be Part of Cognitive Science?" *Topics in Cognitive Science* 4(3):342–53. https://doi.org/10.1111/j.1756-8765.2012.01196.x.

Bender, Andrea, Edwin Hitchins, and Douglas Medin. 2010. "Anthropology in Cognitive Science." *Topics in Cognitive Science* 2(3):374–85. https://doi.org/10.1111/j.1756-8765.2010.01082.x.

Bhatia, Sudeep, and Russell Richie. 2022. "Transformer Networks of Human Conceptual Knowledge." *Psychological Review*, first published October 27, 2022. https://doi.org/10.1037/rev0000319.

Bonikowski, Bart, Yuchen Luo, and Oscar Stuhler. 2022. "Politics as Usual? Measuring Populism, Nationalism, and Authoritarianism in U.S. Presidential Campaigns (1952–2020) with Neural Language Models." *Sociological Methods & Research* 51(4):1721–87. https://doi.org/10.1177/00491241221122317.

Cerulo, Karen A., Vanina Leschziner, and Hana Sheperd. 2021. "Rethinking Culture and Cognition." *Annual Review of Sociology* 47:63–85. https://doi.org/10.1146/annurev-soc-072320-095202.

D'Andrade, Roy. 1995. *The Development of Cognitive Anthropology*. Cambridge University Press.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." Preprint, arXiv:1810.04805 [cs.CL]. https://arxiv.org/abs/1810.04805.

DiMaggio, Paul J. 1997. "Culture and Cognition." *Annual Review of Sociology* 23:263–87. https://doi.org/10.1146/annurev.soc.23.1.263.

DiMaggio, Paul J., Manish Nag, and David M. Blei. 2013. "Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding." *Poetics* 41(6):570–606. https://doi.org/10.1016/j.poetic.2013.08.004.

Durand, Rodolphe, and Pierre-Antoine Kremp. 2016. "Classical Deviation: Organizational and Individual Status as Antecedents of Conformity." *Academy of Management Journal* 59(1):65–89. https://doi.org/10.5465/amj.2013.0767.

Feldman, Naomi H., Thomas L. Griffiths, and James L. Morgan. 2009. "The Influence of Categories on Perception: Explaining the Perceptual Magnet Effect as Optimal Statistical Inference." *Psychological Review* 116(4):752–82. https://doi.org/10.1037/a0017196.

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. "Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes." *Proceedings of the National Academy of Sciences* 115(16):E3635–44. https://doi.org/10.1073/pnas.1720347115.

Hannan, Michael T. 2010. "Partiality of Memberships in Categories and Audiences." *Annual Review of Sociology* 36:159–81. https://doi.org/10.1146/annurev-soc-021610-092336.

Hannan, Michael T., Gaël Le Mens, Greta Hsu, Balázs Kovács, Giacomo Negro, Lászlo Pólos, Elizabeth G. Pontikes, and Amanda J. Sharkey. 2019. *Concepts and Categories: Foundations for Sociological and Cultural Analysis*. New York: Columbia University Press.

Hannan, Michael T., László Pólos, and Glenn R. Carroll. 2007. *Logics of Organization Theory: Audiences, Codes, and Ecologies*. Princeton University Press.

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation* 9(8):1735–80. https://doi.org/10.1162/neco.1997.9.8.1735.

Hsu, Greta. 2006. "Jacks of All Trades and Masters of None: Audiences' Reactions to Spanning Genres in Feature Film Production." *Administrative Science Quarterly* 51(3):420–50. https://doi.org/10.2189/asqu.51.3.420.

Hsu, Greta, Michael T. Hannan, and Özgeçan Koçak. 2009. "Multiple Category Memberships in Markets: An Integrative Theory and Two Empirical Tests." *American Sociological Review* 74(1):150–69. https://doi.org/10.1177/000312240907400108.

Jones, Karen Sparck. 1972. "A Statistical Interpretation of Term Specificity and Its Application in Retrieval." *Journal of Documentation* 28(1):11–21. https://doi.org/10.1108/eb026526.

Kelly, Bryan, Dimitris Papanikolaou, Amit Seru, and Matt Taddy. 2021. "Measuring Technological Innovation over the Long Run." *American Economic Review: Insights* 3(3):303–20. https://doi.org/10.1257/aeri.20190499.

Konovalova, Elizaveta, Gaël Le Mens, and Nikolas Schöll. 2022. "Social Media Feedback and Extreme Opinion Expression." Working paper.

Kovács, Balázs, and Michael T. Hannan. 2010. "The Consequences of Category Spanning Depend on Contrast." In *Categories in Markets: Origins and Evolution*, edited by Greta Hsu, Giagomo Negro, and Özgecan Koçak, volume 31 of *Research in the Sociology of Organizations*, pp. 175–201. Bingley, United Kingdom: Emerald Group Publishing. https://doi.org/10.1108/S0733-558X(2010)0000031008.

Kovács, Balázs, and Michael T. Hannan. 2015. "Conceptual Spaces and the Consequences of Category Spanning." *Sociological Science* 2:252–86. https://doi.org/10.15195/v2.a13.

Kovács, Balázs, Greta Hsu, and Amanda Sharkey. 2021. "The Stickiness of Category Labels: Audience Perception and Evaluation of Change in Creative Markets." Working paper.

Kovács, Balázs, and Rebeka Johnson. 2014. "Contrasting Alternative Explanations for the Consequences of Category Spanning: A Study of Restaurant Reviews and Menus in San Francisco." *Strategic Organization* 12(1):7–37. https://doi.org/10.1177/1476127013502465.

Kozlowski, Austin C., Matt Taddy, and James A. Evans. 2019. "The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings." *American Sociological Review* 84(5):905–49. https://doi.org/10.1177/0003122419877135.

Le Mens, Gaël, Balázs Kovács, Michael T. Hannan, Cecilia Nunes, and Guillem Pros. 2020. "How Do Categories Affect Valuation?" Working paper.

Lewis, Molly, and Gary Lupyan. 2020. "Gender Stereotypes Are Reflected in the Distributional Structure of 25 Languages." *Nature Human Behaviour* 4:1021–28. https://doi.org/10.1038/s41562-020-0918-6.

Maron, Melvin Earl. 1961. "Automatic Indexing: An Experimental Inquiry." *Journal of the ACM* 8(3):404–17. https://doi.org/10.1145/321075.321084.

Mohr, John W., Christopher A. Bail, Margaret Frye, Jennifer C. Lena, Omar Lizardo, Terrence E. McDonnell, Ann Mische, Iddo Tavory, and Frederick F. Wherry. 2020. *Measuring Culture*. New York: Columbia University Press.

Nangia, Nikita, and Samuel R. Bowman. 2019. "Human vs. Muppet: A Conservative Estimate of Human Performance on the GLUE Benchmark." Preprint, arXiv:1905.10425 [cs.CL]. https://arxiv.org/abs/1905.10425.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. "GloVe: Global Vectors for Word Representation." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–43. Association for Computational Linguistics. http://www.aclweb.org/anthology/D14-1162.

Pontikes, Elizabeth G. 2008. *Fitting In or Starting New? An Analysis of Invention, Constraint, and the Emergence of New Categories in the Software Industry*. Ph.D. thesis, Stanford University.

Pontikes, Elizabeth G. 2022. "Category Innovation in the Software Industry: 1990–2002." *Strategic Management Journal* 43(9):1697–727. https://doi.org/10.1002/smj.3383.

Pontikes, Elizabeth G., and Michael T. Hannan. 2014. "An Ecology of Social Categories." *Sociological Science* 1:311–43. 10.15195/v1.a20.

Porac, Joseph F., Howard Thomas, Fiona Wilson, Douglas Paton, and Alaina Kanfer. 1995. "Rivalry and the Industry Model of Scottish Knitwear Producers." *Administrative Science Quarterly* 40(2):203–27. https://doi.org/10.2307/2393636.

Rosch, Eleanor H. 1973. "On the Internal Structure of Perceptual and Semantic Categories." In *Cognitive Development and the Acquisition of Language*, edited by T. E. Moore, pp. 111–44. New York: Academic Press.

Ruef, Martin. 2000. "The Emergence of Organizational Forms: A Community Ecology Approach." *American Journal of Sociology* 106(3):658–714. https://doi.org/10.1086/318963.

Sanborn, Adam N., Thomas L. Griffiths, and Richard M. Shiffrin. 2010. "Uncovering Mental Representations with Markov Chain Monte Carlo." *Cognitive Psychology* 60(2):63–106. https://doi.org/10.1016/j.cogpsych.2009.07.001.

Schöll, Nikolas, Aina Gallego, and Gaël Le Mens. 2023. "How Politicians Learn from Citizens' Feedback: The Case of Gender on Twitter." *American Journal of Political Science*, forthcoming.

Smith, Edward Bishop. 2011. "Identities as Lenses: How Organizational Identity Affects Audiences' Evaluation of Organizational Performance." *Administrative Science Quarterly* 56(1):61–94. https://doi.org/10.2189/asqu.2011.56.1.061.

Vaisey, Stephen. 2021. "Welcome to the Real World: Escaping the Sociology of Culture and Cognition." *Sociological Forum* 36(S1):1297–315. https://doi.org/10.1111/socf.12770.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008.

Vincianza, Paul, Amir Goldberg, and Sameer B. Srivastava. 2020. "Who Sees the Future? A Deep Learning Language Model Demonstrates the Vision Advantage of Being Small." Working Paper No. 3869, Stanford Graduate School of Business.

Vogel, Tobias, Evan W. Carr, Tyler Davis, and Piotr Winkielman. 2018. "Category Structure Determines the Relative Attractiveness of Global versus Local Averages." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 44(2):250–67. https://doi.org/10.1037/xlm0000446.

Wan, Mengtin, and Julian J. McAuley. 2018. "Item Recommendation on Monotonic Behavior Chains." In *RecSys '18: Proceedings of the 12th ACM Conference on Recommender Systems*, edited by Sole Pera, Michael D. Ekstrand, Xavier Amatriain, and John O'Donovan, pp. 86–94. ACM. https://doi.org/10.1145/3240323.3240369.

Wan, Mengting, Rishabh Misra, Ndapa Nakashole, and Julian J. McAuley. 2019. "Fine-Grained Spoiler Detection from Large-Scale Review Corpora." In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, edited by Anna Korhonen, David R. Traum, and Lluís Màrquez, pp. 2605–10. Association for Computational Linguistics. https://doi.org/10.18653/v1/p19-1248.

Zuckerman, Ezra W. 1999. "The Categorical Imperative: Securities Analysts and the Legitimacy Discount." *American Journal of Sociology* 104(5):1398–438. https://doi.org/10.1086/210178.

Acknowledgments: We are grateful to Jerker Denrell, Amir Goldberg, Greta Hsu, Thorbjørn Knudsen, Cecilia Nunes, and Phanish Puranam for discussion of ideas developed in this article and for the detailed feedback we received from them on the earlier versions. We thank conference participants at the 2021 and 2022 Nagymaros Conferences for valuable feedback and discussion. G. Le Mens and G. Pros received financial support from ERC Consolidator Grant #772268 from the European Commission. G. Le Mens also received financial support from grant PID2019-105249GB-I00/AEI/10.13039/501100011033 from the Spanish Ministerio de Ciencia, Innovacion y Universidades (MCIU) and the Agencia Estatal de Investigacion (AEI) and from the BBVA Foundation Grant G999088Q. B. Kovács was supported by Yale School of Management. M. Hannan was supported by the Stanford Graduate School of Business. Data, material, and analysis code for all analyses are available online at https://osf.io/ta273/. We encourage readers to download the shared folder and use the code to compute *BERT typicality* on their own data sets.

Gaël Le Mens: Department of Economics and Business, Universitat Pompeu Fabra (UPF), Barcelona School of Economics, and UPF Barcelona School of Management, Barcelona, Spain. E-mail: gael.le-mens@upf.edu.

Balázs Kovács: School of Management, Yale University, New Haven, CT, USA. E-mail: balazs.kovacs@yale.edu.

Michael T. Hannan: Graduate School of Business, Stanford University, Stanford, CA, USA. E-mail: hannan@stanford.edu.

Guillem Pros: Department of Economics and Business, Universitat Pompeu Fabra, Barcelona, Spain. E-mail: guillem.pros@upf.edu.