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 embeddings models from Natural Language Processing can map text data (e.g. words, sentences, documents) to supposedly meaningful numerical representations called text embeddings. When applied to survey questions, such techniques can benefit survey-based research in many ways, like automating the prediction of responses to new survey questions. However, it is unclear whether such representations of survey questions are valid. Therefore, in our study, we investigate the construct validity of various popular text embeddings methods (e.g. fastText, GloVe, BERT, Sentence-BERT, Universal Sentence Encoder). We specifically focus on four types of construct validity: content, convergent, discriminant and criterion validity. For the analysis of content validity, we "probe" the text embeddings to see if they encode relevant properties of survey questions like the underlying concepts and the formulation. For convergent and discriminant validity, we inspect whether the cosine similarity of the embeddings of survey questions is indicative of the conceptual similarity between survey questions. As for criterion validity, we examine whether the use of text embeddings techniques can improve the prediction of responses to unseen survey questions. We show that despite several limitations, BERT-based embeddings techniques and Universal Sentence Encoder provide more valid representations of survey questions than do others.

**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 embeddings models, which originate from the field of Natural Language Processing (NLP), can map texts (e.g. words, sentences, articles) into 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 are made publicly available (e.g. [1, 2, 3, 4])). This allows other researchers to obtain high-quality off-the-shelf pretrained text embeddings that can be readily used for downstream applications, without the need to spend many computational resources on training the models from scratch (e.g. [5, 6]). Researchers can also further train (i.e. fine-tune) the text embeddings models on additional domain-specific data for even better performance (e.g. [7, 8, 9, 10]).

Increasingly, text embeddings are being applied to survey questions. For instance, quite recently, [5] used a text embeddings model called BERT (Bidirectional Encoder Representations from Transformers) to encode participants' social media posts and the questions from the Big-Five 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. [11] used the skip-gram embedding algorithm 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.

These applications show the promising potential of text embeddings for survey-based research. However, it is unclear whether text embeddings can be high-quality representations of survey questions. For instance, do the embeddings encode relevant information about a survey question, such as the underlying construct of interest and the formulation? Are the embeddings of two conceptually different survey questions located sufficiently far away from each other in the vector space?

The quality of text embeddings for survey questions is an important topic to study, whereby researchers can learn about the strengths and weakness of text embeddings models for survey-related applications. This can also provide helpful directions for the future development of text embeddings methods.

In this paper, we focus on investigating one crucial quality aspect of text embeddings for survey questions: **construct validity**[1]. Construct validity refers to the extent 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 actually measures anxiety would not be considered a valid instrument. In our study, we treat a survey question as the construct of interest and the corresponding embedding as the operationalization. It is therefore important that the embedding actually reflects the original survey question text well.

We consider four types of validity [12]:

1 **Content validity** concerns whether the operationalization adequately covers all the 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 measures theoretically dissimilar constructs. For example, there should be a low correlation

[1]The other aspect of measurement quality is reliability, concerning the consistency of measurements across occasions or raters. This is out of the scope of this study.

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 content, convergent, discriminant and criterion validity of state-of-the-art text embeddings models for survey questions. We focus on pretrained embeddings because of their popularity and convenience. However, our approach to construct validity also applies to fine-tuned embeddings.

Construct validity analysis typically relies on visually examining the measure of interest (i.e. content validity) and evaluating the size of the correlation of scores between two self-reported measures against some expected values (i.e. the other three validity types). However, this approach does not apply to text embeddings, because they are high-dimensional data with opaque features that cannot be directly interpreted and that there are no responses provided by human participants. Therefore, an alternative analytic approach to construct validity is required.

To this end, we propose a novel framework of construct validity analysis for such high-dimensional measures. Namely, we draw from the field of interpretable machine learning & NLP and connect it to the construct validity literature, by borrowing tools like "probing classifiers" and adapting them to our specific research problem. We apply this framework and uncover the strengths and limitations of different text embeddings 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 embeddings techniques for survey research.

Last but not least, 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, language models specific for survey research.

## 2 Background

### 2.1 Classic Count-based Text Representation Techniques

Prior to the introduction of text embeddings techniques, two popular count-based approaches to text representation were: bag-of-words (BOW) and term frequency-inverse document frequency (TF-IDF) [13].

BOW represents a piece of text by describing the occurrence of pre-defined words within the text. Take the survey question "How happy would you say you are?" as an example: Using BOW, we can represent this question as a vector [1, 1, 1, 1, 2, 1], with the numbers corresponding to the frequency of "how", "happy", "would", "you", "say" and "are", respectively. This is illustrated in Figure 1.

However, a problem with word frequency is that highly frequent words (e.g. "the", "and") are not necessarily important words. 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: $w_{i,j} = tf_{i,j} * log(N/df_i)$, where $tf_{i,j}$

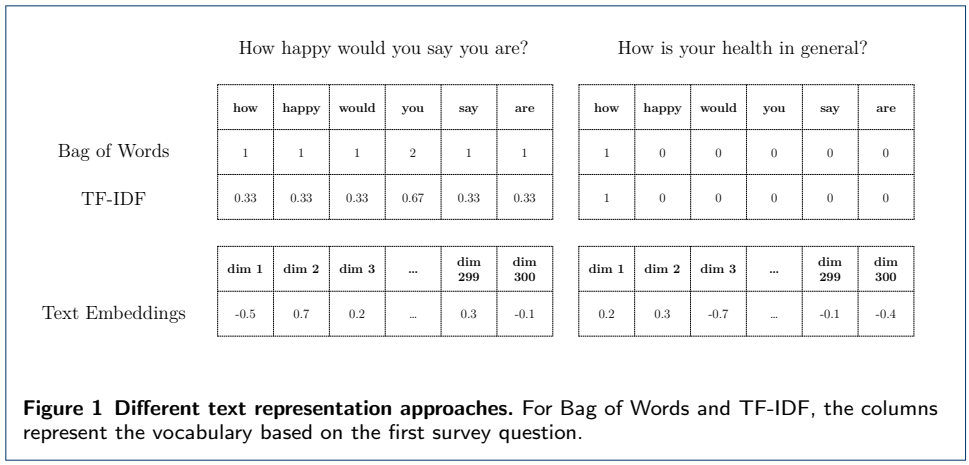


**Figure 1 Different text representation approaches.** For Bag of Words and TF-IDF, the columns represent the vocabulary based on the first survey question.

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 book, etc.

We can see that both approaches share many common strengths and weaknesses. For example, they are both simple and efficient, but this comes as the cost of disregarding 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 BOW and TF-IDF representation, meaning that a large amount of meaningful information from question 2 is lost. We show in the next section that this is less of an issue for modern text embeddings techniques.

## 2.2 Text Embeddings Techniques

### *2.2.1 fastText*

One famous family of text embeddings algorithms is word2vec [1, 14]. 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 [15, 16] that words that occur in similar contents tend to have similar meanings, closely located words in the vector space are expected to be semantically similar, which can be indicated by the cosine similarity between two word vectors.

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 have outperformed its word2vec predecessors across various benchmarks [17].

Compared to BOW and TF-IDF, there are several advantages of word2vec models. First, word2vec produces 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) in texts [19, 20, 21, 22].

Second, like many other text embeddings techniques, pretrained word2vec models are made publicly available (even in various languages). This makes it convenient for researchers to obtain high-quality off-the-shelf text embeddings without having to train the models from scratch, which can be computationally expensive.

Word2vec, like any modern text embeddings technique, suffers from a clear disadvantage compared to classical approaches like BOW and TF-IDF. Namely, the embedding dimensions themselves are not directly interpretable (see Figure 1). Nevertheless, there are methods to probe the information encoded in each dimension, despite requiring extensive research effort [23, 24, 25].

#### *2.2.2 GloVe*

GloVe, which stands for Gloval Vector word representations [26], 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 embeddings techniques have recently become available. A popular and effective one is Sentence-BERT [2], where BERT 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, focuses on sentence- or document-level text representations, and is trained with different objective functions.

BERT and its variants have 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]. The embeddings generated by BERT have been shown to encode syntactic and semantic knowledge about the original texts [27].

Sentence-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].

Like word2vec, many pretrained BERT models are freely and publicly available for use. They differ (to varying degrees) in their model architecture, the training

schema and the training data, allowing them to specialize in different application settings. There is, however, not yet a single pretrained BERT or Sentence-BERT model for survey question representation.

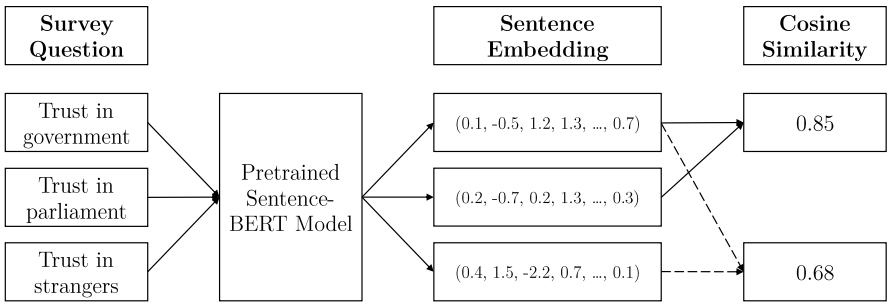


**Figure 2 Illustrating the Use of Pre-trained Sentence-BERT Models**

Figure 2 illustrates how pretrained Sentence-BERT models can be used. For instance, given three survey questions that intend to measure three separate concepts ("trust in government", "trust in parliament" and "trust in strangers"), we can feed them into a pretrained Sentence-BERT model and in return obtain three sentence embeddings that supposedly represent the original question texts. The first two concepts are the most related because both concern institutional trust, while the third one measures generalized trust. Therefore, we would expect the embeddings of the first two questions to be more closely located in the embedding space than, for instance, the embeddings of the first and third questions.

We can measure similarity between two embeddings using cosine similarity, which is a measure of similarity 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.

#### *2.2.4 Universal Sentence Encoder*

Universal Sentence Encoder (USE) is another text embeddings model meant for greater-than-word length texts like sentences, phrases and short paragraphs [28]. USE utilizes 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 [28]. 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 Research Setup

## 3.1 Pretrained Text Embeddings Models

In this paper, we investigate whether text embeddings techniques like fastText, GloVe, BERT and USE can produce valid representations for survey questions. Table 1 summarizes the specific pretrained embedding models we adopted.

**Table 1** Pretrained Text Embeddings Models

| 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 |

Specifically, we use the fastText pretrained model developed by [3]. It is trained on Common Crawl and Wikipedia, and produces word embeddings with 300 dimensions. As for GloVe, we used the model pretrained on Common Crawl with 840B tokens and a vocabulary of 2.2 million words. It also gives 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-V1"" and "All-MPNet-base-V2"", where "DistilRoBERTa" [29, 30] and "MPNet" [31] are two different improved versions 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" (size: 420 MB, dimension: 768) and "BERT-large-uncased" (size: 1.25 GB, dimension: 1024). "Uncased" refers to BERT treating upper and lower cases equally.

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

## 3.2 Aggregating Word-level to Sentence-Level Embeddings

Because word2vec and GloVe models only produce word embeddings, it is necessary to combine the word embeddings into a higher-level representation when the corresponding texts are sentences (like survey questions) or documents. 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 or approximate the performance [32] of more sophisticated methods [33]. Therefore, we use simple averaging to compute sentence-level embeddings for survey questions from fastText and GloVe word embeddings. The resulting representations have the same number of dimensions, 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.

For the original BERT models, we follow the advice of [27, 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.

### 3.3 Analysis Overview

In the remainder of the paper, we will describe how we examine the content validity (Section 4), convergent & discriminant validity (Section 5) and criterion validity (Section 6) of various text representation approaches for survey questions. In each section, we will also describe how we adapt the traditional validity assessment framework in a novel manner to the high-dimensional, opaque nature of text embeddings.

## 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. However, we will consider several important ones.

The most obvious are the underlying **concepts**. According to the typology proposed by [34], most survey questions can be categorized into one of 21 so-called **basic concepts**, such as "feelings", "cognitive judgement" and "expectations". 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 question formulation often apply in survey research: direct request (DR), imperative-interrogative request (ImIn), interrogative-interrogative request (InIn), declarative-interrogative request (DeIn) and interrogative-declarative request (InDe)[2].

**Complexity** is another important aspect of survey questions which can affect how respondents understand and answer a survey question [35]. It is often measured as the length of a survey question [34].

Therefore, we investigate whether text embeddings encode information about the following aspects of survey questions: basic concepts, concrete concepts, formulation and length [3]. We refer to them as **properties** in the remainder of the paper.

### 4.1 Data

We constructed a synthetic data set because in this way, we can have much better control over the properties of the survey questions. In comparison, real survey questions from established surveys (like European Social Survey [36]) often have

[2] Actually, [34] 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.

[3] There are other relevant aspects, such as whether a question is double-barrelled and whether it concerns a sensitive topic. We keep the investigation to the four mentioned because they apply to all survey questions.

correlated properties, which makes it difficult for the analysis of content validity (as well as convergent and discriminant analysis).

Specifically, the data set should satisfy two requirements. First, it should cover a wide, (hopefully) 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. This idea is analogous to the so-called "minimal pairs" in NLP.

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 [34], 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". We thus made sure that the questions in our data set covered 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[4]. Finally, for each concrete concept, we created survey questions that vary in their formulation. [34] 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 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 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).

**Table 2** Example Questions from the Survey Question Data Set

| 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? |

[4] This is done based on our judgment and experience working in the field of survey research

## 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 [37].

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 [38, 39, 40, 41, 42]. Furthermore, it is recommended to always include baselines for comparison [37]. 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 [43, 44, 45, 46], 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 [47]. 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. Namely, using chi-square tests of independence, sentence length is statistically significantly related to basic concepts ($\chi^2 = 3636.7, df = 36, p < 0.01$), concrete concepts ($\chi^2 = 4612.1, df = 348, p < 0.01$) and formulation ($\chi^2 = 1252.7, df = 12, p < 0.01$). 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 we used, 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, we can more confidently attribute any positive difference we observe between the performance of the probing classifier and that of the baseline using random embeddings 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 embeddings model encodes more information about a property than does another model.

### 4.3 Results

**Table 3** Results: Analysis of Content Validity

| | 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 |
| BOW | 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 random embeddings of three dimension sizes, BOW and TF-IDF vectors, 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 serve as an additional baseline.

If the classifier performs better on a particular type of text embeddings than on the baselines (i.e. simple majority and random embeddings) for a survey question property, we can conclude that the corresponding text embeddings of survey questions likely encode information about that specific property.

For sentence length, we can see that the BERT-based and the USE embeddings perform better than all the four baselines as well as BOW and TF-IDF, 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. This is a surprising finding, given that the aggregating word-level embeddings is unlikely to preserve any information about sentence length, which is also supported by the poor performance of both fastText and GloVe.

For basic concepts, none of the pretrained text embeddings seems able to beat the performance of the baseline random embeddings. 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 [34] are too abstract for current embeddings 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 due to the Sentence-BERT embeddings having 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 BOW 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 embeddings 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 embeddings 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. Therefore, we analyse and discuss convergent and discriminant validity jointly.

### 5.1 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 should be able to identify conceptually similar survey questions while differentiating between conceptually dissimilar ones. Two hypotheses naturally follow:

For 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.

For Hypothesis 2, we expect that cosine similarity scores are 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. $E_{\text{similar}}$). 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.

### 5.2 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

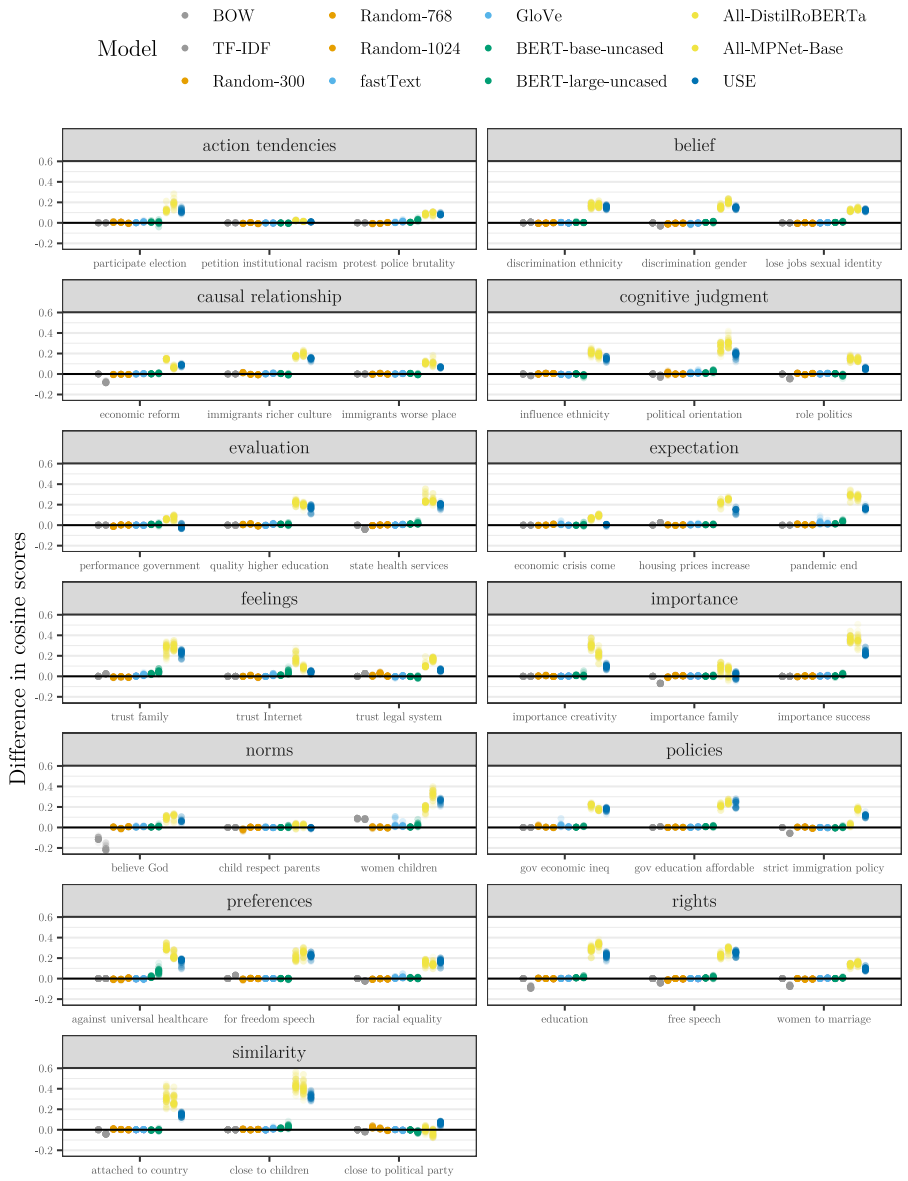


**Figure 3 The distribution of the cosine similarity difference scores for Hypothesis 1 across 13 basic concepts.**

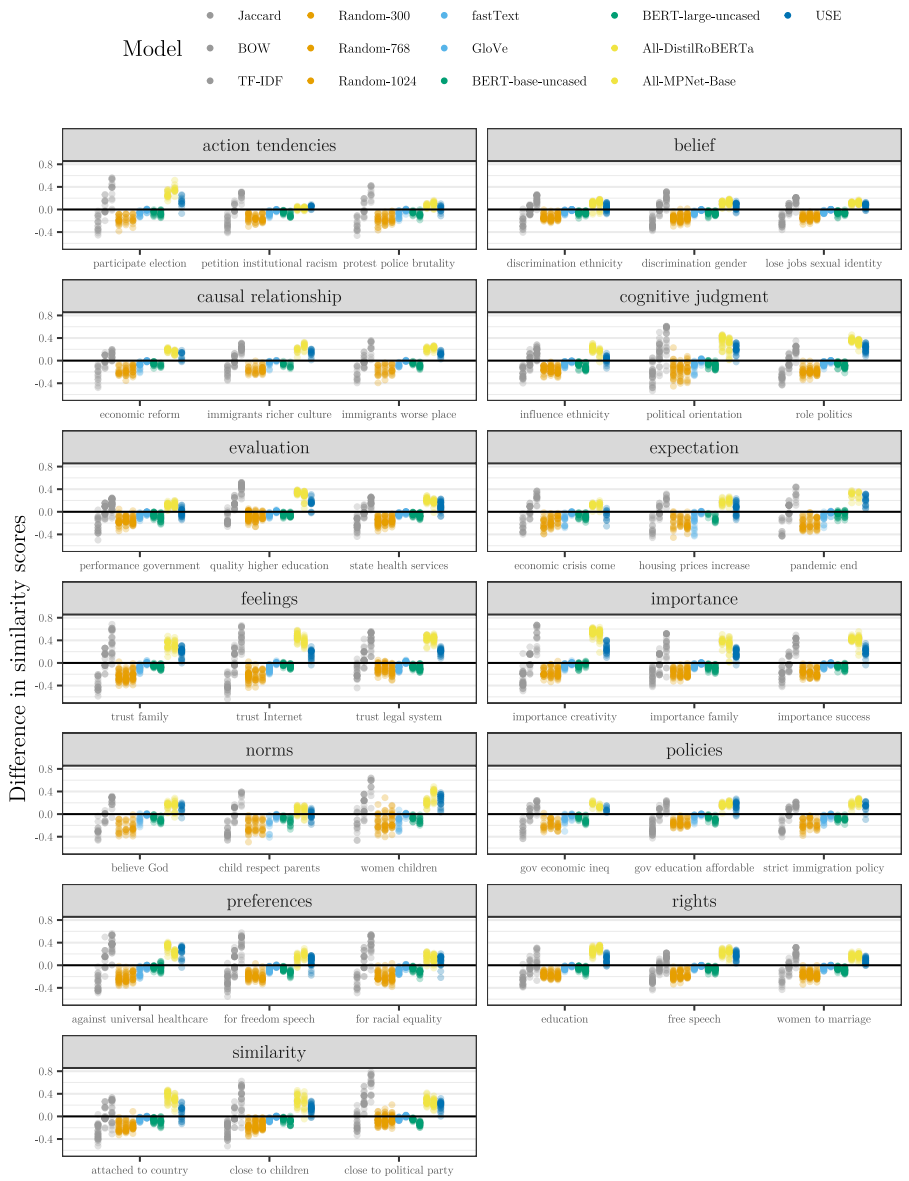


**Figure 4 The distribution of the cosine similarity difference scores for Hypothesis 2 across 13 basic concepts.**

concepts, various baselines and text embeddings approaches. 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 scores, shown as points, are additionally coded in different colour groups (following a colour-blind friendly palette) to ease the reading and interpretation of the figure. The x-axis labels are the (abbreviated) names of the main concrete concepts in the question triads.

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. BOW, 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 text embeddings, one other possible explanation is that cosine similarity might not be a suitable measure, as previous 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. The y-axis indicates the size and direction of the differences. The larger it is, the more evidence for convergent and discriminant analysis. The x-axis labels are the (abbreviated) names of the main concrete concepts in the question triads.

Note that here we also use Jaccard similarity 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].

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 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 the Sentence-BERT and USE embeddings demonstrate convergent and discriminant validity. Meanwhile, there is not enough evidence to suggest that this also applies to 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. Concretely, 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 [48]. The ESS is a research-oriented cross-national survey that is conducted with newly selected, cross-sectional samples every two years since 2001 [36]. 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. This way, we avoid multilingual issues and simplify the research task.

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 embeddings methods. Namely, our data set now consists of the following features/columns: background variables, embeddings of the survey questions generated by text embeddings 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 BOW, TF-IDF and random embeddings of dimension 300, 768 and 1024 as baseline text representation techniques for the text embeddings approaches to compare themselves to. Note that with these baseline approaches, we

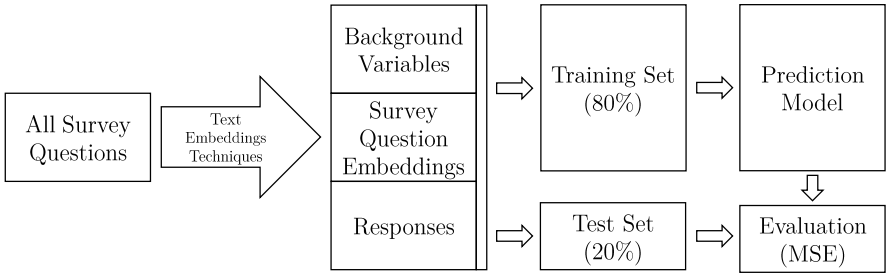


**Figure 5 Prediction Modelling Approach for the Analysis of Criterion Validity.** Note that this exact approach applies only to the text embeddings methods (i.e. excluding BOW, TF-IDF and random embeddings).

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.

### *6.2.1 Lasso and Random Forest*

We adopted two popular prediction models. The first is Lasso regression [49], 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) [50], 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 and fine-tuned these three models on the training set using 10-fold cross-validation and grid search [51]. 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 fine-tuning objective is consistent with the overall training-testing goal.

### *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. To establish a baseline for Pearson's r for the text representation approaches to compare to, we used the average response of each participant in the training data as the prediction for the corresponding participant's responses in the test set.

**Table 4** Results: Analysis of Criterion Validity

| | Baseline $r$ | Lasso $r$ | Lasso $+\Delta\%$ | RF $r$ | RF $+\Delta\%$ |
|---|---|---|---|---|---|
| BOW | 0.187 (0.040) | 0.106 (0.216) | - | 0.337 (0.116) | 80.007 |
| TF-IDF | 0.187 (0.040) | 0.092 (0.231) | - | 0.323 (0.141) | 72.830 |
| Random 300 | 0.187 (0.040) | 0.149 (0.162) | - | 0.331 (0.142) | 77.066 |
| Random 768 | 0.187 (0.040) | 0.116 (0.159) | - | 0.334 (0.126) | 78.614 |
| Random 1024 | 0.187 (0.040) | 0.069 (0.181) | - | 0.338 (0.121) | 80.520 |
| fastText | 0.187 (0.040) | 0.204 (0.121) | 9.261 | 0.356 (0.143) | 90.439 |
| GloVe | 0.187 (0.040) | 0.107 (0.131) | - | 0.347 (0.129) | 85.664 |
| BERT-base-uncased | 0.187 (0.040) | 0.195 (0.204) | - | 0.411 (0.109) | 119.994 |
| BERT-large-uncased | 0.187 (0.040) | 0.151 (0.120) | - | 0.378 (0.126) | 102.260 |
| All-DistilRoBERTa | 0.187 (0.040) | 0.188 (0.159) | - | 0.374 (0.139) | 100.228 |
| All-MPNet-base | 0.187 (0.040) | 0.119 (0.173) | - | 0.406 (0.132) | 117.135 |
| USE | 0.187 (0.040) | 0.186 (0.235) | - | 0.386 (0.143) | 106.272 |

### 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. The numbers in parentheses next to the $r$ scores indicate the 10-fold standard error of $r$. The columns with $\Delta\%$ in the title indicate the percentage positive change in $r$ in comparison to the baseline $r$ (0.187). "-" indicates a non-positive change. Larger scores mean that the observed response scores are more correlated with the predicted response scores than with the baseline. The more positive the value of $\Delta\%$, the more evidence for criterion validity.

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 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 can see that simply using RF with baseline methods (BOW, TF-IDF and random embeddings) can already lead to substantial improvement in prediction compared to the baseline performance. All the embeddings approaches, except for fastText and GloVe, achieved considerable higher $r$ scores than the baseline methods. This suggests that criterion validity holds (to some extent) for the BERT-based and USE pretrained models

Third, all the Lasso and RF $r$ scores are accompanied by relatively large standard errors, suggesting that prediction performance highly varies across different folds of the data. This is not unexpected, as there are in total only 94 unique survey questions in the data, giving way to large data variance and model uncertainty. Nevertheless, the large standard errors do not invalidate our previous conclusions on the criterion validity of different text embeddings model, because the models that perform well on average across the 10 folds of the data also tend to outperform the other models given a specific data fold.

In summary, we see that text embedding approaches 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 the specific prediction model used.

## 7 Conclusion and Discussion

In this paper, we introduce a novel framework of construct validity analysis for high-dimensional data like text embeddings and demonstrate it on the case of survey questions. We argue that it is important to ensure that text embeddings are valid representations of survey questions. Specifically, we investigated the content, convergent, discriminant and criterion validity of various text embeddings methods, including fastText, GloVe, BERT, Sentence-BERT and USE.

We find that different text embeddings techniques demonstrate different degrees of evidence for different types of 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 in 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 and similar 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 such findings, we urge researchers to examine and compare the construct validity of different text embeddings models before deploying them in research. We also hope to see development in text embeddings 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 embeddings techniques to survey questions. We show that text embeddings 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, for instance, 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 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 techniques, we encourage researchers to probe more complex properties like double-barrelled questions, biased questions and social desirability. On the application side, we would like to see text embeddings 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 and extend our construct validity analysis framework to other types of texts, such as Tweets, poems and lyrics.

**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. Joseph, K., Morgan, J.H.: When do word embeddings accurately reflect surveys on our beliefs about people? In: ACL (2020)
7. Abbasi, A., Dobolyi, D.G., Netemeyer, R.G.: Constructing a testbed for psychometric natural language processing. ArXiv **abs/2007.12969** (2020)
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. 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)
10. 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)
11. 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)
12. Heale, R., Twycross, A.: Validity and reliability in quantitative studies. Evidence-Based Nursing **18**, 66–67 (2015)
13. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to information retrieval. (2005)
14. Mikolov, T., Chen, K., Corrado, G.S., Dean, J.: Efficient estimation of word representations in vector space. In: ICLR (2013)
15. Wittgenstein, L.S.: Philosophical investigations = philosophische untersuchungen. (1958)
16. Harris, Z.S.: Distributional structure. WORD **10**, 146–162 (1954)
17. Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., Joulin, A.: Advances in pre-training distributed word representations. ArXiv **abs/1712.09405** (2018)
18. Mikolov, T., Yih, W.-t., Zweig, G.: Linguistic regularities in continuous space word representations. In: NAACL (2013)
19. 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)
20. Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science **356**, 183–186 (2017)
21. Rice, D., Rhodes, J.H., Nteta, T.M.: Racial bias in legal language. Research & Politics **6** (2019)
22. 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)
23. Senel, L.K., Utlu, I., Yücesoy, V., Koç, A., Çukur, T.: Semantic structure and interpretability of word embeddings. IEEE/ACM Transactions on Audio, Speech, and Language Processing **26**, 1769–1779 (2018)
24. Allen, C., Hospedales, T.M.: Analogies explained: Towards understanding word embeddings. ArXiv **abs/1901.09813** (2019)
25. Shin, J., Madotto, A., Fung, P.: Interpreting word embeddings with eigenvector analysis. (2018)
26. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: EMNLP (2014)

27. 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)
28. 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)
29. 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)
30. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. ArXiv **abs/1910.01108** (2019)
31. 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)
32. Tawfik, N.S., Spruit, M.R.: Evaluating sentence representations for biomedical text: Methods and experimental results. Journal of biomedical informatics, 103396 (2020)
33. Rücklé, A., Eger, S., Peyrard, M., Gurevych, I.: Concatenated p-mean word embeddings as universal cross-lingual sentence representations. ArXiv **abs/1803.01400** (2018)
34. 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)
35. 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)
36. 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
37. Belinkov, Y.: Probing classifiers: Promises, shortcomings, and advances. Computational Linguistics (2021)
38. Hupkes, D., Zuidema, W.H.: Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. ArXiv **abs/1711.10203** (2018)
39. Hewitt, J., Liang, P.: Designing and interpreting probes with control tasks. ArXiv **abs/1909.03368** (2019)
40. Alain, G., Bengio, Y.: Understanding intermediate layers using linear classifier probes. ArXiv **abs/1610.01644** (2017)
41. 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)
42. Maudslay, R.H., Valvoda, J., Pimentel, T., Williams, A., Cotterell, R.: A tale of a probe and a parser. In: ACL (2020)
43. Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H., Glass, J.R.: What do neural machine translation models learn about morphology? In: ACL (2017)
44. 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)
45. 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)
46. 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)
47. Belinkov, Y., Glass, J.R.: Analysis methods in neural language processing: A survey. Transactions of the Association for Computational Linguistics **7**, 49–72 (2019)
48. 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
49. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the royal statistical society series b-methodological **58**, 267–288 (1996)
50. Breiman, L.: Random forests. Machine Learning **45**, 5–32 (2001)
51. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer Series in Statistics. Springer, New York, NY (2009)