RESEARCH

Evaluating Pretrained Sentence Embeddings for Survey Research - A Study of Measurement Validity

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

First part title: Text for this section.

Second part title: Text for this section.

Keywords: word embeddings; sentence embeddings; measurement validity; survey questions; computational social science; survey methodology

1 Introduction

Text embeddings techniques, which originate from the field of natural language processing (NLP), are used to map texts (e.g. words, sentences, articles) into semantically meaningful, numeric vectors (i.e. embeddings) (e.g. [1, 2]). This means that, for instance, the embeddings of similar texts (e.g. words like big and large) should be closer to one another than those of dissimilar ones (e.g. words like big and small), as measurable by some distance function (e.g. cosine distance).

Their potential for meaningful text representation has made text embeddings techniques increasingly popular and attracted a growing number of applications in various social science disciplines outside of NLP. For instance, text embeddings have been employed to encode - social media posts for the purpose of suicide risk assessment [3]; Tweets and TV captions for emotion detection [4]; historical texts to quantify societal trends of gender and ethnic stereotypes in the US [5]; interview data for automatic qualitative content analysis [6].

We believe that text embeddings techniques also offer exciting opportunities and directions for survey methodology research. Take modelling and predicting individual-level responses to survey questions as an example (see Figure 1): to build a model capable of estimating responses to new (or even just slightly modified) survey questions, we need to incorporate as predictors not only covariates like demographics and socio-economic indicators, but also the texts of survey questions. However, transforming survey questions into some numerical representation usable by the model remained for a long time a challenge, because the representation needs to be both efficient (i.e. not requiring manual coding) and fine-grained (i.e. encoding fine details like the underlying survey concepts and linguistic styles). This was not possible with traditional approaches (like manual feature extraction) until the introduction of modern text embeddings techniques.

Taking this idea further, we may be able to better estimate effect sizes for the purpose of power analysis prior to data collection using (new) survey instruments.

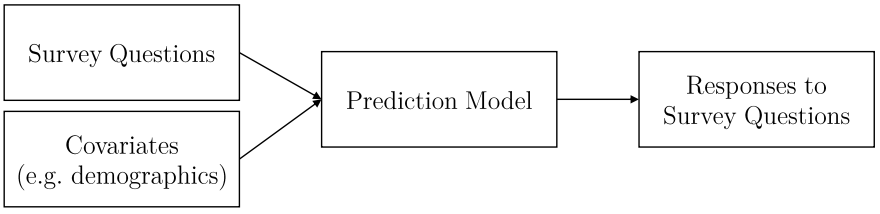

Figure 1 Response Prediction Modelling

We can also swap the outcome variable in Figure 1 for quality indicators of survey questions. In this way, we may be able to automate quality control of new survey questions without conducting expensive experiments, on the premise that we have the available data to train such a model.

Naturally, we are not the only ones who have noticed the potential of text embeddings techniques for survey research. Indeed, very recently, [7] used a text embeddings technique called BERT (Bidirectional Encoder Representations from Transformers) to encode participants' social media posts and the Big-Five questionnaire. They showed that by making use of the generated 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. To the best of our knowledge, this is the only study that has applied text embeddings techniques to the study of survey questions.

Despite the promising potential of text embeddings techniques for survey research, one important question remains to be investigated: **are text embeddings valid representation of survey questions?**

Survey questions are often intended as measures for abstract concepts (or constructs; e.g. trust, emotion, ideology) that are not directly observable ([8]). Because of this, survey questions are subject to measurement quality issues like validity[1], which concerns how well a measure captures the intended construct of interest [9]. A good survey question, therefore, should demonstrate high validity. Likewise, a good transformed representation of survey questions should also preserve the validity of the original survey question text. This entails, for instance, that the new representation encodes all relevant information like the underlying construct of interest, the complexity of the question and the linguistic style. In other words, text embeddings should be valid measures of survey questions to 'qualify' for subsequent use.

Therefore, in this paper, we take a step back from the application side of the story. Instead, we carefully examine the validity of text embeddings of survey questions. We argue that this is an important question to answer before we proceed with any application (so that we can better move on to application research). Specifically, we deal with the following research question: Can state-of-the-art text embeddings techniques provide valid representation of survey questions? We look at four crucial validity indicators: content validity, convergent validity, discriminant validity, and criterion validity.

[1]The other aspect of measurement quality is reliability, which is out of the scope of this work.

The contribution of this study is four-fold: First, we curate a data set of 5436 survey questions covering various survey concepts and linguistic styles. These questions are designed in a way that pairs of questions that differ in only one feature (e.g. concepts, linguistic styles) can be identified and thus allow for assessing the validity of text embeddings. Second, we showcase a novel framework to assess the validity of text embeddings, which are unintuitive, high-dimensional data that cannot be accommodated for by traditional validity assessment approaches. Third, we assess the validity of state-of-the-art text embeddings of survey questions, which informs about the strengths and limitations of text embeddings techniques for survey research. Lastly, we discuss the potential applications and future directions of text embeddings techniques for survey research.

2 Background

2.1 Text Embedding Techniques

Text for this sub-heading ...

2.2 Measurement Validity

Text for this sub-sub-heading ...

3 Research Setup

4 Analysis of Content Validity

4.1 Methods

4.2 Data

4.3 Results

Table 1 Results: Analysis of Content Validity

	Length	Basic Concept	Concrete Concept	Form of Request
Baseline	0.39	0.01	0.03 (0.03)	0.25
Random 300	0.19	0.20	0.40 (0.43)	0.74
Random 768	0.14	0.20	0.58 (0.45)	0.71
Random 1024	0.06	0.20	0.52 (0.45)	0.75
FastText	0.06	0.18	0.76 (-)	0.67
BERT-base-uncased	0.65	0.18	0.81 (-)	0.94
BERT-large-uncased	0.54	0.15	0.74 (-)	0.90
PI-MPNet-base	0.47	0.20	0.97 (-)	0.80
STSB-MPNet-base	0.53	0.20	0.94 (-)	0.82
STSB-RoBERTa-base	0.47	0.20	0.92 (-)	0.77
STSB-RoBERTa-large	0.37	0.20	0.86 (-)	0.73

Competing interests

The authors declare that they have no competing interests.

Author's contributions

DL.O proposed research; Q.F designed and performed 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. 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)
4. 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)
5. 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)
6. 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)
7. 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)
8. Hox, J.J.: From theoretical concept to survey question. In: Lyberg, L.E., Biemer, P.P., Collins, M., de Leeuw, E.D., Dippo, C.S., Schwarz, N., Trewin, D. (eds.) Survey Measurement and Process Quality, pp. 47–69. John Wiley & Sons, Inc., online (2012)
9. Cronbach, L.J., Meehl, P.E.: Construct validity in psychological tests. Psychological bulletin **52 4**, 281–302 (1955)

Figures

Additional Files

Additional file 1 — Sample additional file title

Additional file descriptions text (including details of how to view the file, if it is in a non-standard format or the file extension). This might refer to a multi-page table or a figure.

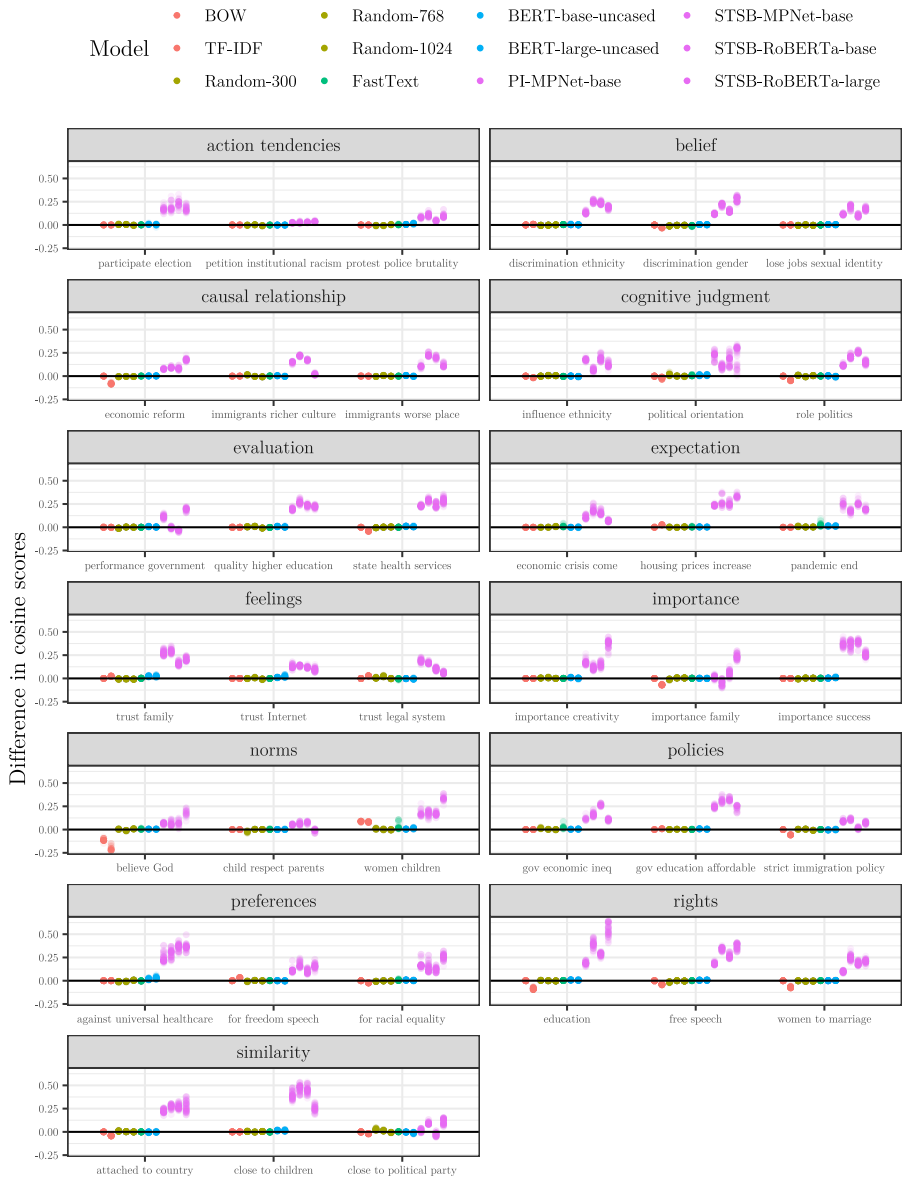

Figure 2 The distribution of the cosine similarity difference scores for Hypothesis 1, across 13 basic concepts. A short description of the figure content should go here.

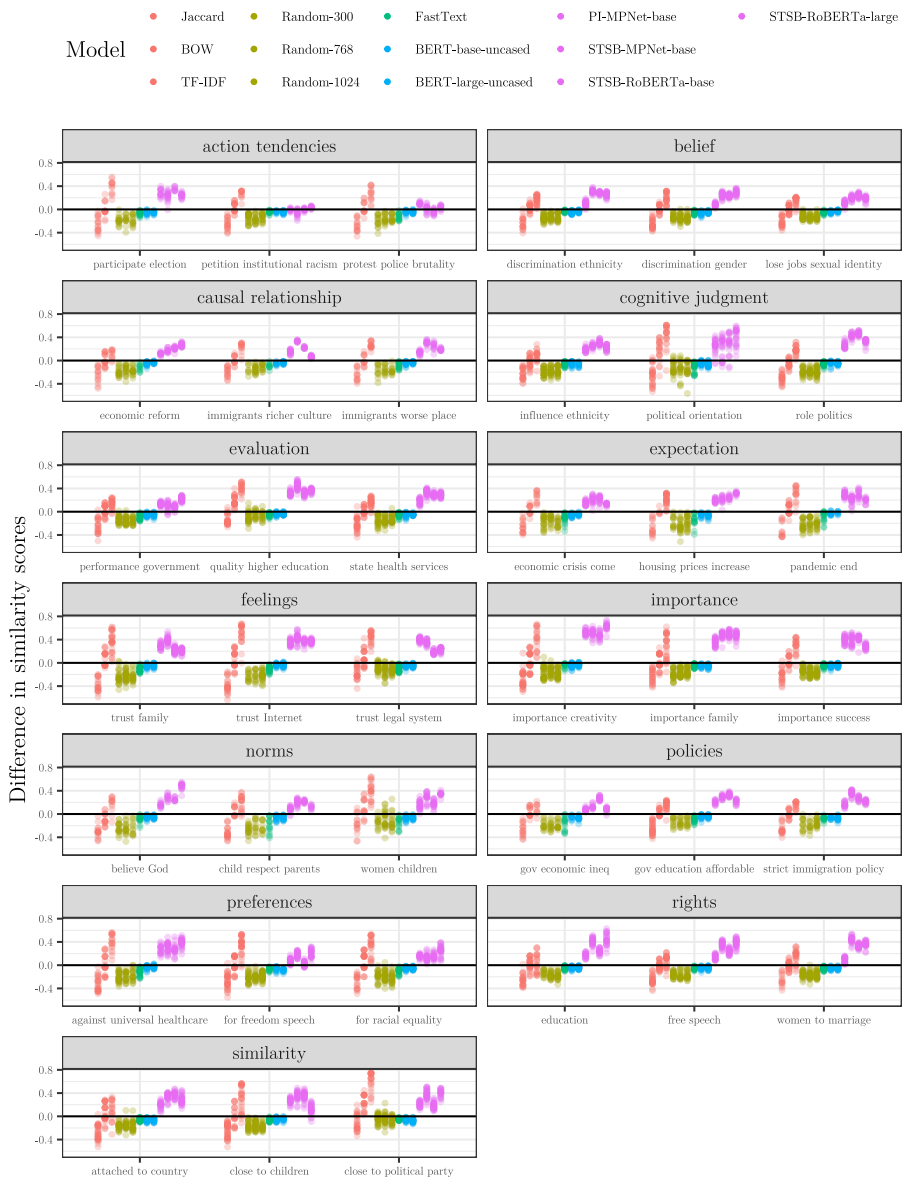

Figure 3 Hypothesis 2. Figure legend text.

Additional file 2 — Sample additional file title
Additional file descriptions text.