Fang et al. *EPJ Data Science* (2022) 11:39
https://doi.org/10.1140/epjds/s13688-022-00353-7

EPJ Data Science
a SpringerOpen Journal

EPJ.org

REGULAR ARTICLE Open Access

Evaluating the construct validity of text embeddings with application to survey questions

Qixiang Fang[1*], Dong Nguyen[2] and Daniel L. Oberski[1,3]

*Correspondence: q.fang@uu.nl
[1]Department of Methodology & Statistics, Utrecht University, Padualaan 14, Utrecht, The Netherlands
Full list of author information is available at the end of the article

Abstract

Text embedding models from Natural Language Processing can map text data (e.g. words, sentences, documents) to meaningful numerical representations (a.k.a. text embeddings). While such models are increasingly applied in social science research, one important issue is often not addressed: the extent to which these embeddings are high-quality representations of the information needed to be encoded. We view this quality evaluation problem from a measurement validity perspective, and propose the use of the classic construct validity framework to evaluate the quality of text embeddings. First, we describe how this framework can be adapted to the opaque and high-dimensional nature of text embeddings. Second, we apply our adapted framework to an example where we compare the validity of survey question representation across text embedding models.

Keywords: Word embeddings; Sentence embeddings; Measurement validity; Content validity; Convergent validity; Discriminant validity; Predictive validity; Survey questions; Survey methodology; Computational social science

1 Introduction

Text embedding models from Natural Language Processing (NLP) can map texts (e.g. words, sentences, articles) to supposedly semantically meaningful, numeric vectors (i.e. embeddings) with typically a few hundred or thousand dimensions (e.g. [1, 2]). Intuitively speaking, this means that the embeddings of similar texts (e.g. words like "big" and "large") would be closer to one another than those of dissimilar texts (e.g. "big" and "paper") in the vector space.

Such models are often *pretrained* on an enormous amount of text data (e.g. Wikipedia, websites, news) and made publicly available (e.g. [1–4])). This allows researchers to obtain off-the-shelf pretrained text embeddings for downstream applications, without the need to spend many computational resources on training the models from scratch. Researchers can also choose to further train the text embedding models on additional task-specific data (i.e. *fine-tuning*) or domain-specific data (i.e. *continued pre-training*) for better performance.

Springer

Because of their capability for meaningful text representation and convenient use, text embedding models have attracted various applications in (computational) social science. For instance, they have been employed to encode *the Big-Five personality questionnaire* for personality trait prediction [5], *social media posts* for suicide risk assessment [6], *Tweets and TV captions* for emotion detection [7], and *relevant terms* (e.g. gender pronouns, names, occupations) to quantify societal trends of gender and ethnic stereotypes in the US [8].

Despite the popularity of text embedding models, an important question to ask is: how good are the text representations, given a research task? High-quality representations are desirable because they encode relevant information, which leads to greater generalization capabilities [9]. Typically, the quality of text embeddings is evaluated by established benchmarks like GLUE [10] and SentEval [9]. GLUE is a collection of common natural language understanding (NLU) tasks like question answering, sentiment analysis and textual entailment recognition. Similarly, SentEval consists of a broad and diverse set of NLU tasks, such as sentiment analysis, semantic textual similarity, natural language inference, and verb tense prediction.

While such benchmarks are useful for evaluating an embedding model's capability for certain NLU tasks, they do not necessarily inform researchers about whether an embedding model is suitable for a specific downstream task at hand. In computational social science research, the goal of using embedding models is often to tackle a substantive question (e.g. personality prediction, bias detection) that requires the embedding models to be capable of handling tasks that are different from or not directly addressed by the tasks in existing benchmarks. Moreover, the type of text data can also be different from the evaluation data sets in those benchmarks. For instance, in [5], the authors used embedding models to encode personality questionnaires and social media posts for personality prediction. This requires that the embedding models should be able to encode *personality-relevant* information from *questionnaire* texts and *social media* text data. However, knowing how well an embedding model performs many of the NLU tasks (e.g. natural language inference, verb tense prediction) on other types of data (e.g. movie reviews, image captions) do not tell us much about whether this embedding model is appropriate for the task of personality prediction.

Therefore, we need an additional framework that can help to guide the quality evaluation of text embedding models for a specific research task, which benefits model selection, debugging and improvement. Concretely, this framework should evaluate whether an embedding model *encodes what it is supposed to* for the research task. Furthermore, it should be easily adaptable to different research tasks.

In this work, we propose such a framework based on the classic *construct validity testing* framework widely used in the social sciences. In measurement theory, *construct validity* concerns whether the measure of a construct "measures what it is supposed to". Constructs are theoretical, latent variables (e.g. personality traits, emotions and biases). They are not directly observable and hence, need to be translated into something measurable. This process is called operationalization, which results in some concrete measure for the construct of interest [11]. For instance, the emotion "anger" can be translated into actionable measures like one's facial expression, one's tone of voice or a simple question "how angry are you?". However, the operationalization procedure inevitably introduces some degree of incongruence between the intended construct and the measure. When this incongruence is

high, we say the construct validity of a measure is low and vice versa. Needless to say, the construct validity of a measure needs to be checked and established before it is adopted in research. To achieve this, social scientists resort to an established construct validity testing framework, which involves examining various types of construct validity such as face validity, content validity, convergent validity, discriminant validity and predictive validity [11]. This testing framework is grounded in measurement theory and is the standard approach to testing the validity of a measure in the social sciences.

That an embedding model should "encode what it is supposed to" is comparable to the idea of construct validity (i.e. a measure should measure what it is supposed to). Specifically, we can treat what an embedding model needs to encode as the construct of interest, the embedding model as the measure, and the embeddings as the measurements. This allows us to approach the quality evaluation of text embedding models from the perspective of construct validity testing.

Thus, in this paper, we propose using the construct validity testing framework to evaluate the quality of text embedding models, given a research task. This framework has the advantages of being theory-driven and having established testing procedures. It is also easily adaptable to different research tasks. Furthermore, we focus on the validity of the text embeddings (rather than model predictions), because this can help researchers choose more "valid" models, which encode information relevant for the task at hand. This can help to lower the risk of overfitting due to the model learning from noises.

This paper has three main parts. First, we review several popular text embedding models in Sect. 2. Second, in Sect. 3, we describe the classic construct validity testing framework and how it can be adapted to the quality evaluation of text embedding models, given a specific research task. The main challenge is that established procedures for construct validity analysis only apply to low-dimensional measurements (e.g. responses to survey questions), while text embeddings are much higher-dimensional. We show how to adapt the classic construct validity framework to high-dimensional settings by, for instance, borrowing tools from interpretable NLP. Finally, in Sect. 4, we apply our adapted construct validity testing framework to a concrete application, where we show how to evaluate the quality of embedding models when they are used to encode survey questions.

2 Background: text embeddings

In this section, we review several popular text embedding models, including word2vec, fastText, GloVe, BERT, Sentence-BERT and Universal Sentence Encoder. We are aware of other (near) state-of-the-art text embedding models, such as GPT-3 [12], ALBERT [13] and XLNet [14]. Nevertheless, it is not our goal to cover all text embedding models.

2.1 word2vec and fastText

One influential family of text embeddings algorithms is word2vec [1, 15]. Simply put, word2vec is a two-layer neural network model that takes as input a large corpus of text and gives as output a vector space. This vector space has typically several hundred dimensions (e.g. 300), with each unique word in the training corpus being assigned a corresponding continuous vector. Such a vector is also called an embedding. The objective of the algorithm is to predict words from surroundings words (or the other way around). In this way, the final word vectors are positioned in the vector space such that words sharing common contexts in the corpus are located closely to one another. Under the Distributional

Hypothesis [16, 17], which states that words occurring in similar contexts tend to have similar meanings, closely located words in the vector space are expected to be semantically similar.[1] The similarity between two word vectors can be measured by cosine similarity. Mathematically, it is simply the cosine of the angle between two vectors, which can be calculated as the dot product of the two vectors divided by the product of the lengths of the two vectors. Cosine similarity scores are bounded in the interval $[-1, 1]$, where -1 indicates complete lack of similarity while 1 suggests the other extreme.

Word2vec has been shown to produce text representations that capture syntactic and semantic regularities in language, in such a way that vector-oriented reasoning can be applied to the study of word relationships [19]. A classic example is that the male/female relationship is automatically learned in the training process, such that a simple, intuitive vector operation like "King – Man + Woman" would result in a vector very close to that of "Queen" in the vector space [19].[2] Many studies have also made use of this characteristic to study human biases (e.g. gender and racial bias) encoded in texts [8, 21–23].

One popular extension of word2vec is fastText [3], which is trained on subwords in addition to whole words. This allows fastText to estimate word embeddings even for words unknown to the training corpora. fastText was shown to outperform its word2vec predecessors across various benchmarks [24].

2.2 GloVe

GloVe, which stands for Global Vector (for word representations) [25], is another popular text embedding model. Similar to word2vec, GloVe also produces word representations that capture syntactic and semantic regularities in language. However, a major difference is that GloVe is trained on a so-called global word-word co-occurrence matrix, where matrix factorisation is used to learn word embeddings of typically 25 to 300 dimensions.

2.3 BERT and sentence-BERT

More sophisticated embedding models have recently become available. A prominent one is BERT, which stands for Bidirectional Encoder Representations from Transformers [4]. Like word2vec and fastText, BERT is a type of neural networks trained on a large text corpus in order to learn a good representation of natural language. Notably, BERT produces context-dependent embeddings. That is, while word2vec, fastText and GloVe models produce a fixed embedding for each word, BERT can produce different embeddings for the same word depending on the context (e.g. the neighbouring words). For instance, the word "bank" can have different meanings (e.g. a financial establishment or land alongside a river) or function (i.e. as a noun or verb), across linguistic contexts.

BERT has achieved state-of-the-art performance on various natural language tasks [4]. BERT embeddings have also been shown to encode syntactic and semantic knowledge about the original texts [26].

Sentence-BERT [2], an extension of BERT, differs from the original BERT in that its architecture is optimised for generating semantically meaningful sentence embeddings that can be compared using cosine similarity [2].

[1] However, as widely applicable as it is, the Distributional Hypothesis' reliance on word contexts can be limiting. For instance, Parasca et al. [18] pointed out that word contexts alone cannot capture implicit or prior knowledge.

[2] Note that this is a somewhat outdated example, as one reviewer pointed out. For instance, Linzen [20] showed that using vector operations to study word relationships is not always warranted.

2.4 Universal sentence encoder

Universal Sentence Encoder (USE) is another text embedding model meant for greater-than-word length texts like sentences, phrases and short paragraphs [27]. USE uses both a Transformer model and a Deep Averaging Network model. The former focuses on achieving high accuracy despite suffering from greater resource consumption and model complexity, while the latter targets efficient inference at the cost of slightly lower accuracy [27]. It is trained on various language understanding tasks and data sets, with the goal to learn general properties of sentences and thus produce sentence-level embeddings that should work well across various downstream tasks. Pretrained USE embeddings have been shown to outperform word2vec-based pretrained embeddings across different language tasks.

3 Construct validity testing

In this section, we first introduce the classic construct validity testing framework. Then, we describe how this framework can be adapted for the quality evaluation of text embedding models.

3.1 The classic framework

As mentioned in the introduction, construct validity concerns the extent to which a measure matches the intended construct in theory. In practice, evaluating the construct validity of a measure means answering questions like: Does the measure capture all relevant aspects of the construct? Do the measurements correlate with other measurements of the same construct, or the measurements of similar constructs? Do the measurements differ from measurements of unrelated constructs? Each of these questions focuses on a particular aspect of construct validity.

Therefore, construct validity testing involves carefully thinking about the relevant aspects of construct validity that apply (i.e. need investigation) and designing the appropriate tests to evaluate those aspects. The literature often refers to such aspects as the (sub)types of construct validity [11]. We describe below the most common and important ones, as well as how they are usually tested. They are: face validity, content validity, convergent validity, discriminant validity, predictive validity and concurrent validity. We use the example of measuring *emotional intelligence* (EI) as a recurring example.

3.1.1 Face validity

Face validity concerns checking whether the measure "on its face" seems like a good operationalization of the construct. For example, EI was defined by Goleman [28] as consisting of five domains: self-awareness (i.e. knowing one's emotions), self-regulation (i.e. managing one's emotions), motivation (i.e. motivating oneself), empathy (i.e. recognizing emotions in others), and social skills (i.e. handling relationships). If a self-report questionnaire is designed to measure this definition of EI, then the questions in the questionnaire should appear to reflect the definition in the eyes of, for instance, experts on this topic.

3.1.2 Content validity

Content validity concerns whether a measure adequately covers all relevant aspects of the intended construct. For example, per the earlier definition of EI, the questionnaire should consist of questions that measure all the five domains of EI. To judge the content validity of such a test, expert opinions can be requested. Often, researchers also ask a sample

of individuals to provide responses to the questionnaire. Then, they use latent variable models like factor analysis to test whether the data supports the theoretical structure of the construct: What is the most likely number of EI domains based on the data? Do those questions intended to measure the same domain also fall under the same and only that domain? Are there questions that do not belong to any domain?

3.1.3 Convergent and discriminant validity

Convergent validity concerns whether the measure of a construct is similar to other measures that it should be similar to. For instance, say there is already an established questionnaire-based measure of EI, but we wish to develop a shorter version of it (to save time and costs associated with filling out the questionnaire). Then, given that the same people respond to both the long and the short questionnaires, the responses to one version (e.g. aggregated scores on an EI domain) should highly correlate with the responses to the other.

In contrast, discriminant validity concerns whether the measure of a construct is different from other measures that it should not be similar to. For instance, EI was proposed to be distinct from constructs like IQ and personality traits. Therefore, measurements of EI should not highly correlate with IQ or personality traits.

3.1.4 Predictive validity

Predictive validity concerns the ability of a measure to predict some future target it should be able to. For instance, Goleman [28] claimed that per his theory about EI, EI is crucial for success in life. Then, EI measurements should be able to predict one's future performances in, for example, career and academics.

3.2 Construct validity testing for text embedding models

To apply the idea of construct validity to text embedding models, we can view what an embedding model needs to encode as the construct of interest, the embedding model as the measure, and the embeddings as the measurements. Then, similar to the regular way of construct validity testing, we need to think about the relevant aspects of construct validity to evaluate. However, because text embeddings are high-dimensional data without interpretable labels for the dimensions, we need to adapt the testing procedures described earlier accordingly, which we describe next.

3.2.1 Face validity

We can evaluate the face validity of a text embedding model by looking at whether some fundamental aspects of the model (e.g. architecture, training data) are suitable for the task at hand. For instance, if the task is to predict hate speech from Tweets, a sentence embedding model is likely more valid than a word embedding model; a model trained on Twitter data is also likely more valid than one trained on Wikipedia entries.

3.2.2 Content validity

Text embeddings are high-dimensional and opaque, which makes it difficult to find out what information is encoded in them. We therefore borrow an approach from interpretable NLP: *probing classifiers*. The idea is to train a classifier that takes text representations as input and predicts some property of interest (e.g. sentence length). If the classifier

performs well, this suggests that the representations encode information relevant to the property [29].

A recommended practice in choosing a classifier is to select a linear model like (multinomial) logistic regression, because a more complex probing classifier may run the risk of inferring properties not actually present in the text representations [30–34]. Furthermore, it is recommended to always include baselines for comparison [29]. The better the probing classifier based on some text representation performs relative to the baselines, the more evidence that the probed property is present and that content validity is supported. Previous studies [35–38] suggest using two forms of baselines: simple majority in the training data and random embeddings.

3.2.3 Convergent and discriminant validity

We define convergent validity as the extent to which a reference embedding is similar to some embedding that it is supposed to be similar to. To investigate convergent validity, we introduce perturbations to a given text (i.e. the reference text) without changing its label or core meaning. For instance, if the research task is to detect hate speech, then a text considered as hate speech should not be modified in a way that would render it no longer hate speech. High cosine similarity between the embeddings of the reference and the modified text indicates convergent validity.

Likewise, we define discriminant validity as the extent to which a reference embedding is dissimilar to some embedding that it is not supposed to be similar to. To investigate discriminant validity, we introduce perturbations to the reference text in such a way that the new text is substantially different from the reference text (e.g. that would lead to changing the label or the core meaning). Accordingly, we would expect low cosine similarity between the two embeddings, which would then indicate discriminant validity.

In the social sciences, there are guidelines to interpret the sizes of correlations for convergent and discriminant validity testing. However, there are no such agreed-upon rules for cosine similarity in NLP. Therefore, it is difficult to say how high or low cosine similarity scores need to be to provide sufficient evidence for convergent or discriminant validity. To partly mitigate this problem, we propose using the difference between the two cosine similarity scores (i.e. one from convergent validity analysis and the other from discriminant validity analysis) as a joint indicator for convergent and discriminant validity. This provides the benefit of having a natural baseline: zero. If the difference score is above zero, it provides some degree of support for both convergent and discriminant validity.

Our proposed testing procedure for convergent and discriminant validity shares some commonality with several existing model testing methods in NLP, including robustness tests [39], adversarial changes [40], invariance checks and directional expectation tests [41]. However, they focus on evaluating model predictions, while we focus on text representations.

3.2.4 Predictive validity

How well an embedding model performs the research task of interest already provides information about the predictive validity of the model. In addition, we can test how well the embeddings can perform tasks that are related to the research task but not part of it. For instance, if an embedding model is designed to detect hate speech in Tweets, then the embeddings should also be able to predict whether a Tweet gets reported later. If an

embedding model is used for political stance detection, then the embeddings should also be able to predict a person's party affiliation or their future voting behaviour.

4 Application: embedding models for survey questions

There is growing research interest in using embedding models to encode subjective survey questions. Such questions are subjective because they aim to measure information that only exists in the respondent's mind (e.g. opinions, feelings). For instance, Vu et al. [5] used pretrained BERT to encode participants' social media posts and the questions from the Big-Five personality questionnaire, to predict individual-level responses to out-of-sample Big-Five questions. Pellegrini et al. [42] used the skip-gram embedding algorithm [1] to encode psychiatric questionnaires and patients' responses to those questionnaires. They showed that the resulting embeddings can be used for effective diagnosis of some mental health conditions.

However, to the best of our knowledge, there is no work on evaluating whether text embedding models can provide valid representations of subjective survey questions. In this section, we demonstrate how our proposed construct validity testing procedures can be used to investigate and compare the validity of text embedding models for representing subjective survey questions. Nevertheless, because survey questions exist in various forms and are used differently across research fields, we are unable to cover all kinds of survey questions in our analysis. In this work, we focus on the ones that are more typical in sociological research.

We begin an overview of the embedding models we use (Sect. 4.1.1), and describe how to generate sentence-level embeddings (Sect. 4.1.2). Then, we describe the data that we use for the analysis, with a focus on the synthetic data set (Sect. 4.2). Next, we present our analysis of face validity (Sect. 4.3), content validity (Sect. 4.4), convergent and discriminant validity (Sect. 4.5), as well as predictive validity (Sect. 4.6).

4.1 From survey questions to pretrained sentence embeddings

4.1.1 Pretrained embedding models

We focus on the following embedding models: fastText, GloVe, BERT and USE. We use their pretrained versions (as opposed to fine-tuned models) because of their widespread and convenient use. However, our approach to construct validity analysis also applies to fine-tuned text embeddings and other vector representations of texts.

Table 1 lists the pretrained embedding models we adopt. For fastText, we use the pretrained model developed by [3]. It is trained on Common Crawl with 600B tokens, and produces word embeddings with 300 dimensions for 2M words. For GloVe, we use the model pretrained on Common Crawl with 840B tokens and a vocabulary of 2.2M words. It also outputs word vectors of 300 dimensions.

Table 1 Overview of Pretrained Text Embedding Models Investigated in this Study

Model	Name	Dimension	File size
fastText	cc.en.300.bin	300	2.44 GB
GloVe	glove.840B.300d	300	2.03 GB
BERT	BERT-base-uncased	768	420 MB
BERT	BERT-large-uncased	1024	1.25 GB
Sentence-BERT	All-DistilRoBERTa-V1	768	292 MB
Sentence-BERT	All-MPNet-base-V2	768	418 MB
USE	USE-V4	512	916 MB

As for pre-trained Sentence-BERT models, there are many to choose from, which differ not only in the specific natural language tasks that they have been optimised for, but also in their model architecture. We select two pretrained models which have been trained on various data sources (e.g. Reddit, Wikipedia, Yahoo Answers; over 1B pair of sentences) and are thus designed as general purpose models [2]. They are "All-DistilRoBERTa" and "All-MPNet-base", where "DistilRoBERTa" [43, 44] and "MPNet" [45] are two different extensions of the original BERT. "Base" indicates that the embedding dimension is 768, as opposed to "Large" where the dimension is 1024. Both "All-DistilRoBERTa" and "All-MPNet-base" have been shown to have the top average performance across various language tasks. For the purpose of comparison, we also include two pretrained models of the original BERT model [4]: "BERT-base-uncased" and "BERT-large-uncased". "Uncased" refers to BERT treating upper and lower cases equally. Both models have been trained on Wikipedia (2.5B words) and BookCorpus (800M words).

As for USE, we use the most recent (i.e. 4th) version of its pretrained model, which outputs a 512 dimensional vector given an input sentence. The sources of training data are Wikipedia, web news, web question-answer pages, discussion forums and the Stanford Natural Language Inference corpus [46].

4.1.2 Sentence-level embeddings

For survey questions, we need to obtain sentence-level representations (as opposed to word-level). Namely, we treat a survey question as one sentence and represent it as a single embedding. Obtaining sentence-level embeddings is straightforward with Sentence-BERT and USE models, because they have been designed for this specific purpose. In contrast, word2vec and GloVe models only produce word embeddings. When using these models, it is therefore necessary to combine the word embeddings into a sentence-level representation. Among various methods, simple averaging across all the word embeddings (e.g. taking the means along each dimension) has been shown to either outperform other methods [47] or approximate the performance of more sophisticated ones [48]. Therefore, we use simple averaging to compute sentence-level embeddings for survey questions from fastText and GloVe word embeddings. The resulting embeddings have the same number of dimensions as the word-level embeddings, as we average the word embeddings along each dimension. However, one disadvantage of this approach is that information like word order is absent in the aggregated representation.

As for the original BERT models, we follow the advice of [2, 26] to average the word embeddings produced at the last layer of BERT to form sentence-level embeddings. This way, the resultant sentence-level representation has the same dimension as that of the word embeddings.

4.2 Data sets

4.2.1 Synthetic data set

For the analysis of content validity, convergent validity and discriminant validity, we use a data set of subjective survey questions which we generate by ourselves (hence the name "synthetic"). The reason for using a synthetic data set over existing survey questions is two-fold. First, because we would like to study how well a text embedding model encodes subjective survey questions in general, it is important that the questions we study cover a diverse range of concepts and formulations. Existing questionnaires like the European

Social Surveys only concern a small set of concepts and formulations. Second, for the analysis of convergent and discriminant validity, we need to introduce small perturbations to a reference survey question to create similar and dissimilar survey questions. This is difficult when the survey questions themselves are problematic, which is the case for many existing survey questions in use. For instance, many surveys use this question to measure generalized trust: "Generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people?". The problem with this question is that it is unclear what this question is measuring: trust or caution [49]? This makes it difficult to make variants of the question that are similar or different in terms of the underlying concept. Using synthetic questions, in contrast, allows us to focus on high-quality questions that are also easier to vary.

Therefore, we create a synthetic data set of survey questions for the analysis of content, convergent and discriminant validity. To generate a diverse set of high-quality survey questions, we follow the three-step procedure described in [50]. The full details are described in Appendix A. We summarize below three main characteristics of this data set.

First, we focus on covering a wide selection of survey questions. According to Saris and Gallhofer [50], subjective survey questions normally fall under one (or more) of the following 13 basic concepts: "evaluation", "importance", "feelings", "cognitive judgment", "causal relationship", "similarity", "preferences", "norms", "policies", "rights", "action tendencies", "expectation", and "beliefs" (see Appendix B for definitions and examples of these 13 basic concepts). The questions in our data set therefore cover these 13 basic concepts.

Second, for every subjective concept, we assign three reference concrete concepts. Take the basic concept "evaluation" as an example: we specify "the state of health services", "the quality of higher education" and "the performance of the government" as the three corresponding reference concrete concepts. Next, for every reference concrete concept, we specify one similar concrete concept and one dissimilar concrete concept (see Appendix C for a complete list of concrete concepts). Finally, for each concrete concept, we create survey questions that vary in their formulation. [50] provided many templates for each type of formulation. We adopt 19 templates (see Appendix D) and thus create differently formulated survey questions for each concrete concept. Our final data set contains 5436 unique subjective survey questions.

Table 2 shows six example questions from the data set. They all fall under the basic concept "evaluation". The main concrete concept here is evaluation about "the state of

Table 2 Example Questions from Our Survey Question Data Set. InDe: interrogative-declarative request. DR: direct request

ID	Concrete concept	Similarity	Formulation	Survey question
1	state of health services	reference	DR	How good is the state of health services in your country?
2	state of health services	reference	InDe	Do you agree that the state of health services in your country is good?
3	state of medical services	high	DR	How good is the state of medical services in your country?
4	state of medical services	high	InDe	Do you agree that the state of medical services in your country is good?
5	state of religious services	low	DR	How good is the state of religious services in your country?
6	state of religious services	low	InDe	Do you agree that the state of religious services in your country is good?

health services", while the corresponding similar and dissimilar concepts are evaluation about "the state of medical services" and "the state of religious services". Each concrete concept has two differently formulated questions in the table: DR (i.e. direct request) and InDe (i.e. interrogative-declarative request).

4.2.2 European social survey wave 9

For the analysis of predictive validity, we use the publicly available European Social Survey (ESS) Wave 9 data [51]. More information is available in Sect. 4.6.1.

4.3 Analysis of face validity

Survey questions are carefully curated sentences that are used to measure some concepts or constructs. This means that an embedding model should ideally: (1) provide sentence-level representations; (2) have been trained on texts similar to survey questions. Therefore, to examine the face validity of the embedding models (i.e. fastTest, GloVe, BERT, Sentence-BERT and USE), we can check whether they fulfil these two requirements.

With regard to the first aspect, only Sentence-BERT and USE can generate sentence-level embeddings by design. The other models require aggregating word-level embeddings to form sentence-level embeddings. However, because BERT can produce context-dependent embeddings, their representations of survey questions are likely of higher-quality than those of fastText and GloVe.

For the second aspect, none has been trained specifically on survey questions. However, all the models have been trained on common data sources (e.g. Common Crawl, Wikipedia and Reddit) which certainly contain some survey questions. Among them, Sentence-BERT models are additionally trained on many question answering data sets, which may allow them to represent survey questions better.

Taking both aspects into account, Sentence-BERT seems to show the highest face validity, followed by USE, then BERT, then fastText and GloVe.

4.4 Analysis of content validity

The analysis of content validity concerns whether text embeddings encode information about all relevant aspects of survey questions. Naturally, not all aspects are equally important, and we also cannot provide an exhaustive list of them. In this paper, we consider four such aspects.

The first aspect concerns the underlying *concepts*. According to [50], most subjective survey questions can be categorised into one of 13 so-called *basic concepts*, such as "feelings", "cognitive judgement" and "expectations" (see Appendix B). In addition to the basic concept, a survey question also has a *concrete concept*, such as "happiness" (under the basic concept "feelings") and "political orientation" (under "cognitive judgement") (see Appendix C).

Furthermore, survey questions can differ in terms of *formulation*. Specifically, five types of formulation often apply in survey research [50]: direct request (DR), imperative-interrogative request (ImIn), interrogative-interrogative request (InIn), declarative-interrogative request (DeIn) and interrogative-declarative request (InDe).[3] See Appendix D for examples of these different formulations.

[3] [50] mentioned one more formulation type: direct instruction, which does not apply to most survey questions concerning subjective basic concepts and is thus not considered in our study.

Lastly, *complexity* is another important aspect of survey questions which can affect how respondents understand and answer a survey question [52]. A simple proxy for complexity is the length of a survey question [50].

Therefore, we investigate whether text embeddings encode information about the following aspects of survey questions: basic concepts, concrete concepts, formulation and length. We refer to them as *properties* in the remainder of the paper.

4.4.1 Methods

We use probing classifiers for content validity analysis. Following [35–38], we include two baselines: simple majority in the training data and random embeddings. To generate random embeddings for each survey question, we randomly generate from a uniform distribution (-1,1) a unique fixed size embedding for each word in the training data. Then, we simply average the word embeddings along each dimension to derive sentence-level embeddings for the survey questions.

A common problem with probing classifiers is that the good performance of the classifier could simply be due to it making use of other properties present in the embeddings that are correlated with the properties of interest [53]. For instance, if we want to find out whether text embeddings encode information about basic concepts, our training data should differ only in terms of the probed property (i.e. basic concepts). In other words, for survey questions corresponding to a particular basic concept (e.g. "feelings"), the distribution of other properties should be similar to that of questions belonging to another basic concept (e.g. "expectation"). Otherwise, we cannot conclude that the performance of our classifier is explained by whether the text embeddings encode knowledge about basic concepts.

Unfortunately, with natural language data such as survey questions, it is extremely difficult, if not impossible, to construct a data set where properties like concepts, length and formulation are completely uncorrelated. To mitigate this issue, we need to construct our training and test sets in a way that they do not share the same distribution of the correlated properties. In this way, the probing classifier is less likely to make use of the correlated properties to achieve good performance on the test sets. We explain below how we achieve this.

In our data, we see that sentence length (as a four-level categorical variable, see Table 3) is highly correlated with all the other properties. Chi-square tests of independence show that sentence length is statistically significantly related to basic concepts ($\chi^2 = 3636.7$, $df = 36$, $p < 0.05$), concrete concepts ($\chi^2 = 4612.1$, $df = 348$, $p < 0.05$) and formulation ($\chi^2 = 1252.7$, $df = 12$, $p < 0.05$), with multiple testing corrected for. Therefore, when probing these properties (i.e. basic concepts, concrete concepts and formulation), we split the data into a training set and a test set of similar sizes, where the length of the survey questions in the training set is different from that in the test data (see Fig. 1 for illustration). Likewise, when probing sentence length, we make sure that our training and test data do not share the same concepts or formulation. Furthermore, when probing basic concepts, because concrete concepts are nested within basic concepts (and hence highly correlated), we make sure that the concrete concepts between the training set and the test set do not overlap.

Nevertheless, even separating the training and test set in terms of sentence length is not enough for effective probing of concrete concepts. We find that regardless of whether we use random embeddings or the actual text embeddings, the classifier always achieves

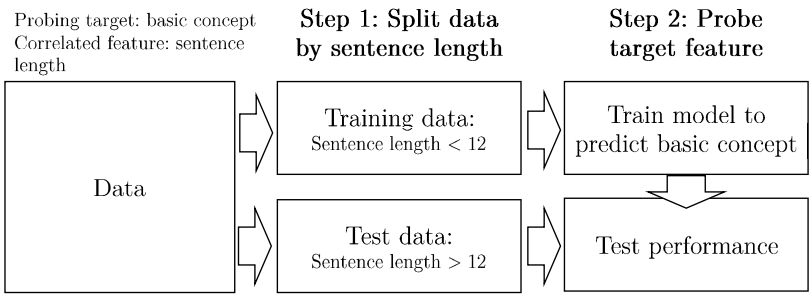

Figure 1 Illustration of Train-test Splitting for the Probing Analysis. Note that in this figure, the probing target is basic concept, while the correlated feature is sentence length

Table 3 Results of Content Validity Analysis: Prediction Accuracy Scores of Probing Classifiers. Note that sentence length is converted into a categorical variable with four levels including "0-10", "11-12", "13-15" and "16-25"; basic concept, concrete concept and formulation are also categorical with 13, 117 and 5 levels, respectively

	Length	Basic concept	Concrete concept	Formulation	Average
Simple Majority	0.389	0.010	0.029	0.255	0.171
Random 300	0.102	**0.198**	0.440	0.742	0.371
Random 768	0.148	**0.198**	0.509	0.694	0.387
Random 1024	0.074	**0.198**	0.548	0.731	0.388
TF	0.148	**0.198**	0.636	0.770	0.438
TF-IDF	0.167	**0.198**	0.493	0.690	0.387
fastText	0.093	0.173	0.711	0.656	0.408
GloVe	0.194	0.192	0.908	0.642	0.484
BERT-base-uncased	**0.657**	0.175	0.815	**0.944**	**0.648**
BERT-large-uncased	0.620	0.153	0.739	0.908	0.605
All-DistilRoBERTa	0.407	**0.198**	0.916	0.776	0.574
All-MPNet-base	0.481	**0.198**	**0.929**	0.805	0.603
USE	0.454	**0.198**	0.903	0.853	0.602

perfect performance on the test set. The absence of difference in performance prohibits us from concluding whether the embedding models encode information about concrete concepts to different extent. This is likely due to the fact that the prediction of concrete concepts may rely solely on the presence of certain words, which is a simple task and can be fully captured by even random embeddings. We therefore decided to increase the difficulty of the probing task for concrete concepts. Specifically, we made the classifier predict for a survey question its similar concrete concept (such as "the importance of achievement" and "the importance of success") (which we define in Sect. 4.2.1), while ensuring that the training set and the test set have not seen the exact same concrete concepts.

Using the probing approaches above, if we observe any positive difference between the performance of the probing classifier and that of the baseline using random embeddings, we can more confidently attribute it to the relevant survey question property being encoded in the text embeddings (on top of simple word-level information). In this way, we can learn about whether one text embedding model encodes more information about a property than does another model.

4.4.2 Results

Table 3 summarises the performance of the probing classifier (multinomial logistic regression) across different embedding models. The baseline classification accuracy scores are

based on simple majority voting, random embeddings of three dimension sizes, two simple count-based text representation approaches: term-frequency (TF) and term frequency-inverse document frequency (TF-IDF).

If the classifier performs better on a particular type of text embeddings than on the random embedding baselines for a survey question property, we can conclude that the corresponding text embeddings of survey questions likely encode information about that property.

For sentence length, we see that the BERT-based and the USE embeddings perform better than all the baselines, which suggests that they likely encode information about sentence length. Among them, both original BERT models have better performance than any of the Sentence-BERT models.

For basic concepts, none of the pretrained text embeddings seems able to beat the performance of the baseline random embeddings, TF and TF-IDF vectors. The fact that all text embeddings (including the random embeddings) have similar performance and perform better than the simple majority baseline suggests that only simple word-level information could be used by the classifier. A possible explanation is that the basic concepts as defined by [50] are too abstract information which are not explicitly encoded in the embeddings.

As for concrete concepts, all types of pretrained text embeddings perform better than the baselines. This suggests that text embeddings likely encode information about concrete concepts of survey questions. We also see that both Sentence-BERT embeddings (0.916 and 0.929) show better performance than do the original BERT embeddings (0.815 and 0.739). Furthermore, USE (0.903) and GloVe (0.908) have similarly good performance.

Lastly, we can see that random embeddings themselves can already achieve good prediction on the types of formulation, likely because single words are indicative of formulation. This holds true also for TF and TF-IDF. The random embeddings even outperform fastText and GloVe, despite the margin being relatively small. The original BERT representations, like with sentence length, perform the best again, suggesting that they encode sentence-level information about formulation. Both Sentence-BERT models and USE also perform better than the random baselines, however, only to a much smaller margin.

To conclude, we find that different text embedding models encode somewhat different kinds of information about survey questions and to different degrees. If we rank the importance of the properties of survey questions in the order of concepts, formulation and sentence length, then USE seems to demonstrate the highest level of content validity with regard to survey questions on average. The sentence-BERT and original BERT models quickly follow. FastText and GloVe as word embedding models do encode some information about survey questions like concrete concepts, but not sentence length or formulation.

4.5 Analysis of convergent & discriminant validity

We analyse convergent validity of text embeddings for survey questions as the extent to which the text embeddings of two conceptually similar survey questions are similar to each other. High convergent validity (as is desired) would be indicated by a high degree of similarity between the two text embeddings. In contrast, discriminant validity concerns the degree to which two conceptually dissimilar survey questions differ in their text embeddings. High discriminant validity (as is desired) is signalled by low similarity between the text embeddings.

4.5.1 Data

We use the same synthetic data set described in Sect. 4.2.1.

4.5.2 Methods

We take a joint approach to examining convergent and discriminant validity. That is, if text embedding models possess both convergent and discriminant validity, the embeddings of conceptually similar survey questions would be closer to one another while the embeddings of conceptually dissimilar survey questions would be further apart. Two hypotheses naturally follow:

With Hypothesis 1, we expect cosine similarity scores to be higher between the embeddings of *conceptually similar* survey questions than between those of *conceptually dissimilar* survey questions, with all other aspects of the survey questions being the same.

With Hypothesis 2, we expect cosine similarity scores to be higher between the embeddings of *conceptually identical but differently formulated* survey questions than between those of *conceptually dissimilar but identically formulated* survey questions.

In the synthetic survey question data set, we can find many pairs of survey questions that only differ in their concrete concepts and those that differ in their formulation but not in their concrete concepts.

For Hypothesis 1, we first calculate the cosine similarity between the embedding of a given survey question (i.e. $E_{\text{reference}}$) and the embedding of the corresponding conceptually similar question (i.e. E_{similar}). Then, we calculate the cosine similarity between $E_{\text{reference}}$ and the embedding of the corresponding conceptually dissimilar question (i.e. $E_{\text{dissimilar}}$). In this way, we obtain $\cos(E_{\text{reference}}, E_{\text{similar}})$ and $\cos(E_{\text{reference}}, E_{\text{dissimilar}})$. We expect the difference between these two scores (i.e. $\cos(E_{\text{reference}}, E_{\text{similar}}) - \cos(E_{\text{reference}}, E_{\text{dissimilar}})$) for a given survey question to be larger than zero. As an example, in Table 2, the two scores of interest are $\cos(E_{\text{ID1}}, E_{\text{ID3}})$ and $\cos(E_{\text{ID1}}, E_{\text{ID5}})$. Note that the two comparison questions differ from the reference question only in terms of the underlying concrete concepts; all other aspects like the formulation and sentence length are identical. This applies to all the question triads, which allows us to attribute any observed differences in similarity scores to the differences in the concrete concepts.

For Hypothesis 2, we first calculate the cosine similarity between the embedding of a given survey question (i.e. $E_{\text{reference}}$) and the embedding of the corresponding conceptually identical but differently formulated question (i.e. $E_{\text{identical}}$). Then, we calculate the cosine similarity between $E_{\text{reference}}$ and the embedding of the corresponding conceptually dissimilar but identically formulated question (i.e. $E_{\text{dissimilar}}$). In this way, we obtain $\cos(E_{\text{reference}}, E_{\text{identical}})$ and $\cos(E_{\text{reference}}, E_{\text{dissimilar}})$. We expect the difference between these two scores (i.e. $\cos(E_{\text{reference}}, E_{\text{identical}}) - \cos(E_{\text{reference}}, E_{\text{dissimilar}})$) for a given survey question to be larger than zero. In the exemplar Table 2, the two scores of interest are $\cos(E_{\text{ID1}}, E_{\text{ID2}})$ and $\cos(E_{\text{ID1}}, E_{\text{ID5}})$. Note that each comparison question differs from the reference question only in terms of one aspect: either concept or formulation.

4.5.3 Results

Figure 2 shows the distribution of the difference between $\cos(E_{\text{reference}}, E_{\text{similar}})$ and $\cos(E_{\text{reference}}, E_{\text{dissimilar}})$ scores for Hypothesis 1 across various baselines and text embedding approaches. The more positive the difference scores are, the more support for convergent and discriminant validity.

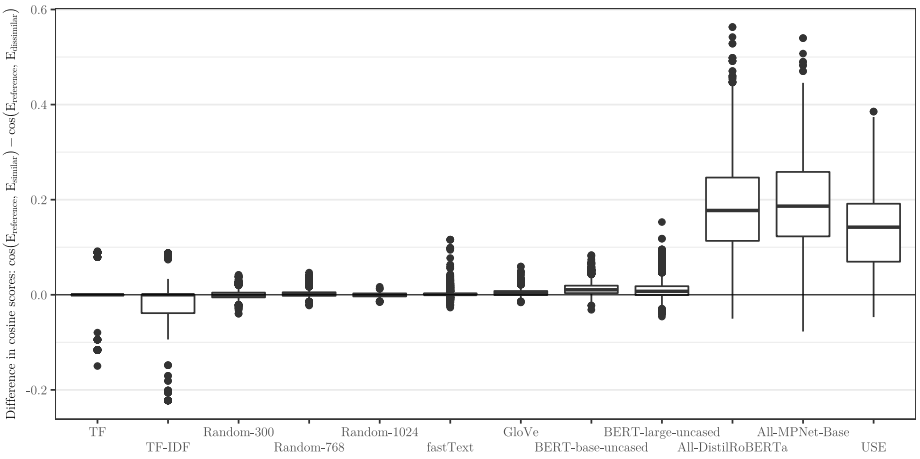

Figure 2 The distribution of the cosine similarity difference scores for Hypothesis 1 across text representation approaches. The *y*-axis indicates the size and direction of the differences. The more positive the difference scores are, the more support for convergent and discriminant validity

We can see in Fig. 2 that the only models that consistently score above zero are the two Sentence-BERT models ("All-DistilRoBERTa" and "All-MPNet-Base") and USE, with the percentages of positive scores being 98.3%, 96.8% and 95.4%, respectively. This result shows evidence of convergent and discriminant validity for the three models. Only in a small percentage of cases does this observation not hold. In stark contrast, none of the baselines models (i.e. TF, TF-IDF, random embeddings) show performance comparable to any of the Sentence-BERT and USE models. To our surprise, this observation holds also for fastText, GloVe and the two original BERT pretrained models, suggesting that these text embeddings lack convergent and discriminant validity. However, for the original BERT embeddings, one other possible explanation is that cosine similarity might not be a suitable measure, as earlier research suggested [2].

Figure 3 shows the distribution of the difference between $\cos(E_{\text{reference}}, E_{\text{identical}})$ and $\cos(E_{\text{reference}}, E_{\text{dissimilar}})$ scores for Hypothesis 2. Note that Jaccard similarity is explicitly included here as an additional baseline of similarity between two survey questions. It is calculated as the ratio of the number of unique overlapping words (i.e. intersection) to the total number of unique words between two survey questions (i.e. union). Naturally, Jaccard similarity scores are bounded in the interval [0, 1]. For Hypothesis 1, because any pair of comparison questions differs in only one word, the difference in Jaccard similarity for any pair of comparison questions would always be zero and therefore not a useful measure.

Similar to Fig. 2, the two Sentence-BERT models and USE again score consistently above zero (98.8%, 97.9% and 87.2% of the cases, respectively). Only in a few cases does this observation not hold (e.g. the concrete concept "petition institutional racism"). We can thus say that for these models, there is reasonable evidence for convergent and discriminant validity. Most of the other approaches (including the baselines, the random embedding models and the original BERT models) score either around or below zero. The only exception is TF-IDF, which in 94.5% cases scores above zero, suggesting evidence for convergent and discriminant validity. However, this conclusion should be treated with great caution, because when we generate the TF-IDF vectors, we build the vocabulary based on all the

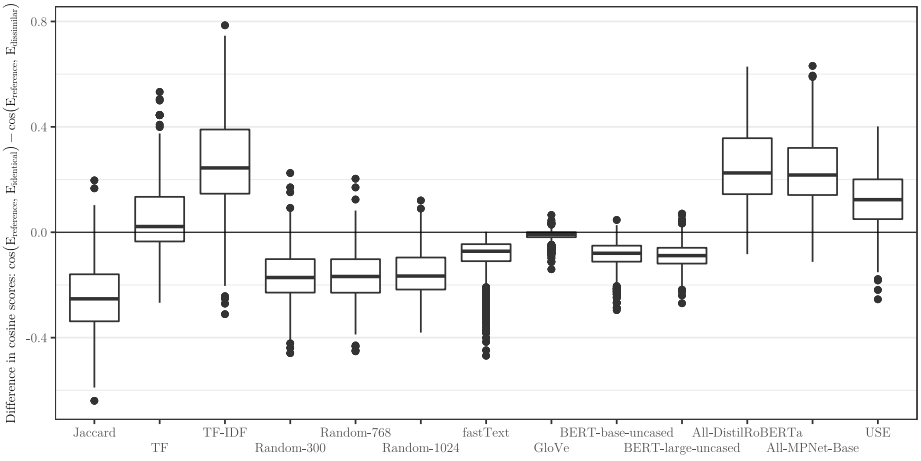

Figure 3 The distribution of the cosine similarity difference scores for Hypothesis 2 across text representation approaches. The *y*-axis indicates the size and direction of the differences. The more positive the difference scores are, the more support for convergent and discriminant validity

survey questions. We follow this approach because in this analysis, it is unclear what the training and testing data should be. In real research applications, TF-IDF may not perform as well due to the difference in the vocabulary between training and test data.

Finally, it is worth noting that the two Sentence-BERT models performed either about equally or better in Hypothesis 2 than in Hypothesis 1, in terms of the percentages of scores above zero. This is somewhat surprising considering that the task in the second hypothesis is supposedly more difficult because the survey questions differ in one extra aspect: formulation.

Overall, we can conclude that text embeddings of survey questions based on Sentence-BERT and USE demonstrate the highest convergent and discriminant validity. Meanwhile, there is not enough evidence to suggest the same for the other approaches.

4.6 Analysis of predictive validity

If embeddings are valid measures of survey questions, we should be able to use them to help with the prediction of people's responses to survey questions. For instance, knowing someone's response to a survey question about gender issues, we should be able to predict that person's response to a question about family values because the two are conceptually related. Also, knowing how someone responds to a particular question formulation may also be helpful in predicting how that person would respond to new questions of similar formulation. Therefore, we test the predictive validity of survey question embeddings by checking whether they can improve the prediction of a respondent's answer to new survey questions, compared to not using text embeddings. Specifically, we can inspect the correlation between the predicted responses and the actual responses. The higher the correlation (compared to some baseline), the more evidence for predictive validity.

4.6.1 Data: *European social survey wave* 9

We use the publicly available European Social Survey (ESS) Wave 9 data [51]. The ESS is a research-oriented cross-national survey that is conducted with newly selected, cross-sectional samples every two years since 2001 [54]. The survey aims to measure attitudes,

beliefs and behaviour patterns of diverse populations in Europe, concerning topics like media and social trust, politics, subjective well-being, human values and immigration. We focus on the UK sample (N = 2204), because the official language of the UK is English, which is consistent with the language of the data on which our text embedding models are pretrained.

Out of more than 200 questions that were asked to the participants, we select only the ones which measure subjective concepts and are continuously or ordinally scaled, totalling 94 questions. This choice is consistent with the type of survey questions we examined previously during the analysis of content, convergent and discriminant validity analysis. To harmonise the difference in response scales across the survey questions, we rescale all the responses to be between 0 and 1.

In addition to these 94 survey questions and the individual responses to each of them, we included the following background variables for each participant: region, gender, education, household income, religion, citizenship, birthplace, language, minority status, marital past and marital status.

The reason for including background variables is two-fold. First, background variables are predictive of many subjective opinions (like the ones measured by the ESS survey questions). Second, we also expect that different people (defined by their background characteristics) will interpret survey questions differently, which in turn influences their responses. This can be modelled by allowing the background variables to interact with the text embedding dimensions, which may help the model to achieve better prediction performance.

Note that we use existing survey questions here (despite some of them having quality problems like unclear underlying concepts), because for this predictive validity analysis, we need response data, which we cannot generate synthetically.

4.6.2 Methods

The data set includes two types of features: (1) various background variables and (2) text representations of survey questions. The prediction target concerns respondents' answers to survey questions. Every respondent answered all or most of the 94 survey questions. Each row in the data corresponds to a unique respondent and survey question combination.

To estimate the average performance of a prediction model, we apply 10-fold cross-validation [55]. Namely, we split the data into 10 folds by survey question IDs, whereby each data partition contains its own unique set of survey questions and responses (see Fig. 4 for illustration). Then, for every one out of the 10 cross-validation loops, we use 9 folds as the training set (covering 90% of the survey questions) and keep the remaining one fold as the test set. We train the prediction model on the whole training set, also using 10-fold cross-validation but for hyperparameter selection (i.e. fine-tuning). Once we have the fine-tuned model, we apply it to the test set and obtain an evaluation score. Because we apply 10-fold cross-validation to the entire data set, we obtain 10 evaluation scores in total. Finally, we average the 10 scores to arrive at the performance estimate of the prediction model.

Text representation methods Apart from the various embedding models, we include TF, TF-IDF and random embeddings of dimension 300, 768 and 1024 as baseline text repre-

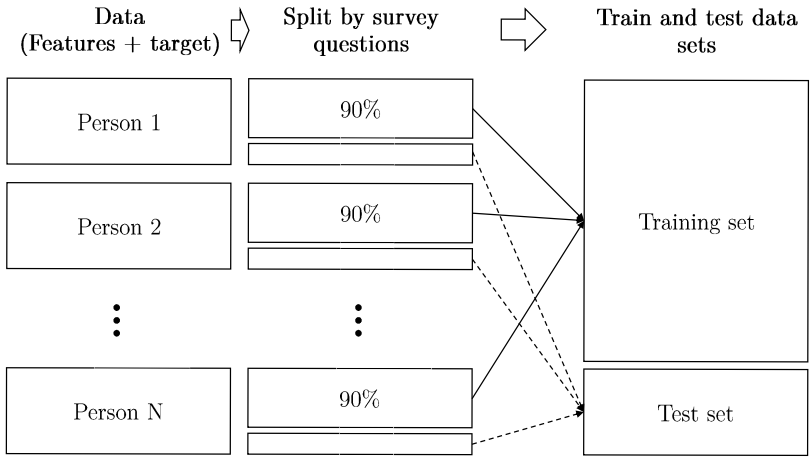

Figure 4 Illustration of Train-test Splitting for the Analysis of Predictive Validity

sentation methods. Note that with these baseline approaches, we build the feature vocabulary based only on the training data, similar to how we conducted the content validity analysis earlier. That is, new words encountered in the test set would be assigned zero weight.

Lasso and random forest We adopt two popular prediction models. The first is Lasso regression [56], which differs from OLS regression by including an additional regularisation term in the loss function. The regularisation term has the advantage of reducing model variance (i.e. lower prediction error, at the cost of slightly higher bias). Furthermore, it can zero-out the parameter estimates of those predictors considered by the model to be "unimportant", thus simplifying the model and easing interpretation.

The second model is Random Forest (RF) [57], which constructs multiple regression trees during training time and outputs the average prediction of all the trees. This approach of combining multiple models falls under the so-called ensemble learning technique, which generally provides the benefit of more powerful prediction. In addition, RF automatically considers interaction among the predictors, which Lasso regression falls short of. This may enable RF to learn more fine-grained patterns from data.

Evaluation metric We use Pearson's correlation r to evaluate the predictive validity of text embeddings. Specifically, we measure the Pearson's correlation between the predicted responses to survey questions and the observed responses. This metric has the advantage that it is bounded between −1 and 1, hence easier to interpret.

As a *prediction baseline* (as opposed to TF, TF-IDF and random embeddings, which are baselines for text embedding models), we use the average response of each participant in the training data as the prediction for that participant's responses in the test set. We choose this baseline because previous research has shown that a respondent's answer behaviour (e.g. choosing a particular answer option) can be quite consistent across survey questions [58]. We also tried using model predictions based on only the background variables of respondents, but the model performance (the best Pearson's r based on RF: 0.187) is similar to using the within-person average baseline.

Table 4 Results of the Predictive Validity Analysis. *r* is the average Pearson's correlation between predicted and observed scores. Δ% in the parentheses indicates the percentage change in *r* in comparison to the baseline *r*: 0.187. 95% CI refers to the 95% confidence interval around *r*

	Lasso *r* (Δ%)	Lasso 95% CI	RF *r* (Δ%)	RF 95% CI
TF	0.106 (−43.316)	[0.102, 0.110]	0.337 (80.007)	[0.333,0.341]
TF-IDF	0.092 (−50.802)	[0.087, 0.096]	0.323 (72.830)	[0.319,0.327]
Random 300	0.149 (−20.321)	[0.144, 0.153]	0.331 (77.066)	[0.327,0.335]
Random 768	0.116 (−37.968)	[0.111, 0.120]	0.334 (78.614)	[0.330,0.338]
Random 1024	0.069 (−63.102)	[0.065, 0.073]	0.338 (80.520)	[0.333,0.342]
fastText	**0.204** (9.261)	[0.200, 0.209]	0.356 (90.439)	[0.352,0.360]
GloVe	0.107 (−42.781)	[0.103, 0.111]	0.347 (85.664)	[0.343,0.351]
BERT-base-uncased	0.195 (4.278)	[0.191, 0.200]	**0.411** (119.994)	[0.407,0.415]
BERT-large-uncased	0.151 (−19.251)	[0.147, 0.155]	0.378 (102.260)	[0.374,0.382]
All-DistilRoBERTa	0.188 (0.535)	[0.183, 0.192]	0.374 (100.228)	[0.370,0.378]
All-MPNet-base	0.119 (−36.364)	[0.115, 0.123]	0.406 (117.135)	[0.402,0.410]
USE	0.186 (−0.535)	[0.182, 0.191]	0.386 (106.272)	[0.382,0.390]

4.6.3 Results

Table 4 summarises the prediction performance of all text representation methods, measured as the average Pearson's correlation *r* across all 10 folds. Δ% in the parentheses indicates the percentage change in *r* in comparison to the prediction baseline *r*. Higher scores mean that the observed response scores are more correlated with the predicted response scores than with the prediction baseline. The more positive the value of Δ%, the more evidence for predictive validity. The 95% confidence intervals (CI) around *r* are also provided. The prediction baseline *r* is 0.187 with a 95% CI of [0.184, 0.190].

We can make three observations. First, RF consistently performs substantially better than Lasso regression. This is not a surprising result, as we know that RF can learn more complex patterns (like interactions) from data and impose stronger model variance reduction, compared to Lasso regression. Furthermore, because Lasso regression performs worse than or about the same as the prediction baseline, we can infer that the interaction between background variables and the survey question representations is crucial for a good prediction performance.

Second, despite RF faring better, the exact *r* scores depend on the text representation methods used. We see that simply using RF with baseline text representation methods (TF, TF-IDF and random embeddings) can already lead to substantial prediction improvement compared to the baseline prediction. All the embeddings approaches achieved higher *r* scores than did TF, TF-IDF and random embeddings. This suggests that predictive validity holds (to some extent) for the embedding models, especially the BERT-based models and the USE.

Lastly, in Appendix E, we summarise the prediction performance in terms of RMSE (Root Mean Square Error) and MAE (Mean Absolute Error). We see that the performance pattern (i.e. which text representation performs better or worse) stays similar. However, the magnitude of improvement over the baseline prediction is much smaller, compared to when Pearson's correlation is used as the evaluation metric. This may suggest that the prediction models are more capable of predicting the trend in survey responses (i.e. higher or lower) rather than the exact values.

In summary, we see that all the text embeddings exhibit some degree of predictive validity. Namely, they can be used to some extent to predict respondents' answers to completely new survey questions. However, the level of predictive validity that can be demonstrated

seems to highly depend on both the specific prediction algorithm, the embedding model and the evaluation metric used.

4.7 Summary

We find that different text embedding models demonstrate different degrees of validity evidence for the purpose of survey question representation. For instance, the USE and the two Sentence-BERT models show the best overall performance across all the construct validity analyses. In contrast, fastText fails to achieve performance comparable to the BERT-based approaches and the USE on all the validity analyses. Furthermore, even for the same text embedding model, its performance varies depending on the specific type of construct validity analysis. For instance, in our probing experiments to assess content validity, the USE and the two Sentence-BERT models show the best overall content validity. However, the original BERT models perform best in probing the formulation of survey questions. GloVe achieves prediction accuracy comparable to the Sentence-BERT models when probing concrete concepts, but it does poorly on the other probing tasks.

5 Conclusion and discussion

In this paper, we propose to view the quality evaluation of text embedding models as a measurement validity problem. That is, we can view what needs to be encoded from texts for a research task as the construct of interest, the text embedding model as the corresponding measure, and the embeddings as the measurements. In this way, we can apply the classic construct validity testing framework to evaluating the quality of text embeddings. This approach has the advantage of being grounded in measurement theory, providing a clear framework about what needs to be tested (i.e. in terms of the different types of validity), and being applicable to many substantive research questions in computational social science.

We describe how we can adapt this framework to suit text embeddings. For face validity, we recommend looking at readily available information about a text embedding model, such as the architecture and the data it is pretrained on. For content validity analysis, we suggest using probing classifiers to investigate explicit information in text embeddings. For convergent and discriminant validity analysis, we recommend constructing test cases where the texts only differ on a main property of interest. Then, we can inspect whether the cosine similarity score for a pair of texts is in line with our theoretical expectation. As for predictive validity, we need to define a relevant prediction task whose observed scores can be seen as prediction target. The better the prediction, the more evidence for predictive validity.

We also demonstrate this framework on the case of survey question representation. More concretely, we investigate the face, content, convergent, discriminant and predictive validity of various popular pre-trained text embedding models, including fastText, GloVe, BERT, Sentence-BERT and USE. Note that despite our use of pretrained embeddings, our approach to construct validity analysis applies equally to fine-tuned and continued pretrained embeddings.

5.1 Limitations

We are aware of several limitations of this study. First, we focus only on the validity of text embeddings. However, validity is only one aspect of measurement quality; the other

aspect is reliability (or instability). Previous work has shown that text embeddings can be unstable (e.g. across random seeds and hyper-parameters) [59–61]. Therefore, this instability aspect can also have an influence on the quality of text representation, and even on the results we obtain in our application study. Nevertheless, it is beyond the scope of our study to address the instability issue of embedding models. Second, the synthetic survey question data set and the ESS data do not cover nearly most survey question types (in terms of, for instance, the concrete concepts, the presence of problematic features), nor do they necessarily resemble the type of survey questions commonly used in research. This means that our findings concerning the construct validity of the investigated text embedding models may not generalise to other/all survey questions. Nevertheless, the goal of our paper is not to perform a comprehensive analysis of the construct validity of text embeddings for all survey questions, but to show how to conduct construct validity analysis for text embeddings, with our specific choice of survey questions as an application example. Last but not least, while we follow Saris and Gallhofer's [50] recommended procedure to generate the synthetic survey question dataset (according to the authors, this procedure, "if properly applied, will always lead to a measurement instrument that measures what is supposed to be measured"), the measurement quality of those synthetic survey questions is assumed rather than empiricallly tested and supported.

5.2 Future directions

Our proposed approach to quality evaluation of text embeddings, based on construct validity testing, can benefit many research areas that make use of text embedding models. For instance, predicting personality traits from Twitter data has been a difficult task, where models often do not outperform simple majority-class baseline [62]. One possible reason is that relevant linguistic cues are not encoded in the text embeddings used to represent Tweets. Our construct validity analysis can help to uncover possible reasons: Do the embeddings encode relevant linguistic cues (i.e. content validity)? Are the embeddings robust to noises and sensitive to altered texts that would indicate a different personality type (i.e. convergent and discriminant validity)? Can the embeddings perform tasks (e.g. prediction of personal achievements and health outcomes) that personality traits can (i.e. predictive validity)? We also hope that our proposed approach can provide insight into the strengths and weaknesses of current text embedding models and thus help the field develop more valid embedding models. Lastly, our paper focuses on the construct validity of text embeddings, while the construct validity of downstream models' prediction is equally important and should be studied in the future.

Appendix A: The synthetic survey question data set

Saris and Gallhofer [50] described a three-step procedure to operationalize a construct into survey questions. First, specify a complex construct by one or more basic concepts. Second, transform the basic concepts into statements. Third, transform the statements into survey questions formulated in a specific way. According to the authors, this procedure, "if properly applied, will always lead to a measurement instrument that measures what is supposed to be measured."

For instance, say we want to measure someone's attitude towards immigrants. This complex construct can be represented by three basic concepts: belief about immigrants,

feelings about immigrants and action tendency towards immigrants. These basic concepts can then be transformed into the following statements, among many other possibilities: "Immigrants have made my country a worse place", "I do not trust immigrants", and "I would not want to be neighbours with immigrants". Then, these statements should be transformed into survey questions following some formulation style, such as imperative-interrogative questions. An example is: "Tell me, do you trust immigrants?".

For researchers who would like to test the quality of text embedding models for survey questions used in a particular area of research (e.g. migration studies; family sociology), we recommend collecting a comprehensive and representative list of constructs relevant to the area of research, and then generate survey questions for these constructs following the three-step procedure. In this way, one can test how well a text embedding model represents survey questions in a particular field.

In our analysis, we would like to cover a wide range of constructs typical in sociological research. However, it is beyond our capacity to cover all of them, nor do we have access to a representative list of sociological constructs. Because of this difficulty and that our focus is on demonstrating our proposed construct validity framework (rather than answering which text embedding model is the best at representing sociological survey questions in general), we take a more practical approach. Specifically, instead of generating survey questions from a predefined list of constructs, we generate concrete concepts for each of the 13 basic subjective concepts defined by Saris and Gallhofer [50], and then generate survey questions based on those concrete concepts. Because most (if not all) subjective survey questions fall under these 13 concepts, we are still able to generate a diverse (and potentially representative) set of survey questions. To generate concrete concepts for each one of the 13 basic concepts, we use the examples of concrete concepts given in [50].

Specifically, for every subjective concept, we assign three reference concrete concepts. Take the basic concept "evaluation" as an example: we specify "the state of health services", "the quality of higher education" and "the performance of the government" as the three corresponding reference concrete concepts. Next, for every reference concrete concept, we specify one similar concrete concept and one dissimilar concrete concept. Then, for each concrete concept, we generate a statement. We make sure that the statements for every triad of concrete concepts (i.e. reference, similar and dissimilar) only differ in one or two words that define the concrete concepts. Finally, we transform these statements into survey questions which vary in their formulation. [50] provided many templates for each type of formulation. We adopt 19 templates, implement them in code, and then automatically create differently formulated survey questions for each concrete concept. Our final data set contains 5436 unique subjective survey questions.

Naturally, our generated data set of survey questions do not cover nearly all constructs common in the social sciences. Nor are they sufficiently varied in the types of questions. For instance, batteries of stimuli are another popular form of survey questions we do not cover. We also do not cover questions that have problematic features (e.g. double-barrelled questions; questions with implicit assumptions).

Appendix B: Definitions of basic subjective concepts

The definitions and examples are taken from Saris and Gallhofer (2007) [50] and the codebook of Survey Quality Predictor [63].

Evaluation Evaluations are questions about attitudes. They are marked by evaluative words like good/bad, positive/negative, perfect/imperfect, superior/inferior, useful/useless etc. For example: "Is the state of health services in your country is good?".

Importance The importance of something concerns questions that include an expression of "importance" as subject complement. Example: "Is personal success important to you?".

Feelings These questions refer to affective evaluations or feelings for something. Example: "Do you trust the legal system?".

Cognitive judgment Cognitive judgments concern questions about someone or something not in evaluative terms (e.g. good/bad, positive/negative) but in neutral terms (e.g. active/passive, small/big, aware/unaware). Example: "How large is the role of politics in your life?".

Causal relationship Such a question concerns a relationship between a subject and an object, with the subject being the cause of the object. Example: "Have immigrants made your country a worse place?".

Similarity These questions concern the similarity/dissimilarity or distance/closeness between objects or connectedness between subjects. Example: "How attached are you to your nationality?".

Preferences Such questions concern one's preferences. They are frequently used in consumer research, election studies and in studies of policies. Example: "Are you for or against universal healthcare?".

Norms Norms questions concern actions deemed by a group of people to be proper or correct. They often contain words like "should" or "have to". Example: "Do you think women should have children?".

Policies These questions are similar to norms questions, but are about what the government or people in power should do. Example: "Do you agree that the government should reduce income inequality?".

Rights These are questions about permission. They often contain words like "accepted", "allowed", "justified" or "the right to". Example: "Do immigrants have the right to social security?".

Action tendency Action tendency questions are about what one intends to do in the future. Example: "Will you vote in the election?".

Expectation These questions are about anticipations of events in which the respondent is not involved. Example: "Do you expect another economic crisis to come soon?".

Belief Such questions aim for an evaluation of the respondent's belief in something without the explicit use of evaluative terms like good/bad, positive/negative, etc. Despite that, they still typically have a positive or negative connotation. Example: "Do people in your country face discrimination because of their gender?".

Appendix C: The concrete concepts

Below we list all the concrete concepts used in our synthetic data set, sorted by the 13 basic concepts. Each basic concept consists of 3 triads of concrete concepts, separated by a semicolon. Each triad is constructed in the following way: a reference concrete concept, a similar concrete concept, and a dissimilar concrete concept. The similarity is defined relative to the reference concrete concept.

Evaluation Evaluation of the state's health services, evaluation of the state's medical services, evaluation of the state's religious services; evaluation of the quality of higher education, evaluation of the quality of public schooling, evaluation of the quality of the transportation infrastructure; evaluation of the performance of the current government, evaluation of the performance of the current leadership, evaluation of the performance of the current economy.

Importance The importance of personal success, the importance of personal achievement, the importance of personal safety; the importance of personal creativity, the importance of personal novelty, the importance of personal loyalty; the importance of personal family, the importance of personal friendship, the importance of personal career.

Feelings Trust in the legal system, trust in the police force, trust in the pharmaceutical industry; trust in family, trust in parents, trust in the president; trust in the Internet, trust in the media, trust in the church.

Cognitive judgment Personal political orientation, personal ideological orientation, personal sexual orientation; the role of politics in one's life, the role of activism in one's life, the role of family in one's life; the influence of ethnicity on one's career, the influence of race on one's career, the influence of gender on one's career.

Causal relationship Immigrants making my country a worse place, refugees making my country a worse place, politicians making my country a worse place; immigrants making my culture richer, foreigners making my culture richer, artists making my culture richer; the economic reform making my life better, the monetary reform making my life better, the education reform making my life better.

Similarity Closeness to a political party, closeness to a government organization, closeness to a religious group; attachment to one's country, attachment to one's nationality, attachment to one's religion; closeness to one's children, closeness to one's kids, closeness to one's colleagues.

Preferences Support racial equality, support ethnic equality, support income equality; against universal healthcare, against universal coverage, against universal demogrant; for the freedom of speech, for the freedom of expression, for the freedom of movement.

Norms Women should have children, women should be mothers, women should stay home; children should respect parents, children should respect elders, children should respect peers; one should believe in God, one should believe in Jesus, one should believe in oneself.

Policies Government should reduce economic inequality, government should reduce income inequality, government should reduce age inequality; government should make education affordable, government should make schooling affordable, government should make housing affordable; government should make stricter immigration policy, government should make stricter naturalisation policy, government should make stricter economic policy.

Rights Women's right to marriage, women's right to divorce, women's right to education; people's right to free speech, people's right to free expression, people's right to free education; one's right to education, one's right to schooling, one's right to life.

Action tendency Petition against institutional racism, petition against systematic racism, petition against individual racism; protest against police brutality, protest against police violence, protest against police reforms; participate in election, participate in referendum, participate in demonstration.

Expectation Housing prices will increase, home prices will increase, food prices will increase; pandemic will end, epidemic will end, war will end; economic crisis will come, economic depression will come, economic boom will come.

Belief Extent of gender discrimination in one's country, extent of sex discrimination in one's country, extent of religion discrimination in one's country; extent of ethnicity discrimination in one's country, extent of racial discrimination in one's country, extent of age discrimination in one's country; people lose jobs over sexual identity, people lose jobs over gender identity, people lose jobs over religious identity.

Appendix D: Survey question formulations

A statement can be transformed into different formulations. We list below the 5 types that we use in this study: direct requests, imperative-interrogative requests, interrogative-interrogative requests, declarative-interrogative request, and interrogative-declarative request.

According to [50], a *direct request* is based on the inversion of the (auxiliary) verb and the subject in the given statement. The other formulations consist of two parts. The first part (also the main clause) mostly indicates a researcher's desire to obtain information, while the second part presents the concept of interest (a subordinate clause). For an *imperative-interrogative request*, the main clause is phrased in the imperative mood (indicated by words like "tell", "specify", "indicate"). For an *interrogative-interrogative request*, the main clause uses the interrogative mood (e.g. "Could you tell me …"), which is more polite than the imperative mood. Polite declarative terms can also be used (e.g. "I would like to ask …"), which makes a *declarative-interrogative request*. Lastly, an *interrogative-declarative request* formulates the main clause in the interrogative mood and the subclause in the declarative form (e.g. "Do you think that …").

See below the 19 templates we used for the study. The direct request is simply the inversion of the (auxiliary) verb and the subject.

Imperative-interrogative requests "Tell me …", "Specify …", "Please tell me …", "Please specify …".

Interrogative-interrogative requests "Will you tell me …", "Can you tell me …", "Can you please tell me …", "Could you tell me …", "Could you please tell me …", "Would you tell me …", "Would you please tell me …", "Would you like to tell me …", "Would you mind telling me …", "Would you be so kind as to tell me …".

Declarative-interrogative request "I ask you …", "I would like to ask you …".

Interrogative-declarative request "Do you think that …", "Do you believe that …", "Do you agree or disagree that …".

Appendix E: Additional results of the predictive validity analysis

Table 5 and Table 6.

Table 5 Results of the Predictive Validity Analysis. MAE is the mean absolute error between predicted and observed scores. Δ% in the parentheses indicates the percentage change in MAE in comparison to the baseline MAE: 0.240. 95% CI refers to the 95% confidence interval around MAE

	Lasso 0.240 (Δ%)	Lasso 95% CI	RF 0.240 (Δ%)	RF 95% CI
TF	0.247 (2.970)	[0.233, 0.261]	0.226 (–5.958)	[0.211, 0.24]
TF-IDF	0.248 (3.478)	[0.233, 0.264]	0.228 (–5.061)	[0.211, 0.245]
Random 300	0.245 (2.247)	[0.233, 0.258]	0.231 (–3.568)	[0.219, 0.244]
Random 768	0.245 (1.899)	[0.232, 0.258]	0.231 (–3.842)	[0.218, 0.243]
Random 1024	0.247 (3.082)	[0.234, 0.261]	0.230 (–4.206)	[0.218, 0.242]
fastText	0.240 (–0.056)	[0.229, 0.251]	0.227 (–5.299)	[0.216, 0.239]
GloVe	0.245 (2.239)	[0.234, 0.257]	0.227 (–5.273)	[0.215, 0.240]
BERT-base-uncased	0.243 (1.391)	[0.230, 0.257]	0.222 (–7.527)	[0.211, 0.233]
BERT-large-uncased	0.245 (2.067)	[0.232, 0.258]	0.225 (–6.051)	[0.215, 0.236]
All-DistilRoBERTa	0.240 (–0.102)	[0.228, 0.252]	0.224 (–6.601)	[0.213, 0.235]
All-MPNet-base	0.245 (1.975)	[0.232, 0.258]	0.223 (–7.088)	[0.211, 0.235]
USE	0.241 (0.309)	[0.228, 0.254]	0.224 (–6.478)	[0.213, 0.236]

Table 6 Results of the Predictive Validity Analysis. RMSE is the root-mean-square error between predicted and observed scores. Δ% in the parentheses indicates the percentage change in RMSE in comparison to the baseline RMSE: 0.288. 95% CI refers to the 95% confidence interval around RMSE

	Lasso RMSE (Δ%)	Lasso 95% CI	RF RMSE (Δ%)	RF 95% CI
TF	0.296 (2.847)	[0.283, 0.310]	0.279 (–3.059)	[0.263, 0.296]
TF-IDF	0.298 (3.525)	[0.283, 0.314]	0.284 (–1.598)	[0.264, 0.303]
Random 300	0.294 (2.207)	[0.282, 0.307]	0.280 (–2.769)	[0.268, 0.292]
Random 768	0.293 (1.658)	[0.280, 0.306]	0.280 (–2.923)	[0.267, 0.292]
Random 1024	0.297 (3.101)	[0.283, 0.311]	0.279 (–3.170)	[0.266, 0.292]
fastText	0.290 (0.781)	[0.278, 0.303]	0.277 (–3.772)	[0.266, 0.289]
GloVe	0.295 (2.425)	[0.283, 0.308]	0.276 (–4.200)	[0.263, 0.289]
BERT-base-uncased	0.294 (1.962)	[0.279, 0.309]	0.270 (–6.197)	[0.259, 0.282]
BERT-large-uncased	0.297 (2.986)	[0.282, 0.311]	0.274 (–5.018)	[0.263, 0.285]
All-DistilRoBERTa	0.287 (–0.238)	[0.275, 0.300]	0.272 (–5.434)	[0.260, 0.284]
All-MPNet-base	0.294 (1.947)	[0.280, 0.308]	0.270 (–6.327)	[0.258, 0.282]
USE	0.290 (0.662)	[0.277, 0.303]	0.273 (–5.405)	[0.261, 0.284]

Appendix F: Reproducibility checklist

We check our paper against the ACL Reproducibility Checklist. We report additional details concerning the reproducibility of our experiments that are not mentioned in the main body of the paper.

The computing infrastructure Most analyses were done on a personal laptop, which runs on a 11th Gen Intel(R) Core(TM) i7-11800H processor with 32 GB of RAM. These analyses concern generating sentence embeddings from pretrained embedding models, generating survey questions, running probing models for the content validity analysis, computing cosine similarity scores for the convergent and discriminant validity analysis, and training the lasso regression models for the predictive validity analysis.

For training the Random Forest models (during the predictive validity analysis), we used the High Performance Computing service available at Utrecht University. We used 50 GB of RAM on a CPU with 12 parallel nodes.

Average runtime Each Random Forest model took on average 5 to 6 hours. All the other analyses took only a few seconds on average.

The number of parameters For the Lasso regression, the number of parameters was up to about 1300. For the Random Forest models, the number of parameters was up to about 5000.

Hyperparameter search For the Lasso regression, one hyperparameter was used: alpha. The alpha search was done automatically by the scikit-learn implementation of LassoCV [64]. We only specified the length of the regularization path (0.001), which is defined as the ratio of the smallest alpha to the largest. The criterion was mean squared error.

For the Random Forest models, three hyperparameters were used: the number of estimators (between 100 and 500 with an increment of 5), the maximum number of features in a tree (either log 2 or the square root of the total number of features), and the maximum depth of a tree (between 10 and 100 in a step of 10). The hyperparameter search was done using grid search. The evaluation criterion was mean squared error.

Acknowledgements
The authors thank the following colleagues for their useful feedback (in alphabetical order): Ayoub Baghari, Laura Boeschoten, Anastasia Giachanou, Erik-Jan van Kesteren.

Funding
This work was supported by the Dutch Research Council (NWO) (grant number VI.Vidi.195.152 to D.L. Oberski; grant number VI.Veni.192.130 to D. Nguyen).

Abbreviations
NLP, natural language processing; NLU, natural language understanding; ESS, European Social Survey; BoW, bag-of-words; TF, term frequency; TF-IDF, term frequency-inverse document frequency; GloVe, Global Vectors for Word Representation; USE, Universal Sentence Encoder; BERT, Bidirectional Encoder Representations from Transformers; DR, direct request; InDe, interrogative-declarative request; RF, random forest; CI, confidence interval.

Availability of data and materials
Our code and synthetic data set of survey questions are available at https://github.com/fqixiang/Survey-Embedding-Validity. The ESS Wave 9 data set is available for download at https://www.europeansocialsurvey.org/download.html?file=ESS9e03_1&y=2018. The GloVe pretrained embeddings are available at https://nlp.stanford.edu/projects/glove/. The other pretrained embedding models can be downloaded using the specific Python packages and commands provided in our code.

Declarations

Competing interests
The authors declare that they have no competing interests.

Author contributions
DLO proposed the research project; QF designed and performed the research; DN and DLO supervised the research; QF wrote the manuscript. All authors read, proofread and approved the final manuscript.

Author details
[1]Department of Methodology & Statistics, Utrecht University, Padualaan 14, Utrecht, The Netherlands. [2]Department of Information & Computing Sciences, Utrecht University, Princetonplein 5, Utrecht, The Netherlands. [3]Department of Biostatistics and Data Science, Julius Center, University Medical Center Utrecht (UMCU), Universiteitsweg 100, Utrecht, The Netherlands.

Publisher's Note

Received: 21 February 2022 Accepted: 22 June 2022 Published online: 07 July 2022

References

1. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, vol 26. Curran Associates, Lake Tahoe Nevada
2. Reimers N, Gurevych I (2019) Sentence-BERT: sentence embeddings using Siamese BERT-networks. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, pp 3982–3992
3. Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. Trans Assoc Comput Linguist 5:135–146
4. Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. Association for Computational Linguistics, Minneapolis, pp 4171–4186
5. Vu H, Abdurahman S, Bhatia S, Ungar L (2020) Predicting responses to psychological questionnaires from participants' social media posts and question text embeddings. In: Findings of the association for computational linguistics: EMNLP 2020. Association for Computational Linguistics, pp 1512–1524, Online
6. Matero M, Idnani A, Son Y, Giorgi S, Vu H, Zamani M, Limbachiya P, Guntuku SC, Schwartz HA (2019) Suicide risk assessment with multi-level dual-context language and BERT. In: Proceedings of the sixth workshop on computational linguistics and clinical psychology. Association for Computational Linguistics, Minneapolis, pp 39–44
7. De Bruyne L, De Clercq O, Hoste V (2021) Emotional RobBERT and insensitive BERTje: combining transformers and affect lexica for Dutch emotion detection. In: Proceedings of the eleventh workshop on computational approaches to subjectivity, sentiment and social media analysis. Association for Computational Linguistics, pp 257–263, Online
8. Garg N, Schiebinger L, Jurafsky D, Zou J (2018) Word embeddings quantify 100 years of gender and ethnic stereotypes. Proc Natl Acad Sci 115(16):3635–3644
9. Conneau A, Kiela D (2018) Senteval: an evaluation toolkit for universal sentence representations. In: Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)
10. Wang A, Singh A, Michael J, Hill F, Levy O, Bowman SR (2019) GLUE: a multi-task benchmark and analysis platform for natural language understanding. In: International conference on learning representations
11. Trochim WMK, Donnelly JP, Arora K (2015) Research methods: the essential knowledge base. Cengage Learning, Boston
12. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan TJ, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D (2020) Language models are few-shot learners. ArXiv. arXiv:2005.14165
13. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R (2020) Albert: a lite bert for self-supervised learning of language representations. In: International conference on learning representations
14. Yang Z, Dai Z, Yang Y, Carbonell JG, Salakhutdinov R, Le QV (2019) XInet: generalized autoregressive pretraining for language understanding. In: NeurIPS
15. Mikolov T, Chen K, Corrado GS, Dean J (2013) Efficient estimation of word representations in vector space. In: ICLR
16. Wittgenstein LS (1958) Philosophical investigations = philosophische untersuchungen
17. Harris ZS (1954) Distributional structure. Word 10:146–162
18. Parasca I-E, Rauter AL, Roper J, Rusinov A, Bouchard G, Riedel S, Stenetorp P (2016) Defining words with words: beyond the distributional hypothesis. In: Proceedings of the 1st workshop on evaluating vector-space representations for NLP, pp 122–126
19. Mikolov T, Yih W-T, Zweig G (2013) Linguistic regularities in continuous space word representations. In: Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies. Association for Computational Linguistics, Atlanta, pp 746–751
20. Linzen T (2016) Issues in evaluating semantic spaces using word analogies. In: Proceedings of the 1st workshop on evaluating vector-space representations for NLP, pp 13–18
21. Caliskan A, Bryson JJ, Narayanan A (2017) Semantics derived automatically from language corpora contain human-like biases. Science 356:183–186
22. Rice D, Rhodes JH, Nteta TM (2019) Racial bias in legal language. Res Polit 6

23. Kumar V, Bhotia TS, Chakraborty T (2020) Nurse is closer to woman than surgeon? Mitigating gender-biased proximities in word embeddings. Trans Assoc Comput Linguist 8:486–503
24. Mikolov T, Grave E, Bojanowski P, Puhrsch C, Joulin A (2018) Advances in pre-training distributed word representations. In: Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018). European language resources association (ELRA), Miyazaki, Japan
25. Pennington J, Socher R, Manning CD (2014) Glove: global vectors for word representation. In: EMNLP
26. Rogers A, Kovaleva O, Rumshisky A (2020) A primer in BERTology: what we know about how BERT works. Trans Assoc Comput Linguist 8:842–866
27. Cer DM, Yang Y, Kong S-Y, Hua N, Limtiaco N, John RS, Constant N, Guajardo-Cespedes M, Yuan S, Tar C, Sung Y-H, Strope B, Kurzweil R (2018) Universal sentence encoder. ArXiv. arXiv:1803.11175
28. Goleman D (1995) Emotional intelligence. A Bantam book. Bantam Books, New York
29. Belinkov Y (2021) Probing classifiers: promises, shortcomings, and advances. Computational Linguistics
30. Liu NF, Gardner M, Belinkov Y, Peters ME, Smith NA (2019) Linguistic knowledge and transferability of contextual representations. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). Association for Computational Linguistics, Minneapolis, pp 1073–1094
31. Hupkes D, Zuidema WH (2018) Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. J Artif Intell Res 61:907–926
32. Hewitt J, Liang P (2019) Designing and interpreting probes with control tasks. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, pp 2733–2743
33. Alain G, Bengio Y (2017) Understanding intermediate layers using linear classifier probes. ArXiv. arXiv:1610.01644
34. Maudslay RH, Valvoda J, Pimentel T, Williams A, Cotterell R (2020) A tale of a probe and a parser. In: ACL
35. Belinkov Y, Durrani N, Dalvi F, Sajjad H, Glass JR (2017) What do neural machine translation models learn about morphology? In: ACL
36. Conneau A, Kruszewski G, Lample G, Barrault L, Baroni M (2018) What you can cram into a single $&!#* vector: probing sentence embeddings for linguistic properties. In: ACL
37. Zhang KW, Bowman SR (2018) Language modeling teaches you more than translation does: lessons learned through auxiliary syntactic task analysis. In: BlackboxNLP@EMNLP
38. Tenney I, Xia P, Chen B, Wang A, Poliak A, McCoy RT, Kim N, Durme BV, Bowman SR, Das D, Pavlick E (2019) What do you learn from context? Probing for sentence structure in contextualized word representations. In: International conference on learning representations
39. Belinkov Y, Bisk Y (2018) Synthetic and natural noise both break neural machine translation. In: International conference on learning representations. https://openreview.net/forum?id=BJ8vJebC
40. Ribeiro MT, Singh S, Guestrin C (2018) Semantically equivalent adversarial rules for debugging NLP models. In: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers). Association for Computational Linguistics, Melbourne, pp 856–865
41. Ribeiro MT, Wu T, Guestrin C, Singh S (2020) Beyond accuracy: behavioral testing of NLP models with CheckList. In: Proceedings of the 58th annual meeting of the association for computational linguistics. Association for Computational Linguistics, pp 4902–4912, Online
42. W AS, Pellegrini AM, Chan S, Brown HE, Rosenquist JN, Vuijk PJ, Doyle AE, Perlis RH, Cai T (2020) Integrating questionnaire measures for transdiagnostic psychiatric phenotyping using word2vec. PLoS ONE 15(4):e0230663
43. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) RoBERTa: a robustly optimized BERT pretraining approach. arXiv:1907.11692 [cs]
44. Sanh V, Debut L, Chaumond J, Wolf T (2019) DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In: 5th workshop on energy efficient machine learning and cognitive computing at NeurIPS'19
45. Song K, Tan X, Qin T, Lu J, Liu T-Y (2020) MPNet: masked and permuted pre-training for language understanding. arXiv:2004.09297 [cs]
46. Bowman SR, Angeli G, Potts C, Manning CD (2015) A large annotated corpus for learning natural language inference. In: EMNLP
47. Tawfik NS, Spruit MR (2020) Evaluating sentence representations for biomedical text: methods and experimental results. J Biomed Inform 104:103396
48. Rücklé A, Eger S, Peyrard M, Gurevych I (2018) Concatenated p-mean word embeddings as universal cross-lingual sentence representations. ArXiv. arXiv:1803.01400
49. Miller AS, Mitamura T (2003) Are surveys on trust trustworthy? Soc Psychol Q 66(1):62–70
50. Saris WE, Gallhofer IN (2007) Design, evaluation, and analysis of questionnaires for survey research. Wiley, Hoboken
51. Norwegian Centre for Research Data (2018) Norwegian centre for research data: European social survey round 9 data. Data file edition 3.1. Norway. https://doi.org/10.21338/NSD-ESS9-2018
52. Yan T, Tourangeau R (2008) Fast times and easy questions: the effects of age, experience and question complexity on web survey response times. Appl Cogn Psychol 22:51–68
53. Belinkov Y, Glass JR (2019) Analysis methods in neural language processing: a survey. Trans Assoc Comput Linguist 7:49–72
54. Norwegian Centre for Research Data (2021) Norwegian centre for research data: European social survey: ESS-9 2018 documentation report. Edition 3.1. Norway. https://doi.org/10.21338/NSD-ESS9-2018
55. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning. Springer series in statistics. Springer, New York
56. Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc, Ser B, Methodol 58:267–288
57. Breiman L (2001) Random forests. Mach Learn 45:5–32
58. Bais F, Schouten B, Toepoel V (2020) Investigating response patterns across surveys: do respondents show consistency in undesirable answer behaviour over multiple surveys? Bull Soc Method 147(1–2):150–168
59. Wendlandt L, Kummerfeld JK, Mihalcea R (2018) Factors influencing the surprising instability of word embeddings. In: Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long papers). Association for Computational Linguistics, New Orlean, pp 2092–2102

60. Burdick L, Kummerfeld JK, Mihalcea R (2021) Analyzing the surprising variability in word embedding stability across languages. In: Proceedings of the 2021 conference on empirical methods in natural language processing, pp 5891–5901
61. Mosbach M, Andriushchenko M, Klakow D (2020) On the stability of fine-tuning bert: misconceptions, explanations, and strong baselines. In: International conference on learning representations
62. Štajner S, Yenikent S (2021) Why is mbti personality detection from texts a difficult task? In: Proceedings of the 16th conference of the European chapter of the association for computational linguistics: main volume, pp 3580–3589
63. Saris WE, Oberski DL, Revilla M, Zavala-Rojas D, Lilleoja L, Gallhofer IN, Gruner T (2011) The development of the program sqp 2.0 for the prediction of the quality of survey questions
64. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830