Beyond the Ballot: Progress Report

Furkan Kadıoğlu

17 Jun 2024

1 Background

1.1 Survey Imputation

Survey imputation is a crucial process in statistical analysis, used to address missing or incomplete data in survey responses. Missing data can introduce bias, reduce the statistical power of analyses, and compromise the validity of research findings [10, 15]. Imputation techniques help provide a complete dataset, enabling more accurate statistical inferences and robust conclusions.

Listwise and pairwise deletion were among the earliest methods used. Listwise deletion involves excluding any cases (respondents) with missing data from the analysis, which can lead to significant data loss and reduced sample size. This reduction can result in biased results because the remaining data may not be representative of the original population. Pairwise deletion involves using only the available data pairs for analysis, which retains more data but can lead to inconsistent results and complications in statistical analysis due to varying sample sizes across different analyses [8].

Simple imputation methods include replacing missing values with the mean, median, or mode of the observed data. For instance, if a survey respondent did not answer a question about their age, the missing value might be replaced with the average age of all other respondents. While simple and easy to use, these methods do not account for the variability in the data. They can underestimate the variance and potentially lead to biased estimates, as they do not consider the underlying distribution or relationships between variables. For example, imputing the mean for a skewed variable would distort the distribution and relationships [10].

Hot deck imputation was developed as a method to replace missing values with observed responses from similar respondents. This technique involves finding a respondent (or several respondents) with similar characteristics to the one with missing data and using their observed value to fill in the missing value. For example, if the income data for a respondent is missing, the method might use the income of another respondent with a similar age, education level, and job type. This method improves upon simple imputation by considering respondent similarity, but it can still introduce bias if the matching process is not well-executed, as it assumes that the chosen respondents are truly representative of the missing cases [1].

Cold deck imputation involves using external sources or prior data for imputation. For example, if a survey is conducted annually, missing values in the current year's survey might be filled in using data from the previous year's survey. This method relies on the availability and relevance of external data, which can sometimes be a limitation if the external data is not comparable to the survey data. Additionally, it assumes that the previous data is accurate and that the relationship between the variables has remained constant over time [10].

Regression imputation uses regression models to predict missing values based on other observed variables. For example, a regression model could be developed to predict income based on variables such as education, age, and job type. This model is then used to estimate the missing income values. This method provides more nuanced imputation than simple mean substitution but assumes that the regression model is correctly specified. If the model is misspecified, it can lead to biased estimates.

Moreover, it tends to reduce the variability in the imputed data because the imputed values are predicted by a deterministic function [8].

The Expectation-Maximization (EM) algorithm, introduced for iteratively estimating parameters in the presence of missing data, refines estimates to improve imputation accuracy. The EM algorithm involves two steps: the Expectation step (E-step), where missing data are estimated based on observed data, and the Maximization step (M-step), where the estimated values are used to update the model parameters. This process is repeated until convergence. Despite its power, the EM algorithm can be computationally intensive and sensitive to initial values. It also assumes that the data are missing at random (MAR) [3].

Multiple imputation (MI) involves creating multiple datasets with different imputed values, analyzing each dataset separately, and then combining the results. This method accounts for the uncertainty inherent in the imputation process and provides more robust statistical inferences. In MI, the missing values are imputed multiple times to create several complete datasets. Each dataset is analyzed separately, and the results are combined to produce estimates and confidence intervals that reflect the uncertainty due to missing data. MI requires careful implementation and computational resources, particularly with large datasets, but it provides a robust framework for dealing with missing data [9].

K-Nearest Neighbors (KNN) imputation imputes missing values based on the average of the nearest neighbors' values, considering the similarity of other variables in the data. For example, if a respondent's income is missing, KNN imputation might find the k respondents with the most similar characteristics (age, education, job type) and use their average income to impute the missing value. This method is computationally feasible with modern technology and handles complex data structures, but it can be sensitive to the choice of the number of neighbors (k) and the distance metric used. Moreover, KNN can struggle with high-dimensional data where distance metrics become less meaningful [12].

Machine learning methods, including decision trees, random forests, and neural networks, have been applied to imputation. These methods can handle large, complex datasets and capture non-linear relationships between variables. For example, a random forest model, which consists of many decision trees, can be trained to predict missing values based on observed data. These methods require significant computational power and expertise in machine learning but offer flexibility and accuracy in handling various types of missing data. However, they can be overfitting and complex to interpret [16].

Advanced statistical techniques like Multiple Imputation by Chained Equations (MICE) and Bayesian methods offer sophisticated ways to handle complex, multivariate missing data patterns. MICE, for instance, imputes missing values by iteratively filling in each variable with missing data using a series of regression models, creating a chained sequence of models. Bayesian methods provide a probabilistic framework for imputation, allowing for the incorporation of prior knowledge and the estimation of uncertainty in a coherent manner. These methods provide flexible and powerful tools for imputation but require careful model specification and can be computationally demanding [17].

Modern imputation methods increasingly leverage AI and deep learning techniques to handle large and complex datasets, providing more accurate and reliable imputed values. For example, denoising autoencoders, a type of deep learning model, can be used to learn a representation of the data that can be used to predict and impute missing values. These methods, while powerful, require extensive computational resources and expertise in AI and deep learning [5].

The development of specialized software and tools, such as the R packages 'mice' and 'Amelia', and Python libraries like 'fancyimpute', has made sophisticated imputation methods more accessible to researchers and practitioners. These tools facilitate the implementation of advanced imputation techniques but still require users to understand the underlying methods to apply them correctly [17, 6].

1.2 Large Language Models

The development of large language models (LLMs) represents a significant milestone in the field of artificial intelligence (AI) and natural language processing (NLP). The journey to modern LLMs has been marked by several key advancements in machine learning, computational power, and linguistic theory.

The earliest attempts at natural language processing can be traced back to the 1950s and 1960s, with efforts to create rule-based systems that could understand and generate human language. The most notable project from this era was the Georgetown-IBM experiment in 1954, which involved a rudimentary machine translation of Russian to English [7]. However, the complexity of human language posed significant challenges to these early systems.

The advent of machine learning in the 1980s introduced statistical methods for NLP, enabling the analysis of large corpora of text. One of the pioneering works was the development of hidden Markov models (HMMs) for speech recognition and part-of-speech tagging. These models utilized probabilities to predict sequences of words, marking a departure from purely rule-based systems [13].

In the 2000s, the introduction of deep learning revolutionized NLP. Neural networks, particularly recurrent neural networks (RNNs) and long short-term memory (LSTM) networks, demonstrated the ability to capture long-range dependencies in text, which was crucial for tasks such as language modeling and machine translation. The development of word embeddings, such as Word2Vec, allowed for the representation of words in continuous vector spaces, preserving semantic relationships [11].

The transformer architecture, introduced by Vaswani et al. in 2017, marked a watershed moment in the development of LLMs. Transformers abstained from the sequential processing of RNNs in favor of a self-attention mechanism, which allowed for greater parallelization and the modeling of long-range dependencies more effectively [18]. This architecture underpinned models such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer).

BERT, introduced by Devlin et al. in 2018, leveraged bidirectional context, improving performance on a variety of NLP tasks through a two-stage process of pre-training and fine-tuning [4]. GPT, initially introduced by Radford et al. in 2018 and significantly scaled up in subsequent versions, demonstrated the power of unsupervised pre-training on large text corpora followed by supervised fine-tuning [14].

The latest iterations of GPT, particularly GPT-3 and beyond, exemplify the capabilities of LLMs with billions of parameters. These models are capable of generating human-like text, performing complex language tasks, and even exhibiting basic reasoning abilities [2]. The success of these models is largely attributed to the scale of training data, advancements in computational resources, and sophisticated training techniques.

Overall, the development of LLMs has been driven by a confluence of advancements in statistical methods, neural network architectures, computational power, and the availability of large-scale datasets. These factors have collectively enabled the creation of models that can understand and generate human language with unprecedented proficiency.

2 Method

Rationale for Survey Imputation with LLMs

There are many advanced methods for survey imputations. However, we aim to explore whether Large Language Models (LLMs) can provide more insights into voters' thought processes while imputing their survey responses. If an LLM can predict party affiliations based on given statements, it might reveal aspects of human reasoning that other survey imputation methods cannot, especially since traditional methods do not inquire about the rationale behind choices.

Prompt Engineering

We can consider various potential directions such as fine-tuning, instruction tuning, representation engineering, and prompt engineering. As a basic initial step, we focus on prompt engineering to maximize the potential of our model by ensuring the quality of prompts.

Methodology

Our study uses survey responses from the European Social Survey (ESS) related to political interests and socio-political orientations to predict party affiliations within the same survey. We employ two models: the first model processes survey responses for input variables and generates *Statements*, which are normal human-like speeches rather than discrete survey numbers. The second model uses these statements to predict party affiliations by analyzing what someone who made those statements might think about the target survey variable. We evaluate the performance by employing accuracy metrics, using ESS data as the gold standard.

Model Choice

For our research, LLaMA 2 is the optimal choice due to its advanced features and suitability for handling complex natural language tasks. The ESS data encompasses a wide range of social variables, including political interest, socio-political orientations, and party allegiance, which are crucial for our study on predicting political affiliations. LLaMA 2's transformer architecture and fine-tuning with reinforcement learning from human feedback (RLHF) enable it to generate coherent, contextually appropriate text from open-ended survey responses. This aligns well with our need to convert survey responses into natural statements and then predict political orientations based on these statements. Furthermore, LLaMA 2's flexibility in fine-tuning on various platforms enhances its adaptability to our specific research requirements, ensuring accurate and nuanced predictions. This approach leverages the comprehensive nature of the ESS data and the advanced capabilities of LLaMA 2, as discussed in recent literature on state-of-the-art language models.

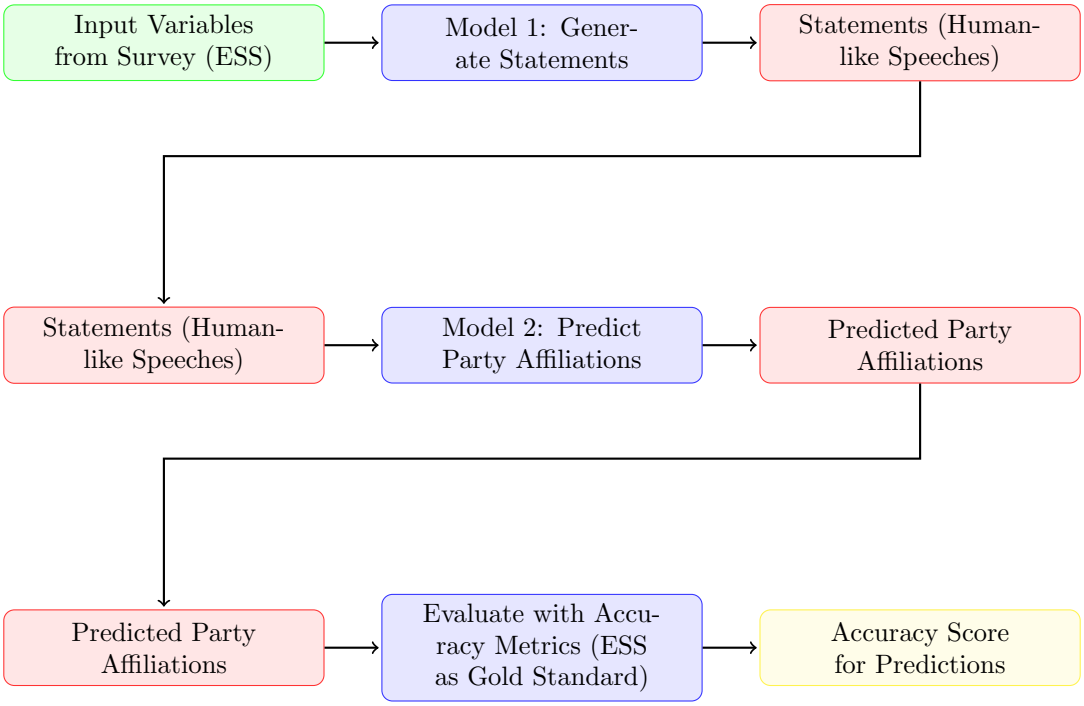

Figure 1: The diagram illustrates the research methodology for survey imputation using Large Language Models (LLMs). Input variables from the ESS survey are transformed into human-like statements by Model 1. These statements are then used by Model 2 to predict party affiliations. The predicted affiliations are evaluated against the gold standard using accuracy metrics, with the final accuracy score for the predicted party affiliations being calculated.

References

[1] Rebecca R Andridge and Roderick JA Little. A review of hot deck imputation for survey non-response. *International statistical review*, 78(1):40–64, 2010.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[3] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[5] Lovedeep Gondara and Ke Wang. Mida: Multiple imputation using denoising autoencoders. In *Advances in Knowledge Discovery and Data Mining: 22nd Pacific-Asia Conference, PAKDD 2018, Melbourne, VIC, Australia, June 3-6, 2018, Proceedings, Part III 22*, pages 260–272. Springer, 2018.

[6] James Honaker, Gary King, and Matthew Blackwell. Amelia ii: A program for missing data. *Journal of Statistical Software*, 45:1–47, 2011.

[7] W John Hutchins. The georgetown-ibm experiment demonstrated in january 1954. In *Conference of the Association for Machine Translation in the Americas*, pages 102–114. Springer, 2004.

[8] Hyun Kang. The prevention and handling of the missing data. *Korean journal of anesthesiology*, 64(5):402, 2013.

[9] Peng Li, Elizabeth A Stuart, and David B Allison. Multiple imputation: a flexible tool for handling missing data. *Jama*, 314(18):1966–1967, 2015.

[10] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

[11] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[12] Utomo Pujianto, Aji Prasetya Wibawa, Muhammad Iqbal Akbar, et al. K-nearest neighbor (k-nn) based missing data imputation. In *2019 5th International Conference on Science in Information Technology (ICSITech)*, pages 83–88. IEEE, 2019.

[13] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[14] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018.

[15] Joseph L. Schafer. Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8:15 – 3, 1999.

[16] Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

[17] Stef van Buuren and Karin G. M. Groothuis-Oudshoorn. Mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45:1–67, 2011.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.