Beyond the Ballot: Final Report

Furkan Kadıoğlu

2 October 2024

Introduction

Large language models (LLMs) have made remarkable strides in recent years, particularly with the introduction of transformer-based architectures, which enabled the scaling of models from millions to trillions of parameters. This increase in size, along with the use of self-supervised learning techniques, has greatly enhanced their ability to perform a wide range of tasks, such as text generation, translation, and reasoning across multiple domains [18, 16, 13]. The shift from task-specific to general-purpose models, as seen with GPT-4 and PaLM, has allowed for versatility and fine-tuning that can adapt to specialized tasks, leading to unprecedented performance across benchmarks and novel use cases [9]. These advancements, while impressive, bring challenges such as scalability, computational demand, and ethical concerns, particularly regarding bias and resource efficiency [9].

With these advancement of large language models (LLMs) and their increasing ability to comprehend natural language, there has been a growing interest in leveraging unstructured data from the internet to extract public opinion [6, 4, 28, 7]. In this context, social media provides an accessible platform, enabling large-scale studies [24, 21]. Despite the increasing use of social media data (SMD) for public opinion analysis, traditional surveys maintain their status as one of the most reliable methods for measuring public sentiment. Research shows that while SMD offers timely insights, it is often limited by the non-representative nature of social media users [20] and the platform's possible tendency to amplify extreme or polarized opinions [22, 17]. Additionally, the complexity of preprocessing and analyzing social media data introduces challenges in drawing clear, valid conclusions from these data sources. For example, constructing meaningful measures from SMD can be both time-consuming and susceptible to biases, as highlighted by studies in the literature [19]. In contrast, surveys provide a more structured and controlled method, ensuring diverse demographic representation, which makes them a more reliable tool for capturing the general population's attitudes. However, they come with their own limitations, such as respondents' reluctance to answer certain questions. Additionally, conducting a comprehensive survey may require posing a large number of questions, which is often impractical. As the number of questions increases, the participation rate tends to decrease [8], consequently undermining the generalizability of the results. Moreover, even if these challenges are addressed, the costs of conducting such surveys can become prohibitively high.

The challenges associated with the high costs of surveys and the difficulties in conducting reliable analyses with social media data may lead to the idea of using LLMs to replace survey participants [23]. However, the potential biases inherent in LLMs and the ongoing debate about their ability to accurately reflect societal diversity suggest that we have not yet reached a point where this solution is feasible [14, 25, 2].

For the reasons mentioned above, an alternative approach has emerged: using LLMs not to replace human participants or rely solely on raw social media data, but to enhance existing surveys [12]. According to this idea, LLMs could be used to predict participants' responses to unanswered questions, questions that were asked to some participants but not others, or even entirely unasked questions. [12, Kim and Lee (2023)] have reported relative success in predictions for the first two types of questions, while predictions for unasked questions have not achieved the same level of accuracy. Furthermore, for prediction and evaluation, converting survey responses from scales like the Likert scale into binary categories, treating them as a two-class classification problem, has limitations. Let's consider two different response distributions. For simplicity, let's have four options that represent levels of agreement from 1 to 4. We have 100 participants. If we examine the participants' distribution as 25-25-25-25 versus 50-0-0-50 from the perspective of a binary classification problem, there is no difference. However, as can be seen, the first distribution exhibits a more uniform distribution, while the second distribution indicates a more polarized society. If we binarize the problem, we cannot see the nuances between these two distributions.

Predicting missing values has been investigated over the years using survey imputation methods. However, current machine learning-based survey imputation techniques often struggle with sparse data [12]. Additionally, merely relying on the statistical correlation between responses overlooks the semantic relationships between the questions themselves. Current deep learning-based imputation methods, while powerful, face significant limitations in terms of computational complexity and scalability. These models require substantial resources, making them impractical for large-scale or real-time applications. Performance inconsistencies arise across datasets with varying levels of missingness, as no single method consistently performs well in all scenarios. Moreover, many approaches focus solely on imputation accuracy, often neglecting the impact of imputed data on downstream tasks, which remains a critical area for improvement [26].

In this study, we aim to understand on whether leveraging semantic relationships among survey questions can help predict answers to unasked questions. To predict responses to unasked questions, we generate an embedding that captures the participants' opinions. This approach not only aims to predict their likely answers but also proposes participants' vector representations that we can use to measure the similarity between participants, enabling a range of computational social science applications. For example, in social network analysis, we can use vector representations to model the formation and dissolution of social ties by calculating the similarity of individuals based on their interaction patterns or shared attributes. This allows researchers to analyze how closely connected individuals are, predict future relationships, and identify influential nodes within a network [27]. In the context of social movements and online activism, these vectors can be used to identify groups with shared interests or ideologies by comparing their behavior or communication data, which helps track how movements grow, evolve, and influence public discourse [1]. In polarization and echo chambers, vector representations can quantify how ideologically similar individuals are, offering insights into how polarized clusters form and interact over time. This allows researchers to model the evolution of polarized opinions and design interventions to mitigate echo chambers [3]. Lastly, in group dynamics and behavioral prediction, vectors enable the comparison of group members based on behavioral attributes, helping to predict how certain group compositions impact outcomes like collaboration success or conflict within teams. By modeling individual similarities, researchers can assess how different personality traits or skill sets affect group performance and outcomes [11].

Data and Methodology

Dataset

The dataset used in this study comes from the European Social Survey (ESS) Round 11, which includes standard political variables collected across 13 European countries with a total of 22,190 participants. The chosen variables focus on political engagement, attitudes, and trust in political institutions. Examples of the variables include political interest (polintr), trust in parliament (trstprl), and placement on the left-right political scale (lrscale). These variables were selected to explore how individuals' political attitudes and trust levels differ across countries and to examine correlations between political participation and socio-political beliefs. For representative data collection, the ESS data follows standard practices for survey sampling, including the use of inclusion probabilities, clustering, stratification, and the calculation of design weights and design effects [10, 15], making it a suitable choice for public opinion analysis.

Methodology

First, questions from the ESS dataset that could be answered using a specific scale were selected, and questions like "Which party did you vote for?" that could not be scaled were excluded. The analysis included questions grouped into three categories. The first category comprised questions that could be answered on a bipolar scale ranging from "strongly disagree" (extremely dissatisfied) to "strongly agree" (extremely satisfied), with responses normalized to a range between -1 and 1. The second category involved questions with a unipolar scale, such as the level of trust in politicians, and responses were normalized between 0 and 1. Lastly, yes/no questions, which do not have intermediate values, were normalized similarly to questions on a bipolar scale.

In the second phase, the questions and their corresponding responses were transformed into statements using a large language model (LLM), remaining as close to the original syntax as possible to prevent additional bias. For example, the question *How able do you think you are to take an active role in a group involved with political issues?* and its pivot response *Completely able* were converted into the statement, *I believe I am completely able to take an active role in a group involved with political issues.* Our goal at this stage was not to generate a statement for every possible question and answer pair, but rather to produce statements for the pivot responses (e.g., 5-*completely able* on a scale from 1 to 5). These pivot statements were used to produce the pivot vectors in the third stage. Additionally, we had three types of missing values in our dataset, and one statement was produced for each of these missing values. In total, for each question, we obtained four question-answer pairs for the *pivot response*, *Don't know*, *Refusal*, *No answer*. Then, four statements were generated to produce a vector for each pair.

In the third phase, we collected the semantic representations of the statements created in the previous phase. To achieve this, we utilized a sentence transformer [5, *BGE-M3*] to derive the mathematical representations of the statements, as it is one of the top performers in the Massive Text Embedding[1] and Thai Sentence Embedding[2] benchmarks as of October 2, 2024. Its open-source nature and availability in the Hugging Face library also facilitated straightforward implementation. Then, these vector representations are aggregated to create an individual belief embedding for each participant. The responses normalized in the first phase are used as coefficients for the vector representations of the pivot statements. If the participant refused to answer, selected "don't know," or didn't have an answer, the embedding of the corresponding statement was used. For example, in the case where the participant had options ranging from 1 (Not at all able) to 5 (Completely able), the pivot statement generated in the second phase from the pivot response *completely able* was used to obtain its pivot vector representation. If the participant selected option 3 (Quite able), the normalized coefficient of 0.5 is applied to the corresponding statement's vector representation. If the participant refused to answer, the vector representation of the statement generated for the refusal option is included with a coefficient of 1. Finally, the participant's individual belief embedding is obtained by summing these vectors with the corresponding coefficients.

Finally, after obtaining the vector representations that reflect each user's individual beliefs based on their stated opinions, we predicted the participants' answers that would be given to questions that were not included in their individual belief embeddings. Pivot statements were generated for these excluded questions, and their vector representations were obtained. To predict the answers to these excluded questions, the projection of the participants' belief vectors onto the target question's vector was calculated. The distribution derived from these projections was used to estimate the participants' responses to these omitted questions.

[1] https://huggingface.co/spaces/mteb/leaderboard
[2] https://huggingface.co/spaces/panuthept/thai_sentence_embedding_benchmark

Algorithm 1 Predicting Responses to Unasked Survey Questions Using Embedding Projections

Input: ESS dataset questions, Large Language Model (LLM)
Output: Belief embeddings, Predicted answers for omitted questions

procedure QuestionSelectionAndNormalization
  Select questions that can be answered on a bipolar/unipolar scale
  Exclude questions that cannot be scaled (e.g., "Which party did you vote for?")
  Group questions into three categories:

- **Bipolar scale questions**: Answers range from "Strongly Disagree" to "Strongly Agree"
- **Unipolar scale questions**: Examples include trust levels in politicians
- **Yes/No questions**: Questions with binary answers

  Normalize responses:

- Bipolar scale → [-1, 1]
- Unipolar scale → [0, 1]
- Yes/No questions: Normalize similar to bipolar scale

procedure StatementGeneration
  for each question and its response **do**
    Use LLM to transform question-response pairs into statements
      ▷ For example: *"How able do you think you are to take an active role*
      ▷ *in a group involved with political issues?"*
    if response is **"missing"** (e.g., "Refusal", "Don't know") **then**
      Generate a statement indicating that the participant did not answer
        ▷ For example: *"I prefer not to answer about my ability to take an*
        ▷ *active role in a group involved with political issues."*
    else if response is an **extreme value** (e.g., "Strongly Agree", "Completely able") **then**
      Generate a positive (pivot) statement
        ▷ For example: *"I believe I am completely able to take an active role*
        ▷ *in a group involved with political issues."*

procedure BeliefEmbeddingDerivation
  for each participant **do**
    Use LLM to generate semantic embeddings for the statements
    Apply normalized coefficients (answers from the first phase) to the embeddings
    Aggregate the weighted embeddings to create an individual belief embedding

procedure PredictUnanskedQuestions
  Exclude a subset of answered questions from embedding calculation
  Use LLM to generate pivot statements of the omitted questions
  Use LLM to generate embeddings of the pivot statements
  Calculate the projection of each participant's belief embedding onto the target question's vector
  Estimate responses based on the distribution of projections

References

[1] Falah Amro and Hemant Purohit. Integrated content-network analysis to discover influential collectives for studying social cyber-threats from online social movements. *Social Network Analysis and Mining*, 13(1):120, 2023.

[2] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

[3] Fabian Baumann, Philipp Lorenz-Spreen, Igor M Sokolov, and Michele Starnini. Modeling echo chambers and polarization dynamics in social networks. *Physical Review Letters*, 124(4):048301, 2020.

[4] Aaditya Bhatia. Advancing policy insights: Opinion data analysis and discourse structuring using llms. 2024.

[5] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 2024.

[6] Xuefan Dong and Ying Lian. A review of social media-based public opinion analyses: Challenges and recommendations. *Technology in Society*, 67:101724, 2021.

[7] Yanxia Dui and Hongchun Hu. Social media public opinion detection using multimodal natural language processing and attention mechanisms. *IET Information Security*, 2024(1):8880804, 2024.

[8] Mirta Galesic and Michael Bosnjak. Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public opinion quarterly*, 73(2):349–360, 2009.

[9] Desta Haileselassie Hagos, Rick Battle, and Danda B Rawat. Recent advances in generative ai and large language models: Current status, challenges, and perspectives. *IEEE Transactions on Artificial Intelligence*, 2024.

[10] Roger Jowell, Caroline Roberts, Rory Fitzgerald, and Gillian Eva. *Measuring attitudes cross-nationally: Lessons from the European Social Survey*. Sage, 2007.

[11] R Kelly Garrett. Protest in an information society: A review of literature on social movements and new icts. *Information, communication & society*, 9(02):202–224, 2006.

[12] Junsol Kim and Byungkyu Lee. Ai-augmented surveys: Leveraging large language models and surveys for opinion prediction. *arXiv preprint arXiv:2305.09620*, 2023.

[13] Pranjal Kumar. Large language models (llms): survey, technical frameworks, and future challenges. *Artificial Intelligence Review*, 57(9):260, 2024.

[14] Sanguk Lee, Tai-Quan Peng, Matthew H Goldberg, Seth A Rosenthal, John E Kotcher, Edward W Maibach, and Anthony Leiserowitz. Can large language models estimate public opinion about global warming? an empirical assessment of algorithmic fidelity and bias. *PLOS Climate*, 3(8):e0000429, 2024.

[15] Peter Lynn, Sabine Hader, Siegfried Gabler, and Seppo Laaksonen. Methods for achieving equivalence of samples in cross-national surveys: the european social survey experience. Technical report, ISER Working Paper Series, 2004.

[16] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023.

[17] Maria Nordbrandt. Affective polarization in the digital age: Testing the direction of the relationship between social media and users' feelings for out-group parties. *New media & society*, 25(12):3392–3411, 2023.

[18] Rajvardhan Patil and Venkat Gudivada. A review of current trends, techniques, and challenges in large language models (llms). *Applied Sciences*, 14(5):2074, 2024.

[19] Maud Reveilhac, Stephanie Steinmetz, and Davide Morselli. A systematic literature review of how and whether social media data can complement traditional survey data to study public opinion. *Multimedia tools and applications*, 81(7):10107–10142, 2022.

[20] Derek Ruths and Juergen Pfeffer. Social media for large studies of behaviour. *Science*, 346:1063–4, 11 2014.

[21] Stefan Stieglitz and Linh Dang-Xuan. Social media and political communication: a social media analytics framework. *Social Network Analysis and Mining*, 3:1277 – 1291, 2012.

[22] Jonathan Stray, Ravi Iyer, and Helena Puig Larrauri. The algorithmic management of polarization and violence on social media. 2023.

[23] Seungjong Sun, Eungu Lee, Dongyan Nan, Xiangying Zhao, Wonbyung Lee, Bernard J Jansen, and Jang Hyun Kim. Random silicon sampling: Simulating human sub-population opinion using a large language model based on group-level demographic information. *arXiv preprint arXiv:2402.18144*, 2024.

[24] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G. Sandner, and Isabell M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *Proceedings of the International AAAI Conference on Web and Social Media*, 2010.

[25] Angelina Wang, Jamie Morgenstern, and John P Dickerson. Large language models cannot replace human participants because they cannot portray identity groups. *arXiv preprint arXiv:2402.01908*, 2024.

[26] Jun Wang, Wenjie Du, Wei Cao, Keli Zhang, Wenjia Wang, Yuxuan Liang, and Qingsong Wen. Deep learning for multivariate time series imputation: A survey. *arXiv preprint arXiv:2402.04059*, 2024.

[27] Xinwei Xu. Studying social networks in the age of computational social science. *EPJ Data Science*, 12(1):61, 2023.

[28] Baoyu Zhang, Tao Chen, Xiao Wang, Qiang Li, Weishan Zhang, and Fei-Yue Wang. Decoding activist public opinion in decentralized self-organized protests using llm. *IEEE Transactions on Computational Social Systems*, 2024.