Beyond the Ballot: Final Report

Furkan Kadıoğlu

11 November 2024

Abstract

With the advent of ChatGPT, interdisciplinary studies on how large language models (LLMs) can be utilized have emerged. Questions such as whether survey participants could be replaced by LLMs, leveraging their human-like response generation, gained popularity. Alongside this, studies aimed at enhancing existing surveys using LLMs before considering participant replacement were also introduced. In this study, we explored whether it is possible to predict responses to questions not asked in a survey using the answers given to asked questions. To achieve this, we employed LLM and embeddings, another component of the transformer architecture that underpins LLM technologies. Discussions about LLMs being biased and functioning as black-box systems led us to integrate LLMs as part of a pipeline where their input and output could be transparently observed, rather than solely relying on LLMs to solve the problem. All the results obtained from the experiments and the related codebase are shared as open-source.[1]

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 [29, 27, 21]. 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 [14]. These advancements, while impressive, bring challenges such as scalability, computational demand, and ethical concerns, particularly regarding bias and resource efficiency [14].

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 [10, 5, 46, 11]. In this context, social media provides an accessible platform, enabling large-scale studies [40, 37]. 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 [34] and the platform's possible tendency to amplify extreme or polarized opinions [38, 28]. 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 [33]. 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 [12], consequently undermining the generalizability of

[1] https://github.com/furkan-kadioglu/Beyond-the-Ballot

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 [39]. 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 [22, 43, 3].

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 [20]. 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. [20, 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 [20]. 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 [44].

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 [45]. 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 [4]. 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 [19].

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 [17, 25], 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 [7, *BGE-M3*] to derive the mathematical representations of the statements, as it is one of the top performers in the Massive Text Embedding[2] and Thai Sentence Embedding[3] 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

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

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.

Implementation

We selected the questions related to politics from the European Social Survey data. After filtering out the questions with categorical answer options, we were left with 41 questions. After manually reviewing these 41 questions with their answer options, we divided them into two groups: unipolar and bipolar. If the responses to a question indicated a degree of opinion, it was classified as unipolar, whereas questions that presented a spectrum between two opposing views were classified as bipolar. For instance, a question aimed at understanding trust in politicians was classified as unipolar, while questions like 'Should one be ashamed if a close family member is homosexual?'—where respondents could indicate agreement or disagreement with two opposing views—were classified as bipolar. Since the metadata of the survey questions was in HTML format, the BeautifulSoup library was used during preprocessing. In this phase, by parsing the metadata file, we extracted information about the countries included in the survey, the questions, and possible response options. These were then labeled as unipolar or bipolar according to the definitions above.

In the second phase, the data was preprocessed. We processed the ESS Round 11 data, which we received in CSV format, using Pandas. At this stage, a unique ID was created for each participant by combining country and idno, as idno was unique within each country but could be the same for different participants in other countries. Next, we established a standard convention for handling missing values. For some questions, responses could range from 1 to 10, with missing values indicated by 77, 88, and 99, while for other questions, missing values were indicated by 7, 8, and 9. For overall consistency, 7 was replaced with 77, 8 with 88, and 9 with 99. In the second step of data processing, normalization was performed based on whether the questions were unipolar or bipolar. For unipolar questions, the highest possible answer was rescaled to 1 and the lowest to 0. For bipolar questions, answers were rescaled between -1 and 1. To illustrate this rescaling, consider a unipolar question where participants indicate their level of trust in politicians on a scale from 0 to 10. If a response was 5, it would be normalized to 0.5; if this question had been bipolar, a value of 5 would correspond to 0. We performed this normalization to later determine the weights that each vector obtained from sentence transformers would hold in calculating participants' individual vectors for each question.

In the third phase, prompts to be provided to the large language model were prepared. We selected the most extreme response to each question as the pivot answer. We believed that choosing an extreme response would help the generated statement convey a clear stance, making it more informative. For example, for the question 'Do you trust politicians?' the pivot response was 'Complete Trust (10).' We used the ChatGPT-4o model to turn the questions and pivot responses into statements as if they were made by participants. Before deciding on ChatGPT-4o, we also considered other open-source language models; however, due to the inference time required for large models, the inconsistency in answers from smaller models, and the cost-effectiveness of the ChatGPT-4 model, we decided to use ChatGPT-4o. Additionally, we minimized bias and ensured that the questions and pivot responses were syntactically converted into statements. To maintain objectivity and consistency in the statements, we set the temperature parameter to 0. In addition to the pivot responses, we also created statements for each missing value. As a result, using ChatGPT-4o, we generated four statements for each question in the survey: pivot and missing values (refusal, don't know, no answer).

In the fourth phase, we use the BGE-M3 sentence transformer to obtain vectors for the statements generated in the third phase. This model allows us to map our sentences into mathematical space, where semantically similar sentences are positioned closer together, while semantically distant ones are farther apart. As a result of this phase, we obtain vector representations of the statements created using the pivot responses. These vectors will be used in the next phase to calculate the vectors that will represent the participants.

In the fifth phase, we calculate the vectors intended to represent participants' individual beliefs. To do this, we aggregated the vectors obtained in the previous phase by taking into account the participants' responses. During data preprocessing, participants' responses were normalized based on their position within the spectrum of possible answers. These normalized values were used as coefficients for the vectors representing each question, and the vectors were then summed to create a vector representing each participant. In this phase, we assumed that participants' responses reflect who they are and convey meaning regarding their overall value systems. For participants with missing values, vectors obtained from statements generated for missing values were included with a coefficient of 1.

We exclude the questions for which we want to predict participants' answers when calculating their individual belief vectors. This allows us to have a 'gold annotation' for the answers we aim to predict. The target questions we want to predict, like other questions, have their pivot statements and pivot vectors calculated, but they are not included in the calculation of vectors representing participants. Ultimately, we have a mathematical space containing both participants and target questions. To make predictions within this space, we project participants' vectors onto the vectors of the target questions. These projections create a distribution for each target question. In obtaining this distribution, we assume that the lengths of the projections relate to their proximity to the pivot responses of the target questions.

After obtaining a projection distribution for each target question, we trim the extreme ends of the distribution on both sides to identify participants at the extremes. We then predict the responses for these participants with the extreme values on the answer spectrum. For the remaining majority of the population, we divide the distribution into intervals based on the number of possible responses. We then predict participants' responses according to the interval in which they fall. For example, suppose that after excluding the extreme values, the lengths of the projection vectors range between 0.5 and 1, and our target question allows for values from 1 to 5. In this case, we divide the range from 0.5 to 1 into five intervals, predicting values as follows: those in the 0.5–0.6 interval are predicted as 1, those in the 0.6–0.7 interval as 2, and so forth, with values exceeding 1 being assigned to 5, representing the extreme.

One of the challenges we faced when predicting participants' responses and testing the accuracy of our predictions was that each question had a different response range and number of possible answers. Additionally, the distribution of participants' responses varied from question to question. The metric we chose needed to show how much better our predictions were than random guessing and take into account that our questions were on a Likert scale. To eliminate the random chance factor, we used Cohen's Kappa metric. Since our predictions were on a scale, there was a meaningful difference between predicting a 3 or 4 for a response that should be a 2. To account for this difference, we also used Spearman's correlation metric. We implemented these metrics using the sklearn library.

Results and Discussion

Through our observations of the ESS data, we found that we could categorize questions into three different types. Accordingly, we selected one variable from each type as a target variable: unipolar (e.g., trstprl), yes/no (e.g., badge), and bipolar (e.g., lrscale). For the unipolar and bipolar variables, we chose those with the maximum number of response options to see how nuanced our observations were. Additionally, we made predictions for every variable in the dataset. While doing this, we excluded the target variable from the individual belief embedding calculation and used the remaining variables to obtain belief embeddings, which were then used for predicting the target variable for participants.

Detailed results for each variable can be found in the appendix section.

The Mean Absolute metric yielded different values for the three different variables we evaluated. For the unipolar and bipolar variables, we observed similar values ranging between 0 and 5. When we investigated the reasons for such a wide range in this score, we found that the yes/no questions produced much lower values. We observed that this was because the limited number of options made it less likely to produce a significant absolute error. Following this, we decided to use Cohen's Kappa metric, which takes into account the randomness factor.

Cohen's Kappa is a metric aimed at eliminating the randomness factor, measuring the level of agreement between two annotators on a scale from 0 to 1. A value of 0 indicates almost no agreement. In our experiments with different variables from the three types, the Cohen's Kappa score yielded values very close to 0. This result did not change even when we selected different variables. Consequently, we removed participants with any missing values, reducing the size of our dataset. Our goal was to determine whether our method of including missing values in the individual belief embedding made a difference. However, since this also gave similar results, we concluded that the issue did not lie there. We then considered whether the low score could be due to a precision error. For example, when using Cohen's Kappa to evaluate a variable with a Likert scale ranging from 1 to 10, predicting a value of 3 instead of 2 or predicting a value of 9 was penalized equally. To better reflect the nuance between these two different cases in our assessment, we decided to use Spearman's correlation metric.

Spearman's correlation metric, similar to Cohen's Kappa, returns a result between 0 and 1, where a value close to 0 indicates almost no correlation. In our experiments using this metric, instead of seeing consistent results across all variables like with Cohen's Kappa, we observed values ranging from -0.2 to 0.47. Additionally, when we compared the average values obtained for unipolar variables with those for bipolar variables, we found that our proposed method performed relatively better for unipolar variables. We believe this difference in performance is due to how bipolar variables were included in the calculation of individual belief embeddings. When a participant's response to a bipolar variable was the opposite extreme rather than the pivot response, the vector representing that variable was included by multiplying by -1. The effectiveness of this method relied on the assumption that a statement with an opposite opinion would be represented in our semantic space by a vector of the same magnitude but in the opposite direction. The difference in performance between unipolar and bipolar variables suggests that this assumption may not hold true. We believe that achieving this could require fine-tuning the sentence transformer so that two opposing statements are represented as exact opposites under vector addition.

Additionally, when we examined the predicted values, we observed that a certain value was returned more frequently than others, and this was not specific to any variable; whether the variable was unipolar or bipolar did not affect this outcome. Upon analyzing the distribution of the predicted values, we found that a distribution similar to a normal distribution repeated across all variables. While the extent to which this distribution was concentrated or spread out varied between variables, they all resembled the same distribution. However, when we examined the distribution of actual responses given by participants, we could observe more diverse distributions among the options. Due to this type of distribution, our predictions tended to yield relatively high Spearman correlation scores when they clustered around the most popular response in the survey, but when they failed to capture this, the scores approached 0. We believe that this distribution explains why the Spearman correlation data had high variance across variables.

In conclusion, it is evident that the extent to which the vectors representing participants in the generated space truly represent them remains open to further investigation. The distribution created using the length of the projection onto the target variable was a naive approach for predicting responses to the target question. Additionally, we believe that fine-tuning the sentence transformer used to create this space is necessary to close the performance gap between bipolar and unipolar variables.

Limitations

Although having a dataset like the ESS provided us with gold annotation for evaluating our predictions at the start of the project, several challenges emerged. Since we had to manually determine whether the questions were unipolar or bipolar, we were limited to conducting experiments only with variables related to politics. It was not enough to decide whether questions were unipolar or bipolar; we also had to exclude questions with nominal response values, such as 'Which party did you vote for?' Additionally, we encountered some difficulties when using open-source language models with the help of the Hugging Face library. While large language models took a long time for inference, smaller models produced inconsistent and irrelevant answers. The variability in the number of response options also made it difficult to approach the problem as a type of classification task by simply adding a standard dense layer when generating predictions.

Appendix

Results

		Mean Absolute Error	Cohen Kappa Score	Spearmanr
Type	Variable			
Unipolar	**polintr**	0.883295	0.026453	0.238750
	psppsgva	2.412656	-0.011598	0.246687
	actrolga	2.801398	-0.001775	0.094148
	psppipla	2.496940	-0.012412	0.275624
	cptppola	2.711854	-0.000960	0.091823
	trstprl	4.487281	-0.007492	0.476543
	trstlgl	3.201438	0.001075	0.324105
	trstplc	2.329718	0.007078	0.322615
	trstplt	4.929352	-0.004681	0.396520
	trstprt	4.805217	-0.006087	0.412937
	trstep	4.081178	-0.011811	0.398581
	trstun	3.682062	-0.008092	0.326568
Bipolar	**contplt**	0.157835	0.001568	0.017867
	donprty	0.067026	0.032551	0.076904
	badge	0.061801	0.066551	0.134577
	sgnptit	0.233863	0.017502	0.086000
	pbldmna	0.059074	0.085699	0.163334
	bctprd	0.211588	0.001169	0.018373
	pstplonl	0.135863	0.033650	0.115589
	volunfp	0.212746	0.028483	0.084148
	clsprty	0.454186	0.001012	0.020542
	lrscale	4.972976	-0.001320	0.032543
	stflife	2.463014	-0.013502	0.177612
	stfeco	3.876578	-0.013553	0.332473
	stfgov	4.419828	-0.010696	0.324073
	stfdem	3.173221	-0.013464	0.433951
	stfedu	2.557322	0.000234	0.278435
	stfhlth	3.254164	-0.009570	0.342289
	gincdif	1.943013	0.001135	-0.043765
	freehms	2.154493	0.000020	0.128433
	hmsfmlsh	0.983447	-0.034093	-0.224392
	hmsacld	1.657760	0.002726	0.097818
	euftf	3.741205	-0.005271	0.205423
	lrnobed	1.856650	0.002430	-0.006947
	loylead	1.521010	-0.003442	-0.087905
	imsmetn	1.472540	-0.013509	0.099564
	imdfetn	0.958960	0.010774	0.169370
	impcntr	1.489998	-0.024520	-0.167590
	imbgeco	2.127999	0.021623	0.384305
	imueclt	2.064119	0.012778	0.409174
	imwbcnt	2.072170	0.002954	0.314608

Table 1: Summary of Results with Missing Values

		Mean Absolute Error	Cohen Kappa Score	Spearmanr
Type	Variable			
Unipolar	**polintr**	0.968477	0.000303	0.184554
	psppsgva	2.284846	-0.010712	0.211624
	actrolga	2.669605	-0.001578	0.042211
	psppipla	2.213117	-0.012358	0.208630
	cptppola	2.517598	-0.002369	0.053133
	trstprl	3.730582	-0.012526	0.472565
	trstlgl	2.994123	0.000621	0.315011
	trstplc	1.798704	0.032405	0.351102
	trstplt	3.758298	-0.021041	0.448366
	trstprt	4.217057	-0.016612	0.413550
	trstep	3.073800	-0.010579	0.395895
	trstun	3.093836	-0.014292	0.384136
Bipolar	**contplt**	0.178922	0.000346	0.005706
	donprty	0.079276	0.025499	0.071086
	badge	0.071128	0.038593	0.095371
	sgnptit	0.265010	0.008007	0.054114
	pbldmna	0.068056	0.039456	0.101254
	bctprd	0.247579	0.000272	0.006760
	pstplonl	0.155012	0.018342	0.083775
	volunfp	0.238362	0.013696	0.056084
	clsprty	0.500634	0.000532	0.016317
	lrscale	4.889868	-0.001741	0.041140
	stflife	1.784813	0.002401	0.146555
	stfeco	3.191678	-0.014983	0.297870
	stfgov	3.905697	-0.012772	0.271658
	stfdem	2.481600	0.003359	0.382774
	stfedu	2.157016	0.012808	0.265109
	stfhlth	3.185734	-0.010900	0.347053
	gincdif	1.899152	0.000413	-0.014047
	freehms	2.207106	-0.001243	0.133389
	hmsfmlsh	0.957991	-0.033302	-0.225402
	hmsacld	1.710279	0.001587	0.073138
	euftf	3.672010	-0.004350	0.202250
	lrnobed	1.825419	0.002235	-0.034651
	loylead	1.766981	-0.009832	-0.107330
	imsmetn	1.656315	-0.015923	0.073245
	imdfetn	1.332131	-0.028476	0.149849
	impcntr	1.574234	-0.021680	-0.192748
	imbgeco	2.031523	0.025904	0.374692
	imueclt	2.005944	0.017193	0.414707
	imwbcnt	1.982368	0.005425	0.317069

Table 2: Summary of Results without Missing Values

	Mean Absolute Error	Cohen Kappa Score	Spearmanr
Type			
Bipolar	1.736360	0.006204	0.135062
Unipolar	3.235199	-0.002525	0.300408

Table 3: Means with Missing Values

	Mean Absolute Error	Cohen Kappa Score	Spearmanr
Type			
Bipolar	1.655925	0.002099	0.117475
Unipolar	2.776670	-0.005728	0.290065

Table 4: Means without Missing Values

Literature Reviews

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 [24, 35]. 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 [18].

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 [24].

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 [2].

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 [24].

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 [18].

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) [8].

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 [23].

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 [30].

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 [36].

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 [41].

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 [13].

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 [41, 15].

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 [16]. 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 [31].

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 [26].

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 [42]. 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 [9]. 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 [32].

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 [6]. 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.

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] 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.

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

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

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

[19] 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.

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

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

[22] 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.

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

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

[25] 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.

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

[27] 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.

[28] 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.

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

[30] 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.

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

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

[33] 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.

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

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

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

[37] 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.

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

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

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

[46] 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.