RESEARCH

Evaluating the Construct Validity of Text Embeddings for Survey Questions

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

[*]Correspondence: q.fang@uu.nl
[1]Department of Methodology & Statistics, Utrecht University, Padualaan 14, Utrecht, NL
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 supposedly meaningful numerical representations, which are also called text embeddings. Such techniques have been used for many scientific applications like disease detection from questionnaire texts and personality prediction based on social media posts. However, such studies typically do not address the extent to which the text embeddings used are valid representations of the original texts. We argue that this is an important research question, because valid measurements (i.e. high construct validity) are crucial to obtaining valid scientific findings. Therefore, in our study, we propose a novel framework of construct validity analysis for text embeddings and demonstrate it on survey questions. Specifically, we investigate the construct validity of several popular text embedding methods (e.g. fastText, GloVe, BERT, Sentence-BERT, Universal Sentence Encoder) for survey questions. We show that in general, BERT-based embedding techniques and Universal Sentence Encoder provide more valid representations of survey questions than do others. We therefore argue that it is necessary to examine the construct validity of text embeddings before deploying them in research.

Keywords: word embeddings; sentence embeddings; measurement validity; content validity; convergent validity; discriminant validity; criterion validity; survey questions; survey methodology; computational social science

1 Introduction

Text embedding models/techniques, which originate from the field of 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 dimensions (~~e.g.~~ [1, 2]). Intuitively, 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, 2, 3, 4])). This allows other researchers to obtain off-the-shelf pretrained text embeddings for quick 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-tune*) or domain-specific data (i.e. *continue pre-training*) for better performance.

Because of their capability for meaningful text representation and convenient use, text embedding techniques have become increasingly popular and attracted

a growing number of applications in social science. For instance, text embeddings 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], *interview data* for automatic qualitative content analysis [8], and *historical texts* to quantify societal trends of gender and ethnic stereotypes in the US [9].

While such applications show the promising potential of text embeddings for social science research, it remains unclear to what extent existing text embeddings can provide **valid** representations for the texts of interest. Take [5] as an example, where the authors converted the questions from the Big-Five questionnaire into text embeddings. It is relevant to ask: Do the embeddings encode relevant information about the questions, such as the underlying personality concept of interest and the formulation? Are the embeddings of questions about empirically unrelated personality traits (e.g. openness vs. neuroticism [10]) located farther away from each other in the vector space than do the embeddings of questions about closely related personality traits (e.g. conscientiousness vs. agreeableness [10])? Notably, even when a model that uses text embeddings achieves good performance, there is no guarantee that the embeddings themselves are accurate representations of the original texts. This concern is corroborated by the research of [11], showing that even assigning random vectors to texts can sometimes achieve prediction performance close to using (pre)trained text embeddings.

Precisely, by "valid representations" we mean that text embeddings should demonstrate high *construct validity*. Construct validity concerns the degree to which a construct's operationalization in a study matches the construct in theory [12]. For instance, a survey question designed to detect depression in patients but instead measures anxiety would not be considered a valid instrument, which would subsequently cast doubt on findings based on this survey question. Applying the idea of construct validity, we can view a piece of text as the construct of interest and the corresponding embedding the operationalization. Good construct validity is highly regarded especially in social science, where measurement and explanation are often the research goals. This is also consistent with the aim to develop more accurate text representation methods in the NLP community.

We consider the following four types of construct validity in our study [12]:

1 **Content validity** concerns whether the operationalization adequately covers all relevant aspects of a construct. For instance, a language test with high content validity should cover all the topics relevant to the mastery of the language (e.g. listening, speaking, reading and writing skills).

2 **Convergent validity** concerns whether the operationalization of a construct is highly correlated with the operationalization of other constructs that it theoretically should be similar to. For instance, a psychological test on stress levels should highly correlate with a test on anxiety.

3 **Discriminant validity**, in contrast, concerns whether the operationalization of a construct is poorly correlated with operationalizations that measure theoretically dissimilar constructs. For example, there should be a low correlation between an instrument that measures intelligence and one that measures generalized trust.

4 **Criterion validity** concerns checking the performance of some operationalization against a criterion. For instance, we can assess the operationalization's ability to predict something it should theoretically be able to predict (e.g. IQ and school performance, where the former is the operationalization and the latter the criterion).

Thus, the goal of our paper is to examine the construct validity (specifically, content, convergent, discriminant and criterion validity) of several most commonly used text embedding models. However, there are two challenges to analysing construct validity for text embeddings. For one, content validity analysis typically relies on the researcher's ability to directly interpret the measure of interest. With text embeddings, this is difficult because there is no a priori interpretation about the embedding dimensions. For another, studying the other three types of validity often requires checking the correlation between a single numerical summary of a measure (e.g. the mean of a questionnaire scale) and that of another. A text embedding, nevertheless, is a high-dimensional numeric vector, to which this approach does not apply. Therefore, an alternative analytic approach to construct validity is needed.

To this end, we propose a novel framework of construct validity analysis for text embeddings. Specifically, we borrow tools from the field of interpretable NLP (e.g. [1, 2, 13]) and adapt them to the analysis of construct validity. We demonstrate this framework on one particular type of texts: **survey questions**. We choose to focus on analysing text embeddings for survey questions because of two reasons. First, survey questions are a popular and important measurement tool in social science research. Second, they are usually concise and precise texts. This means that there are likely fewer aspects that text embeddings need to encode, compared to texts that are longer and/or follow a more free style (e.g. Tweets; conversations). Therefore, survey questions can be seen as a "simple baseline" for the evaluation of the construct validity of text embeddings.

Through our analysis, we uncover the strengths and limitations of different text embedding models for survey questions. We thus show that it is necessary to examine the construct validity of such measures before using them. Building on the findings, we discuss the potential applications and future directions of text embedding techniques for survey and social science research. Our study also contributes an original data set consisting of 5,436 survey questions covering various survey concepts and linguistic properties. This data set can be used in future research for studying, for example, text embedding models specific for survey research.

Our paper is structured as follows. In Section 2, we introduce both classic count-based text representation ~~techniques~~ and text embedding techniques. We also review current work on applying text embeddings to survey questions. In Section 3, we outline the specific text embedding methods used in our study and explain how embeddings for survey questions can be computed. Then, we describe how we examine the content validity (Section 4), convergent & discriminant validity (Section 5) and criterion validity (Section 6) of various text embedding approaches for survey questions, ~~respectively.~~ We also present the analysis results in each corresponding section. Lastly, we summarize and discuss the findings in Section 7.

2 Background

2.1 Classic Count-based Text Representation Techniques

Prior to the introduction of text embeddings, a popular count-based approach to text representation was: bag-of-words (BoW) [14]. BoW represents a piece of text by describing the occurrence of words within the text, without taking word order into account (and hence the name). One way to use BoW is to simply count raw term frequency (TF). Take the survey question "How happy would you say you are?" as an example, we can represent this question as a vector [1, 1, 1, 1, 2, 1], with the numbers corresponding to the frequency of the terms "how", "happy", "would", "you", "say" and "are", respectively (see Figure 1).

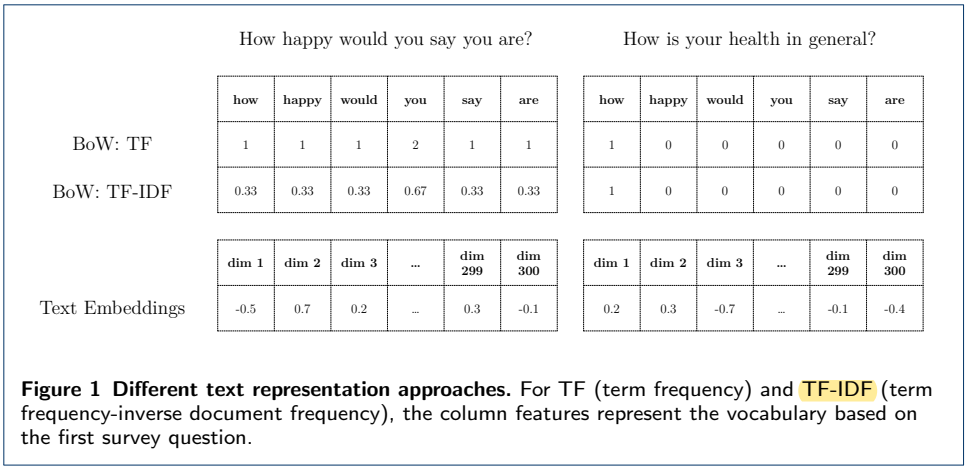

Figure 1 Different text representation approaches. For TF (term frequency) and TF-IDF (term frequency-inverse document frequency), the column features represent the vocabulary based on the first survey question.

However, a problem with raw term frequency is that highly frequent words (e.g. "the", "and") are not necessarily important words. A different approach, term frequency-inverse document frequency (TF-IDF), mitigates this issue by rescaling words according to how often they appear in all documents (i.e. document frequency). In this way, frequent but often uninformative words like "the" are penalized. Specifically, TF-IDF is calculated as: $tf_{i,j} * log(N/df_i)$, where $tf_{i,j}$ refers to the number of occurrences of word i in document j, df_i is the number of documents containing i, and N is the total number of documents. Note that a document can be a sentence, a paragraph, a book, etc.

We can see that both approaches share many strengths and weaknesses. For example, they are simple and efficient, but at the cost of potentially relevant information like word relations, grammar and word order. They also require specifying a priori a list of words to define the features, which is normally based on the vocabulary in the training data. This can lead to the problem that out-of-sample words cannot be accounted for. Figure 1 illustrates this using a new survey question: "How is your health in general?". If we define the feature vocabulary solely based on the first question, then for the new question, all the words except "how" are missing from both TF and TF-IDF representations, meaning that a large amount of meaningful information from question 2 is lost.

2.2 Text Embedding Techniques

2.2.1 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 other words in their context (or the other way around). In this way, the final word vectors (or embeddings) are positioned in the vector space such that words that share common contexts in the corpus are located closely to one another. Under the Distributional Hypothesis assumption [16, 17] that words occurring in similar contents tend to have similar meanings, closely located words in the vector space are expected to be semantically similar. The semantic distance between two word vectors can be measured by cosine similarity, which is a measure of distance between two n-dimensional non-zero vectors in an n-dimensional space. 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 [18]. 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 [18]. Many studies have also made use of this characteristic of word2vec to study human biases (e.g. gender and racial bias) encoded in texts [9, 19, 20, 21].

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 [22].

2.2.2 GloVe

GloVe, which stands for Gloval Vector word representations [23], 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 factorization is used to learn word embeddings of lower dimensions.

2.2.3 BERT and Sentence-BERT

More sophisticated embedding techniques have recently become available. A prominent one is BERT, which stands for Bidirectional Encoder Representations from Transformers [4]. Similar to word2vec, BERT is a special type of neural networks trained over a large size of text data in order to learn a good representation of natural language. The main difference is that BERT has a much more complex model

architecture, is trained with different objective functions and ultimately, produces context-dependent embeddings. That is, while word2vec and GloVe models produce a fixed, global embedding for each word, BERT can produce different embeddings for the same word depending on the context (e.g. the neighbouring words).

BERT has achieved state-of-the-art performance on various natural language tasks such as Semantic Textual Similarity, Paraphrase Identification, Question Answering, and Recognizing Textual Entailment [4]. BERT embeddings have also been shown to encode syntactic and semantic knowledge about the original texts [24].

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

2.2.4 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 [25]. 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 [25]. 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.

2.3 Research Applications of Text Embeddings to Survey Questions

Like many other areas in social science, survey research has also been seeing an increasing number of applications using text embedding techniques. We have identified two such studies that applied text embedding techniques specifically to survey questions. [5] used ~~a text embedding model called~~ BERT to encode participants' social media posts and the questions from the Big-Five personality questionnaire. They showed that by making use of the generated pretrained text embeddings, they were able to moderately improve the prediction of individual-level responses to out-of-sample Big-Five questions, compared to not using any embeddings. [26] used the skip-gram embedding algorithm [1] to represent the questions in 9 different questionnaires about psychiatric symptoms. The embeddings of the questions were then weighted by the numerical responses from psychiatric patients, indicating the severity of specific disease symptoms. In this way, the authors created so-called embeddings profiles unique for every patient. They showed that by applying clustering and classification techniques, such embeddings profiles can be used for effective diagnosis of axis I disorders.

3 Pretrained Embeddings for Survey Questions

3.1 Pretrained Text Embedding Models

In this paper, we investigate whether current text embedding techniques like fastText, GloVe, BERT and USE can produce valid representations for survey questions.

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

We focus on pretrained embedding models (as opposed to fine-tuned models) because of their widespread use. However, note that our approach to construct validity analysis applies to fine-tuned text embeddings as well.

Table 1 summarizes the specific pretrained embedding models we adopted. For fastText, we used the pretrained model developed by [3]. It is trained on Common Crawl and Wikipedia, and produces word embeddings with 300 dimensions. For GloVe, we used the model pretrained on Common Crawl with 840B tokens and a vocabulary of 2.2 million words. It also outputs word vectors of 300 dimensions.

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 optimized for, but also in their model architecture. We selected two pretrained models which according to [2] have been trained on various training data and are thus designed as general purpose models. They are "All-DistilRoBERTa"" and "All-MPNet-base"", where "DistilRoBERTa" [27, 28] and "MPNet" [29] 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 models have been shown to have the top average performance across various language tasks. For the purpose of comparison, we also included two pretrained models of the original BERT [4]: "BERT-base-uncased" and "BERT-large-uncased". "Uncased" refers to BERT treating upper and lower cases equally.

As for USE, we used the most recent version of its pretrained model, which outputs a 512 dimensional vector given an input sentence.

3.2 Using Pretrained Embedding Models for Sentence-level Embeddings

For texts like survey questions, we need to obtain sentence-level representations (as opposed to word-level) from pretrained text embedding models. Figure 2 illustrates this process. For instance, given three survey questions that measure three different concepts ("trust in government", "trust in parliament" and "trust in strangers"), we can feed them into a pretrained embedding model and in return obtain three sentence-level embeddings that supposedly represent the original questions. Then, we can perform our desired evaluation on certain property of these embeddings.

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. Therefore, it is necessary to combine the word embeddings into a sentence-level representation. Various methods to do so have been proposed. Among them, simple averaging across all the word embeddings (e.g. taking the means along each dimension) has been shown to be either outperform other approaches [30] or approximate the performance of more sophisticated methods [31]. Therefore, we use simple averaging to compute

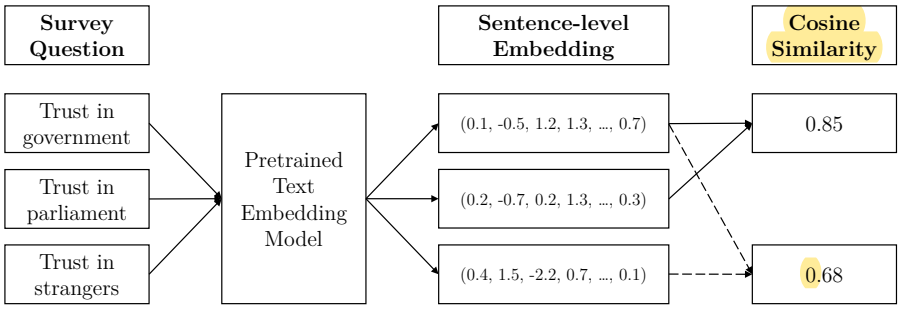

Figure 2 Illustrating the Use of Pre-trained Text Embedding Models

sentence-level embeddings for survey questions from fastText and GloVe word embeddings. The resulting representations 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 likely absent in the aggregated representation.

As for the original BERT models, we follow the advice of [24, 2] 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 Analysis of Content Validity

Our 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 one are the underlying **concepts**. According to the typology proposed by [32], most survey questions can be categorized into one of 21 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").

Furthermore, survey questions can differ in terms of **formulation**. Specifically, five types of formulation often apply in survey research [32]: direct request (DR), imperative-interrogative request (ImIn), interrogative-interrogative request (InIn), declarative-interrogative request (DeIn) and interrogative-declarative request (InDe)[1]. See Appendix C for examples of these different formulations.

Lastly, **complexity** is another important aspect of survey questions which can affect how respondents understand and answer a survey question [33]. It can be measured as the length of a survey question [32].

[1] [32] 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.

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.

4.1 Data

We constructed a synthetic data set of survey questions that satisfies two requirements. First, it should cover a wide, (ideally) representative range of survey concepts. Second, for every survey question, there should be corresponding survey questions that differ in only the concepts, or only the formulation, or both (similar to the idea of "minimal pairs" in NLP [34]). In this way, we have better control over the properties of the survey questions, which will benefit our validity analysis.

For the first requirement, we focus on covering a wide selection of concepts for subjective survey questions, which aim to measure information that only exists in the respondent's mind (e.g. opinions). According to [32], such questions normally fall under one of the following 13 basic concepts: "evaluation", "importance", "feelings", "cognitive judgment", "causal relationship", "similarity", "preferences", "norms", "policies", "rights", "action tendencies", "expectation", and "beliefs". The questions in our data set therefore cover these 13 basic concepts.

To satisfy the second condition, for every subjective concept, we assigned three reference concrete concepts. Take the basic concept "evaluation" as an example: we specified "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 specified one similar concrete concept and one dissimilar concrete concept[2]. Finally, for each concrete concept, we created survey questions that vary in their formulation. [32] provided many templates for each type of formulation. We adopted 19 templates and thus created differently formulated survey questions for each concrete concept. Our final data set contains 5436 unique survey questions.

Table 2 Example Questions from the 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?

Table 2 shows six example questions from the data set. They all fall under the basic concept "evaluation". The main concrete concept here is "the state of health services", while the corresponding similar and dissimilar concepts are "the state of

[2]Based on our judgment and experience working in the field of survey research.

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 Methods

4.2.1 Probing Classifiers

~~As we saw in Figure 1,~~ text embeddings are high-dimensional and opaque. This makes it difficult to learn what information is encoded in them. Luckily, there has been promising development in NLP methodology to achieve this goal. A very popular approach are so-called **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 text embedding technique has learned information relevant to the property [13].

A recommended practice in choosing a classifier is to select a linear model like (multinomial) logistic regression, because a more complex probe may run the risk that the classifier infers properties not actually present in the text representation [35, 36, 37, 38, 39]. Furthermore, it is recommended to always include baselines for comparison [13]. The better the probing classifier based on some text representation performs relative to the baselines, the more evidence that the probed property is present. Following studies like [40, 41, 42, 43], 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.

4.2.2 Adapting Probing Classifiers to Survey Questions

A common problem with probing classifiers is that the good performance of the model could simply be due to the model making use of other properties present in the embeddings that are correlated with the properties of interest [44]. 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 can be 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 features, length and formulation are completely uncorrelated. To mitigate this issue, we construct our training and test sets such that they do not share the same distribution of the correlated properties. In this way, the probing classifier can no longer make use of the correlated properties to achieve good performance on the test sets.

In our data, we see that sentence length is highly correlated with all the other properties. Using chi-square tests of independence, 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 those properties, we constrain our training data to contain only survey questions that have different lengths than the ones in the test data. 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.

Unfortunately, even separating the training and test set in terms of sentence length was not enough for effective probing of concrete concepts. We found that regardless of whether we used random embeddings or the actual text embeddings, the classifier always achieved perfect performance on the test set. The absence of difference in performance prohibits us from concluding whether there is any information about concrete concepts encoded in the text embeddings. 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 defined in Section 4.1), 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.

4.3 Results

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

	Sentence Length	Basic Concept	Concrete Concept	Formulation
Simple Majority	0.389	0.010	0.029	0.255
Random 300	0.102	0.198	0.440	0.742
Random 768	0.148	0.198	0.509	0.694
Random 1024	0.074	0.198	0.548	0.731
TF	0.148	0.198	0.636	0.770
TF-IDF	0.167	0.198	0.493	0.690
fastText	0.093	0.173	0.711	0.656
GloVe	0.194	0.192	0.908	0.642
BERT-base-uncased	0.657	0.175	0.815	0.944
BERT-large-uncased	0.620	0.153	0.739	0.908
All-DistilRoBERTa	0.407	0.198	0.916	0.776
All-MPNet-base	0.481	0.198	0.929	0.805
USE	0.454	0.198	0.903	0.853

Table 3 summarizes the performance of the probing classifier (multinomial logistic regression) across fastText and GloVe embeddings, two types of original BERT embeddings, two different Sentence-BERT embeddings and the USE embeddings.

Classification accuracy scores based on simple majority voting, random embeddings of three dimension sizes, TF and TF-IDF vectors serve as baselines.

If the classifier performs better on a particular type of text embeddings than on the 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 counterparts.

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 [32] are too abstract for current embedding techniques to "comprehend".

As for concrete concepts, all types of pretrained text embeddings perform much 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 show better performance than do the original BERT embeddings. This may be because the Sentence-BERT models have been trained on tasks like Semantic Similarity and Paraphrase Identification, which is arguably similar to identifying sentences with similar or identical underlying concrete concepts. USE and GloVe also 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 techniques 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 techniques do encode some information about survey questions like concrete concepts, but not sentence length or formulation.

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. By the same

logic, 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. As we can see, convergent and discriminant validity are two sides of the same coin. A measure is only properly defined in relation to other measures when both types of validity are established.

5.1 Data

We used the same synthetic data set of survey questions presented in Section 4.

5.2 Methods: Cosine Similarity Analysis

We take a joint approach to examining convergent and discriminant validity. That is, if text embeddings possess both convergent and discriminant validity, they would find conceptually similar survey questions closer to one another while conceptually dissimilar ones 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.

We use the same survey question data set we created for the analysis of content validity, where 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. This allows us to examine the two proposed hypotheses.

Specifically, 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. $E_{\text{dissimilar}}$). 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 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 when evaluating Hypothesis 1, which allows us to attribute any observed differences in similarity scores to the differences in the underlying 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. $E_{\text{dissimilar}}$). 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 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.

5.3 Results

Figure 3 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 the 13 subjective basic concepts, various baselines and text embedding approaches. The more positive the difference scores are, the more support for the presence of convergent and discriminant validity.

We can see in Figure 3 that the only models that consistently score above zero are "All-DistilRoBERTa", "All-MPNet-Base" and USE, with the percentages of positive scores being 98.3%, 96.8% and 95.4%, respectively. This shows evidence of convergent and discriminant validity. Only in a small percentage of cases does this observation not hold (e.g. both Sentence-BERT models for the concrete concept "close to political party"). 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. This would suggest that these text embeddings approaches 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 4 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 overlapped 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 differ in only one word, Jaccard similarity would always be zero and therefore not a useful measure.

Similar to Figure 3, we can see that the two Sentence-BERT models and USE again consistently score 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 convergent and discriminant validity likely hold for these models. Most of the other approaches (including the baselines, the random embeddings and the original BERT) 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 generated the TF-IDF vectors, we built the vocabulary based on all the survey questions. We adopted this approach because in this analysis, it is unclear what the training and testing data should be. In real research applications, TF-IDF is unlikely to perform so well due to the difference in the vocabulary between training and test data.

Last but not least, it is worth noting the two Sentence-BERT models performed either about equally or better in Hypothesis 2 than in Hypothesis 1. This is some-

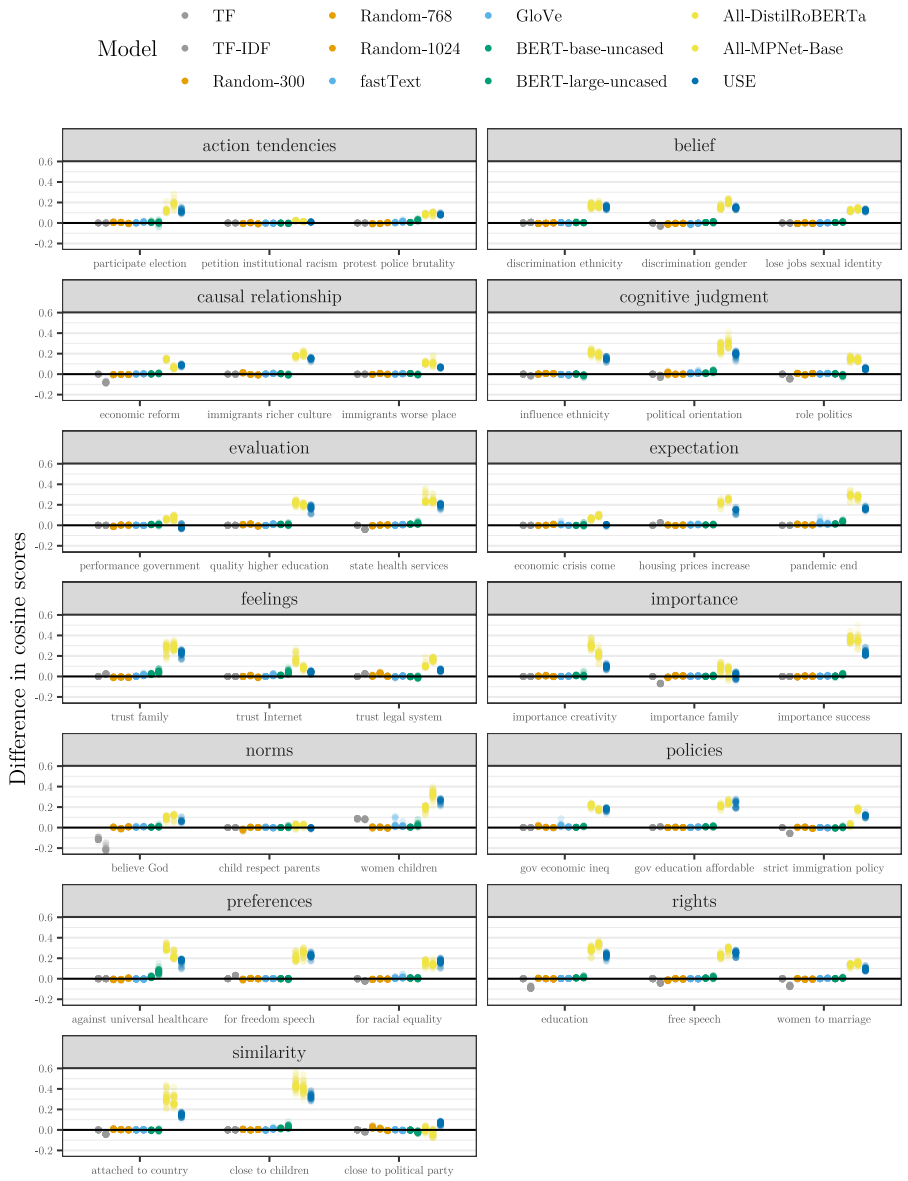

Figure 3 The distribution of the cosine similarity difference scores for Hypothesis 1 across 13 basic concepts. The y-axis indicates the size and direction of the differences. The more positive the difference scores are, the more support for the presence of convergent and discriminant validity. The x-axis labels are the (abbreviated) names of the main concrete concepts in the question triads.

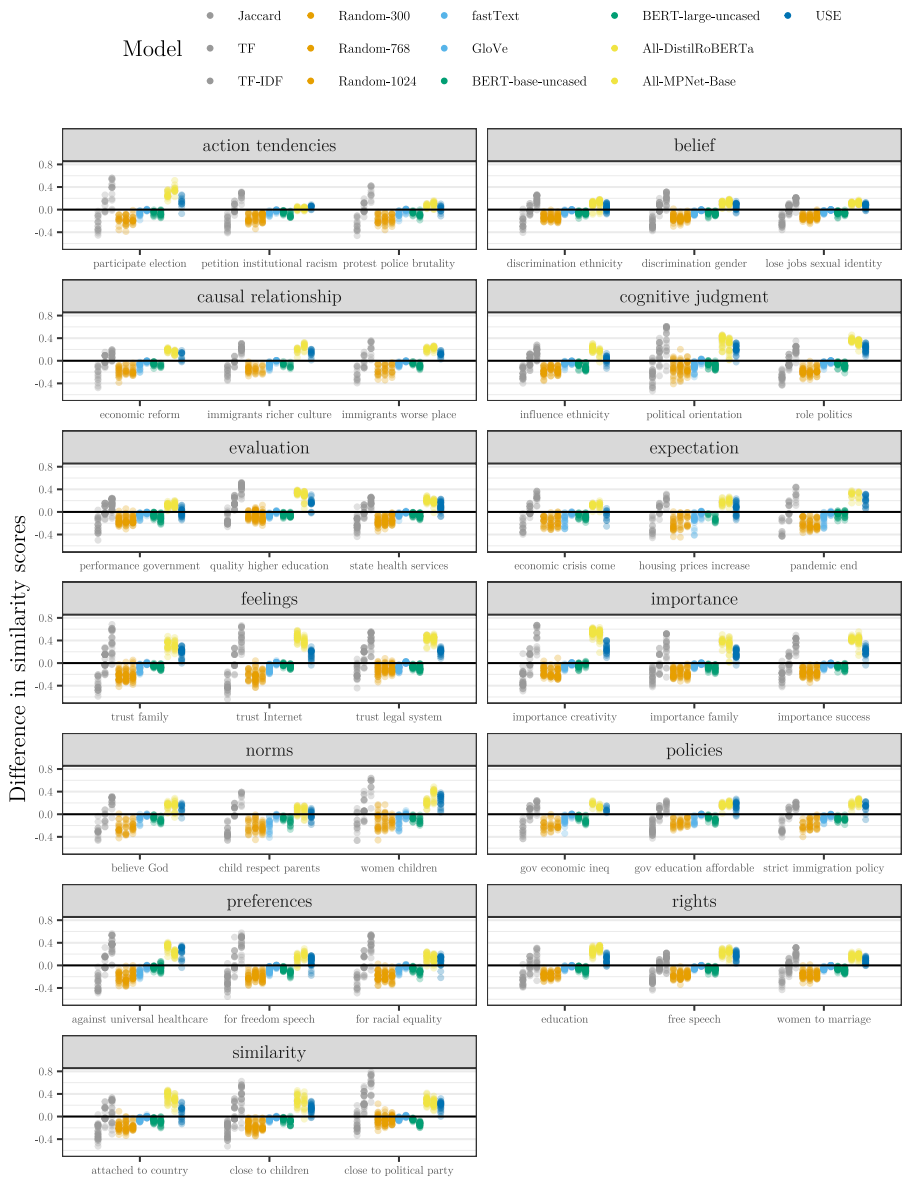

Figure 4 The distribution of the cosine similarity difference scores for Hypothesis 2 across 13 basic concepts. The y-axis indicates the size and direction of the differences. The more positive the difference scores are, the more support for the presence of convergent and discriminant validity. The x-axis labels are the (abbreviated) names of the main concrete concepts in the question triads.

what 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 convergent and discriminant validity. Meanwhile, there is not enough evidence to suggest the same for the other approaches.

6 Analysis of Criterion Validity

Criterion validity concerns how well a measure of interest relates to some criterion. Here, we define the criterion as observed individual responses to survey questions. That is, if text embeddings exhibit good criterion validity, they should improve the prediction of responses 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 criterion validity.

6.1 Data: European Social Survey Wave 9

We used the publicly available European Social Survey (ESS) Wave 9 data [45]. The ESS is a research-oriented cross-national survey that is conducted with newly selected, cross-sectional samples every two years since 2001 [46]. 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 focused 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 selected 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 analysis. To harmonize the difference in response scales across the survey questions, we rescaled 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.

6.2 Methods: Prediction Modelling

Figure 5 illustrates the prediction modelling processes for cases where we apply text embedding methods. Namely, our data set now consists of the following features/columns: background variables, embeddings of the survey questions generated by text embedding models, and survey question responses. Each row corresponds to a unique respondent & survey question combination. We then randomly partitioned our data set into an 80% training set and a 20% test set, while ensuring that the survey questions in the training set are different from the ones in the test set.

We also included TF, TF-IDF and random embeddings of dimension 300, 768 and 1024 as baseline text representation techniques. Note that with these baseline approaches, we built 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.

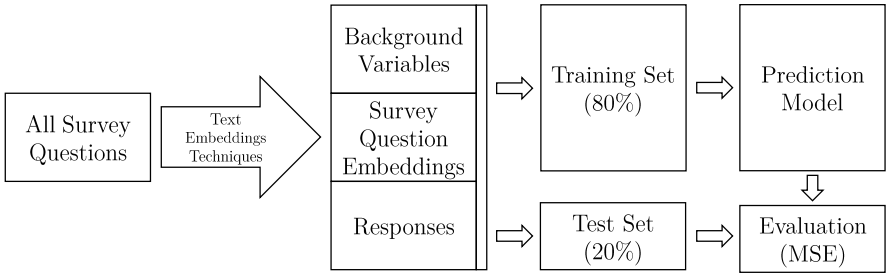

Figure 5 Prediction Modelling Approach for the Analysis of Criterion Validity. Note that this exact approach applies only to the text embedding methods (i.e. excluding TF, TF-IDF and random embeddings).

6.2.1 Lasso and Random Forest

We adopted two popular prediction models. The first is Lasso regression [47], which differs from OLS regression by including an additional regularization term in the loss function. This has the advantage of reducing model variance (i.e. lower prediction error, at the cost of slightly higher bias). Furthermore, the regularization term 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) [48], 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, Random Forest automatically considers interaction among the predictors, which Lasso regression falls short of. This may enable Random Forest to learn more fine-grained patterns from data.

We trained these three models on the training set, using 10-fold cross-validation and grid search for hyperparameter selection [49]. We ensured that the training-validation splits during the cross-validation procedure are done in such a way that the survey questions in a training partition are different from those in the corresponding validation set. In this way, the training objective stays consistent.

6.2.2 Evaluation Metric

We adopted Pearson's correlation r to evaluate the criterion validity of text embeddings. Specifically, we measured the Pearson's correlation between the predicted responses to survey questions and the observed responses. As a **prediction baseline** (as opposed to TF, TF-IDF and random embeddings, which are baselines for text embedding techniques), we used the average response of each participant in the training data as the prediction for that participant's responses in the test set.

6.3 Results

Table 4 summarizes the prediction performance of all text representation methods, measured as the average Pearson's correlation r across all 10 folds. $+\Delta\%$ in the parentheses indicates the percentage positive change in r in comparison to the prediction baseline r. Larger scores mean that the observed response scores are more

Table 4 Results of the Criterion Validity Analysis. r is the average Pearson's correlation between predicted and observed scores. $+\Delta\%$ in the parentheses indicates the percentage positive change in r in comparison to the baseline r: 0.187. "-" indicates a non-positive change. 95% CI refers to the 95% confidence interval around r.

	Lasso r $(+\Delta\%)$	Lasso 95% CI	RF r $(+\Delta\%)$	RF 95% CI
TF	0.106 (-)	[0.102, 0.110]	0.337 (80.007)	[0.333,0.341]
TF-IDF	0.092 (-)	[0.087, 0.096]	0.323 (72.830)	[0.319,0.327]
Random 300	0.149 (-)	[0.144, 0.153]	0.331 (77.066)	[0.327,0.335]
Random 768	0.116 (-)	[0.111, 0.120]	0.334 (78.614)	[0.330,0.338]
Random 1024	0.069 (-)	[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 (-)	[0.103, 0.111]	0.347 (85.664)	[0.343,0.351]
BERT-base-uncased	0.195 (-)	[0.191, 0.200]	0.411 (119.994)	[0.407,0.415]
BERT-large-uncased	0.151 (-)	[0.147, 0.155]	0.378 (102.260)	[0.374,0.382]
All-DistilRoBERTa	0.188 (-)	[0.183, 0.192]	0.374 (100.228)	[0.370,0.378]
All-MPNet-base	0.119 (-)	[0.115, 0.123]	0.406 (117.135)	[0.402,0.410]
USE	0.186 (-)	[0.182, 0.191]	0.386 (106.272)	[0.382,0.390]

correlated with the predicted response scores than with the prediction baseline. The more positive the value of $\Delta\%$, the more evidence for criterion 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 two 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 much 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, except for fastText and GloVe, achieved considerably higher r scores than the baseline text representation methods. This suggests that criterion validity holds (to some extent) for the BERT-based and USE pretrained models.

In summary, we see that text embeddings exhibit some degree of criterion validity, when the criterion concerns the prediction of responses to survey questions. However, the level of criterion validity that can be demonstrated seems to highly depend on both the specific prediction algorithm and the embedding model used.

7 Conclusion and Discussion

In this paper, we argue that it is important to ensure that text embeddings are valid representations of the original texts. Therefore, we introduce a novel framework of construct validity analysis for text embeddings and demonstrate it on survey questions. More concretely, we investigated the content, convergent, discriminant and criterion validity of various popular text embedding models, including fastText, GloVe, BERT, Sentence-BERT and USE.

For content validity analysis, we proposed using probing classifiers from the field of interpretable NLP. Evidence of content validity is indicated by a positive difference in the classifier's performance between using text embeddings and random

embeddings. For the analysis of convergent and discriminant validity, we recommend constructing minimal pairs where the texts only differ on a main semantic 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 criterion validity, we need to define a relevant prediction task whose observed scores (or gold standards) can be seen as the criterion. Then, we compute the correlation between the predicted and the observed scores. The higher the correlation, the more evidence for criterion validity.

When it comes to survey questions, we find that different text embedding techniques demonstrate different degrees of evidence for different types of construct validity. For instance, while USE and the two Sentence-BERT models ("All-DistilRoBERTa" and "All-MPNet-Base") show the best overall content validity, the original BERT models have the best performance 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. Both Sentence-BERT models and USE demonstrate high convergent and discriminant validity. They also, together with the two original BERT models, have overall the best scores for criterion validity. In contrast, fastText failed to achieve performance comparable to the BERT-based approaches and USE on all the validity analyses. In light of these findings, we urge researchers to examine and compare the construct validity of different text embedding models before deploying them in research. We also hope to see development in text embedding models that can further improve the encoding of abstract information like basic concepts.

It is also worth noting that our approach to criterion validity can be seen as an example of application of text embedding techniques to survey questions. We show that text embedding techniques, when used with the right prediction model, can help to substantially improve the prediction of individual responses to survey questions. This is an exciting result, but we should also realize that the best prediction score (r) is still only 0.411, which may be insufficient for production (e.g. in official statistics bureaus). We can likely improve this score by having a larger sample of survey questions.

A limitation of our study is that the synthetic survey question dataset and the ESS data set do not cover nearly most survey question types (in terms of, for instance, the concrete concepts, the formulation, the types of scale). This means that our findings concerning the construct validity of the adopted text embedding methods may only apply to the particular survey questions that we studied and cannot generalize to others.

For future research on the content validity of text embeddings for survey questions, we encourage researchers to probe more complex properties like double-barrelled questions, biases and social desirability. On the application side, we would like to see text embedding techniques being used for, for instance, 1) predicting the validity and reliability of new survey questions, 2) detecting problematic survey questions, and 3) generating new survey questions. We also encourage researchers to apply our construct validity analysis framework to other types of text data.

Appendix

Appendix A: Code and Data

To be added.

Appendix B: The 21 Basic Concepts

To be added.

Appendix C: The Six Types of Survey Question Formulation

To be added.

Appendix D: Sample Sizes

To be added.

Appendix E: Model Hyperparameters

To be added.

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

Author's contributions
DL.O proposed the research project; Q.F designed and performed the research; D.N and DL.O supervised the research; Q.F wrote the manuscript. All authors read, proofread and approved the final manuscript.

Acknowledgements
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).

Author details
[1]Department of Methodology & Statistics, Utrecht University, Padualaan 14, Utrecht, NL. [2]Department of Information & Computing Sciences, Utrecht University, Princetonplein 5, Utrecht, NL.

References

1. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality. In: Advances in Neural Information Processing Systems, vol. 26. Curran Associates, Inc., Lake Tahoe Nevada, the United States (2013)
2. Reimers, N., Gurevych, I.: 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), pp. 3982–3992. Association for Computational Linguistics, Hong Kong, China (2019)
3. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016)
4. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019)
5. Vu, H., Abdurahman, S., Bhatia, S., Ungar, L.: Predicting Responses to Psychological Questionnaires from Participants' Social Media Posts and Question Text Embeddings. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 1512–1524. Association for Computational Linguistics, Online (2020)
6. Matero, M., Idnani, A., Son, Y., Giorgi, S., Vu, H., Zamani, M., Limbachiya, P., Guntuku, S.C., Schwartz, H.A.: Suicide Risk Assessment with Multi-level Dual-Context Language and BERT. In: Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, pp. 39–44. Association for Computational Linguistics, Minneapolis, Minnesota (2019)
7. De Bruyne, L., De Clercq, O., Hoste, V.: 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, pp. 257–263. Association for Computational Linguistics, Online (2021)
8. Grandeit, P., Haberkern, C., Lang, M., Albrecht, J., Lehmann, R.: Using BERT for Qualitative Content Analysis in Psychosocial Online Counseling. In: Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science, pp. 11–23. Association for Computational Linguistics, Online (2020)
9. Garg, N., Schiebinger, L., Jurafsky, D., Zou, J.: Word embeddings quantify 100 years of gender and ethnic stereotypes. Proceedings of the National Academy of Sciences **115**(16), 3635–3644 (2018)
10. Linden, D.V.D., Nijenhuis, J.T., Bakker, A.B.: The general factor of personality: A meta-analysis of big five intercorrelations and a criterion-related validity study. Journal of Research in Personality **44**, 315–327 (2010)
11. Arora, S., May, A., Zhang, J., Ré, C.: Contextual embeddings: When are they worth it? In: ACL (2020)

12. Heale, R., Twycross, A.: Validity and reliability in quantitative studies. Evidence-Based Nursing **18**, 66–67 (2015)
13. Belinkov, Y.: Probing classifiers: Promises, shortcomings, and advances. Computational Linguistics (2021)
14. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to information retrieval. (2005)
15. Mikolov, T., Chen, K., Corrado, G.S., Dean, J.: Efficient estimation of word representations in vector space. In: ICLR (2013)
16. Wittgenstein, L.S.: Philosophical investigations = philosophische untersuchungen. (1958)
17. Harris, Z.S.: Distributional structure. WORD **10**, 146–162 (1954)
18. Mikolov, T., Yih, W.-t., Zweig, G.: Linguistic regularities in continuous space word representations. In: NAACL (2013)
19. Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science **356**, 183–186 (2017)
20. Rice, D., Rhodes, J.H., Nteta, T.M.: Racial bias in legal language. Research & Politics **6** (2019)
21. Kumar, V., Bhotia, T.S., Chakraborty, T.: Nurse is closer to woman than surgeon? mitigating gender-biased proximities in word embeddings. Transactions of the Association for Computational Linguistics **8**, 486–503 (2020)
22. Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., Joulin, A.: Advances in pre-training distributed word representations. ArXiv **abs/1712.09405** (2018)
23. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: EMNLP (2014)
24. Rogers, A., Kovaleva, O., Rumshisky, A.: A Primer in BERTology: What We Know About How BERT Works. Transactions of the Association for Computational Linguistics **8**, 842–866 (2020)
25. Cer, D.M., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., Kurzweil, R.: Universal sentence encoder. ArXiv **abs/1803.11175** (2018)
26. W, A.S., Pellegrini, A.M., Chan, S., Brown, H.E., Rosenquist, J.N., Vuijk, P.J., Doyle, A.E., Perlis, R.H., Cai, T.: Integrating questionnaire measures for transdiagnostic psychiatric phenotyping using word2vec. PLoS ONE **15** (2020)
27. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs] (2019)
28. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. ArXiv **abs/1910.01108** (2019)
29. Song, K., Tan, X., Qin, T., Lu, J., Liu, T.-Y.: MPNet: Masked and Permuted Pre-training for Language Understanding. arXiv:2004.09297 [cs] (2020)
30. Tawfik, N.S., Spruit, M.R.: Evaluating sentence representations for biomedical text: Methods and experimental results. Journal of biomedical informatics, 103396 (2020)
31. Rücklé, A., Eger, S., Peyrard, M., Gurevych, I.: Concatenated p-mean word embeddings as universal cross-lingual sentence representations. ArXiv **abs/1803.01400** (2018)
32. Saris, W.E., Gallhofer, I.N.: Design, Evaluation, and Analysis of Questionnaires for Survey Research. Design, evaluation, and analysis of questionnaires for survey research. Wiley-Interscience, Hoboken, NJ, US (2007)
33. Yan, T., Tourangeau, R.: Fast times and easy questions: the effects of age, experience and question complexity on web survey response times. Applied Cognitive Psychology **22**, 51–68 (2008)
34. Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., Bowman, S.R.: Blimp: A benchmark of linguistic minimal pairs for english. Transactions of the Association for Computational Linguistics **8**, 377–392 (2020)
35. Hupkes, D., Zuidema, W.H.: Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. ArXiv **abs/1711.10203** (2018)
36. Hewitt, J., Liang, P.: Designing and interpreting probes with control tasks. ArXiv **abs/1909.03368** (2019)
37. Alain, G., Bengio, Y.: Understanding intermediate layers using linear classifier probes. ArXiv **abs/1610.01644** (2017)
38. Liu, N.F., Gardner, M., Belinkov, Y., Peters, M.E., Smith, N.A.: Linguistic knowledge and transferability of contextual representations. ArXiv **abs/1903.08855** (2019)
39. Maudslay, R.H., Valvoda, J., Pimentel, T., Williams, A., Cotterell, R.: A tale of a probe and a parser. In: ACL (2020)
40. Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H., Glass, J.R.: What do neural machine translation models learn about morphology? In: ACL (2017)
41. Conneau, A., Kruszewski, G., Lample, G., Barrault, L., Baroni, M.: What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In: ACL (2018)
42. Zhang, K.W., Bowman, S.R.: Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. In: BlackboxNLP@EMNLP (2018)
43. Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R.T., Kim, N., Durme, B.V., Bowman, S.R., Das, D., Pavlick, E.: What do you learn from context? probing for sentence structure in contextualized word representations. ArXiv **abs/1905.06316** (2019)
44. Belinkov, Y., Glass, J.R.: Analysis methods in neural language processing: A survey. Transactions of the Association for Computational Linguistics **7**, 49–72 (2019)
45. Norwegian Centre for Research Data: European Social Survey Round 9 Data. Data File Edition 3.1. Norway (2018). doi:10.21338/NSD-ESS9-2018. Norwegian Centre for Research Data
46. Norwegian Centre for Research Data: European Social Survey: ESS-9 2018 Documentation Report. Edition 3.1. Norway (2021). doi:10.21338/NSD-ESS9-2018. Norwegian Centre for Research Data
47. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the royal statistical society series b-methodological **58**, 267–288 (1996)
48. Breiman, L.: Random forests. Machine Learning **45**, 5–32 (2001)
49. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer Series in Statistics.

Springer, New York, NY (2009)