Predicting Public Opinions through the Integration of Large Language Models and Nationally Representative Surveys

Keywords: public opinion, social survey, large language model, artificial intelligence, machine learning

Extended Abstract

Social scientists have explored how the advancement of large language models could improve and innovate research practices in quantitative and qualitative studies, but their integration with social survey data, such as simulating surveys and predicting public opinions, has yet to be thoroughly investigated (Evans 2022). Recent studies show that machine learning techniques, such as matrix factorization, can fill in missing responses in the existing survey data better than traditional imputation models (Sengupta et al. 2023). However, this approach has a clear limitation: it cannot predict responses to new questions that were not asked before. The emerging literature has attempted to show that large language models trained on a large corpus of text data can understand the context and meaning of words and phrases, which enables researchers to measure "average" humans' opinions (Kozlowski, Taddy, and Evans 2019). However, the validity and reliability of these large language models have been contested because they are known to produce biased opinions that are not representative to the population due to the lack of socio-demographic diversity in their training data.

In this paper, we investigate the feasibility of combining large language models, such as BERT, with the General Social Survey, a nationally representative survey that has been conducted regularly since 1972 in the United States to predict public opinions at an unprecedented scale. We develop a novel methodological framework that leverages open-source large language models to predict public opinions both at individual and population levels. Specifically, we incorporate three most important embeddings for predicting opinions – sentence embedding, belief embedding, and period embedding – into fine-turning the pre-trained large language model building on recent works (e.g., jury model) (Gordon et al. 2022) (Figure 1). We fine-tune our BERT-GSS model to predict 64,818 Americans' responses across 2,950 survey questions in the General Social Surveys (GSS) from 1972 to 2018 that has collected a wealth of data on various topics such as political attitudes, cultural activities, social values, and demographic characteristics. Due to the varying number of response options in the survey questions, we binarized response options into "agree" or "disagree". Our model architecture is implemented using Huggingface's TensorFlow API for BERT (i.e., RoBERTa).

We employed two methods to evaluate the performance of our model using pooled cross-sectional and panel data of GSS. First, we assessed the model's ability to recover responses to missing survey questions in the pooled cross-sectional data. To do this, we randomly selected 10% of the questions for each survey year and removed the entire responses to these questions from the training data. We then used our BERT-GSS model to predict the missing responses, and we evaluated the accuracy of the predictions. Second, we evaluated the model's predictive capability for future opinions of individuals in the panel data using three-wave panel datasets, where each sample of individuals was surveyed three times for four years (i.e., 2006-2008-2010, 2008-2010-2012, 2010-2012-2014). In doing so, we use the initial round of the panel data to predict the responses in the subsequent round. We evaluated the model's capability to predict both known and unknown questions.

To assess the model's performance in predicting individual opinions in the pooled cross-sectional data, we trained the model using 90% of the survey questions and tested its

predictive accuracy using the remaining 10%. Figure 2 presents the outcomes, demonstrating that the model accurately predicted individual-level responses, achieving an AUC of 84.10%, accuracy of 76.34%, and F1-score of 74.06%. significantly outperforming state-of-the-art methods in missing response imputation (i.e., matrix factorization). Additionally, we show that our model can predict the same individuals' opinions on new questions in the future, achieving an AUC of 84.66% in two years and 80.17% in four years in the panel data. We further show that our model can predict about two-thirds of survey responses with a 3% margin of error and predict past and future counterfactual trends even for survey questions asked only once based on the near-perfect correlation (r=0.98) between predicted and observed responses in the entire 1974-2018 GSS data (See Figure 3).

Figure 4 shows two different cases that demonstrate a potential use of our predictive models. Panel A shows how our model prediction can be useful to fill in missing responses for questions that were not asked all the time. We can successfully fill the missing responses for the proportion of Americans who think a wife should help her husbands' career than to have oneself since 2000 when those questions were not asked in the survey. In addition, our predictive models can be useful for predicting trends of an opinion even when it is asked only once. Panel B shows that in 1994, about 42% Americans think that a woman should get an abortion for any reason. If our model that could successfully predict the aggregate response in 1994 will be able to fill in the missing responses in other years, support for women's abortion for any reason would have been continuously increasing since 1993.

Although our results show that fine-tuned language models have the potential to be used for predictive purposes and provide new insights into human language and societal attitudes, they do not perform consistently across different individuals and across different variables. First, our model shows that socio-demographic characteristics are significantly associated with predictive accuracy; we observe higher accuracy for those who are highly educated, middle-aged and old, Whites and have higher income. Second, our models do not predict every opinion equally well. For example, our models make a better prediction on opinions that are answered by many individuals, which highlights the importance of having large training data that can inform the model about the meaning of opinions. We also find that aggregate-level prediction errors are higher for questions that are not ideologically sorted and that lack consensus. Overall, our paper demonstrates potentials as well as limitations of our novel model framework that integrates large language models with nationally representative surveys to predict public opinions.

References

Evans, James. 2022. "From Text Signals to Simulations: A Review and Complement to Text as Data by Grimmer, Roberts & Stewart (PUP 2022)." *Sociological Methods & Research* 51(4):1868–85.

Gordon, Mitchell L., Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. "Jury Learning: Integrating Dissenting Voices into Machine Learning Models." Pp. 1–19, *CHI '22*. New York, NY, USA: Association for Computing Machinery.

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

Sengupta, Nandana, Madeleine Udell, Nathan Srebro, and James Evans. 2023. "Sparse Data Reconstruction, Missing Value and Multiple Imputation through Matrix Factorization." *Sociological Methodology* 53(1):72–114. doi: 10.1177/00811750221125799.

Figure 1. Model architecture

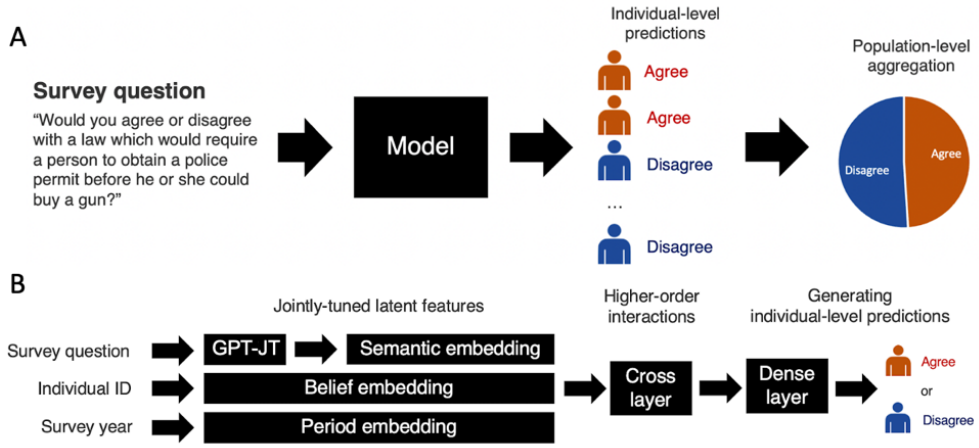
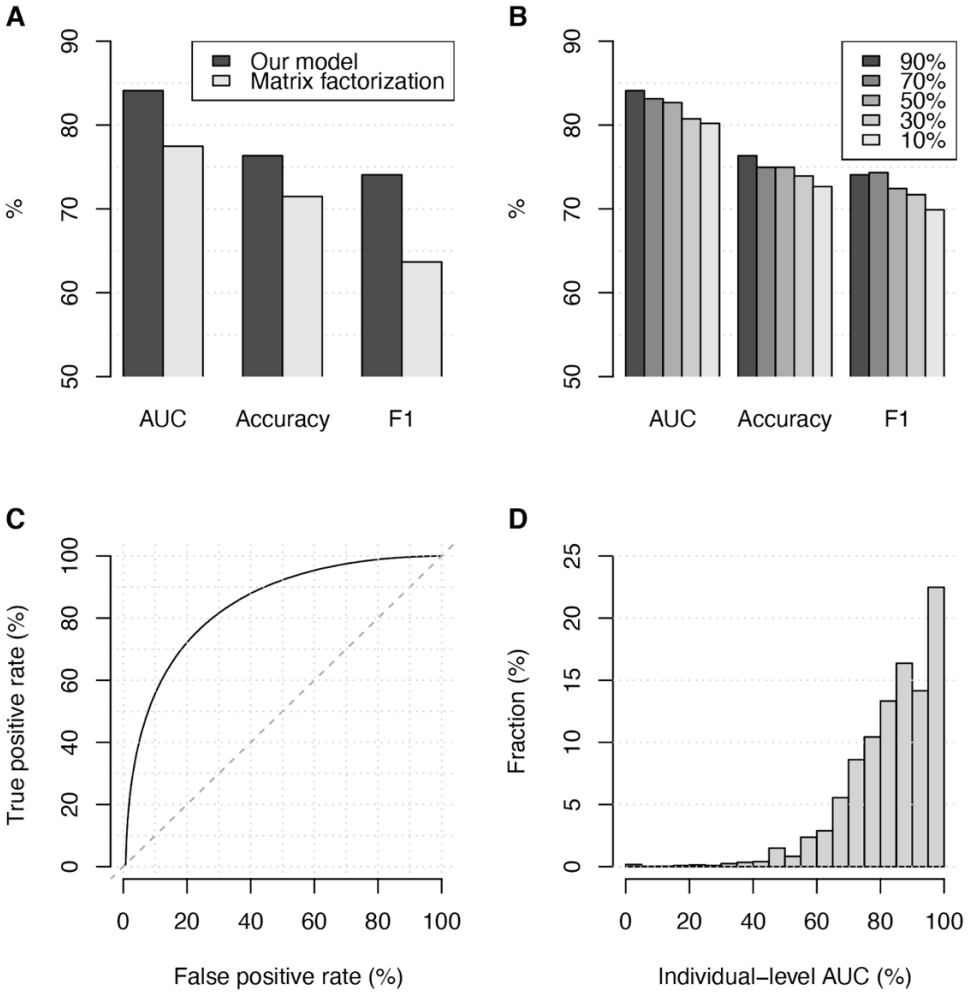
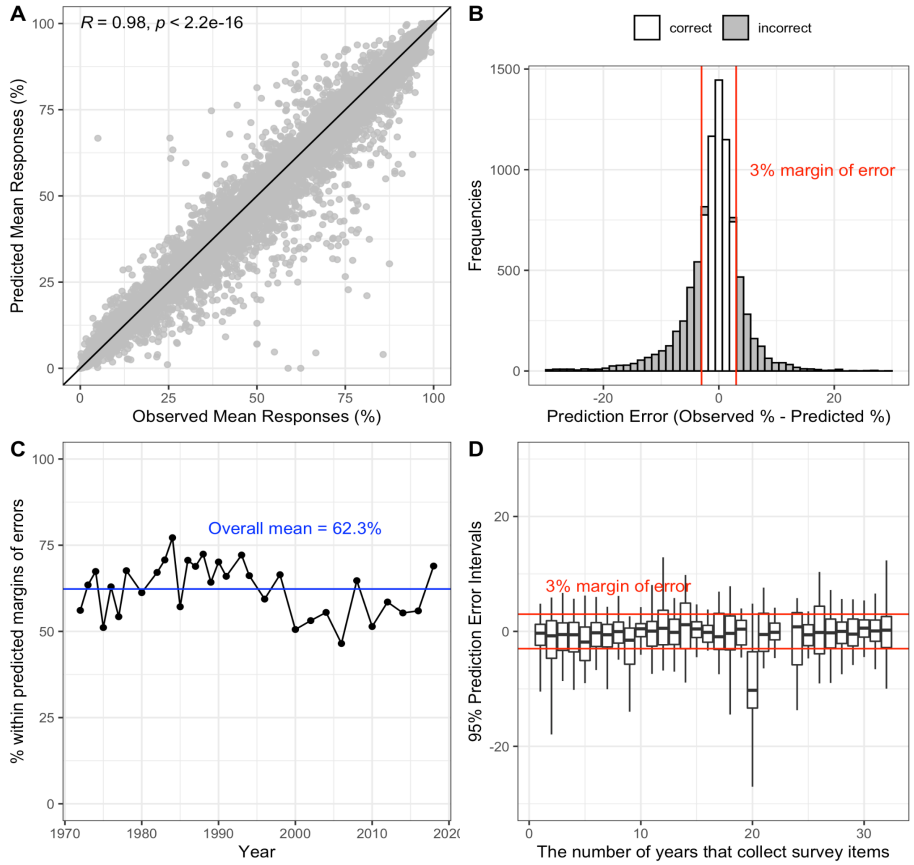


Figure 2. Prediction of opinions at the level of individuals



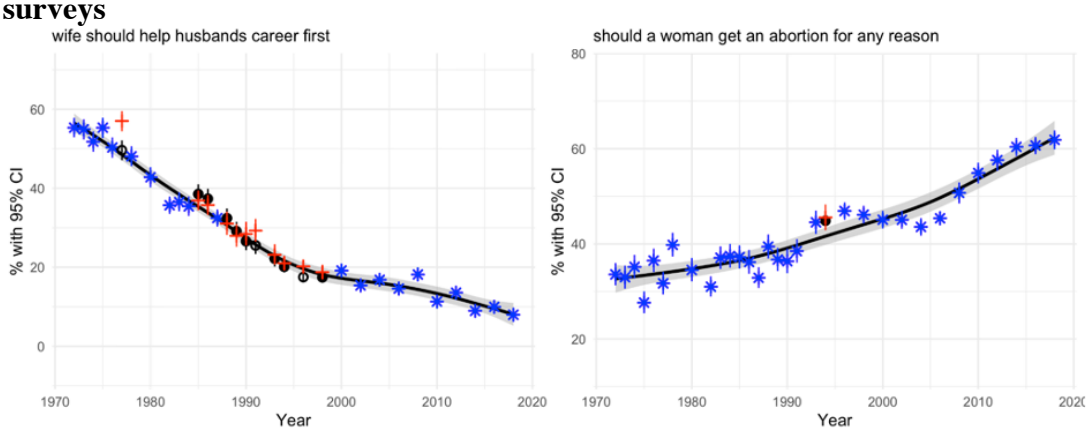
Notes: *Panel A shows the model performance of our model and matrix factorization model. Our model shows the* AUC of 84.10%, Accuracy of 76.34%, and F1-score of 74.06%, whereas matrix factorization shows the AUC of 77.47%, Accuracy of 71.48%, and F1-score of 63.68%. Panel B shows the model performance by the sparsity of training data (i.e., *how many questions are kept in the training data*). Panel C shows the *receiver operating characteristic (ROC) curve showing the model's ability to predict individuals' opinions*. Panel D shows the distribution of prediction performance (AUC) across individuals.

Figure 3. Prediction of aggregated opinions at the level of population



Note: Panel A show the relationship between the original mean and the predicted mean. Panel B shows the distribution of the prediction error. Panel C shows the percentage of survey items that were within the predicted margin of errors (+/- 3%) across survey years. Panel D shows the distribution of prediction errors across the number of survey years.

Figure 4. Illustration of potential application of the LLM for predicting trends in the GSS surveys



Notes. The generalized additive model was used to estimate the counter-factual trends. We define correct prediction when the 95% prediction intervals include the observed estimate. The variable name, response option, and wording of questions for each panel are followed: Panel A. "Now I'm going to read several more statements. As I read each one, please tell me whether you strongly agree, agree, disagree, or strongly disagree with it. For example, here is the statement: B. It is more important for a wife to help her husband's career than to have one herself. (fehelp), Panel B. "Do you agree or disagree: A pregnant woman should be able to obtain a legal abortion for any reason whatsoever, if she chooses not to have the baby." (abchoose)