Assessing the Predictive Power of Social Media Data-Fed Large Language Models on Voter Behavior

Ehsan Barkhordar
ebarkhordar23@ku.edu.tr
Koç University
Istanbul, Turkey

Şükrü Atsizelti
satsizelti22@ku.edu.tr
Koç University
Istanbul, Turkey

ABSTRACT

This article explores how large language models (LLMs) can reflect human preferences and exhibit biases based on the diversity and type of input data. Utilizing survey data linked with tweets, we compare the predictive performance and bias manifestations of LLMs under three different data inclusion strategies: (1) using only demographic information, (2) combining demographic information with tweets, and (3) exclusively using tweets. The study finds that prompts enriched with tweets notably improve the predictive accuracy of models compared to those relying solely on demographic data. More importantly, the inclusion of dynamic, user-generated content like tweets not only reduces the oversimplification of individual identities but also lessens inherent biases, leading to more accurate and representative simulations of voter behavior. These findings underscore the critical role of data variety in LLM-based simulations, suggesting that integrating richer, real-time data sources can effectively diminish biases and enhance the models' ability to simulate complex human characteristics.

CCS CONCEPTS

• **Applied computing** → **Sociology**; • **Computing methodologies** → **Simulation evaluation**.

KEYWORDS

large language models, social media, voter behavior, predictive analytics

ACM Reference Format:
Ehsan Barkhordar and Şükrü Atsizelti. 2024. Assessing the Predictive Power of Social Media Data-Fed Large Language Models on Voter Behavior. In *Proceedings of 16th ACM Web Science Conference 2024 (WebSci '24).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3630744.3659831

1 INTRODUCTION

This study aims to assess the predictive accuracy and bias of Large Language Models (LLMs) when fed with social media data in forecasting voter behavior. It has been argued that biases in the training data of large language models reflect the real tendencies in society [3]; simulation studies have been conducted based on the capacity of large language models to reflect society [1]. In this field, existing surveys or experiments are replicated using large language models, and their compatibility with existing survey and experiment results is examined [1, 2].

Discussions on whether large language models can replace actual survey and experiment participants have also brought up objections regarding the capacities of LLMs. For example, Ollion et al. [4] have made warnings revolving around the replicability, privacy, and language bias issues, cautioning researchers about using it for research without considering those problems. One significant problem of these studies is the representativeness of the created silicon participants. In their meaningfully named article "large language models cannot replace human participants because they cannot portray identity groups", Wang et al. [5] argue that these LLM-based studies suffer from misportrayal, group flattening, and identity essentialization. They attribute these problems to the training data, where the group-related information generally comes from out-group remarks not from in-group and the loss function that 'rewards the most likely output' [5].

This disposition to essentialize identities may lead to a nonproductive line of research where identities are linked with unchanging or fixed characteristics, behaviors, or ideas. Adding self-produced information to the simulation process instead of just demographic information may enrich the results and solve some of the problems that arose due to the lack of information about the related simulated identities. Secondly, feeding the simulations with real-world information about the user may prevent or balance the essentializing tendency of large language models by providing unfixed stances and opinions. On top of that, adding user-generated information may be the cure to the problem of the source of group-related informations [5], that is since they are expressed by the user they may reflect the real tendencies of a population, instead of the prejudices towards them.

In this study, we utilized large language models (LLMs) to assess their predictive power on voter behavior by employing different types of data inputs. Specifically, we compared the following three data input strategies:

- **Demographic Data Only:** The model was prompted using basic demographic information such as age, education, ethnicity, gender, and location.
- **Demographic Data with Tweets:** The model received a combination of demographic information and a random selection of 30 tweets, including favorites and retweets, from the user's Twitter account.
- **Tweets Only:** The model's input consisted solely of random tweets from the user's Twitter account.

WebSci '24, May 21–24, 2024, Stuttgart, Germany

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/3630744.3659831

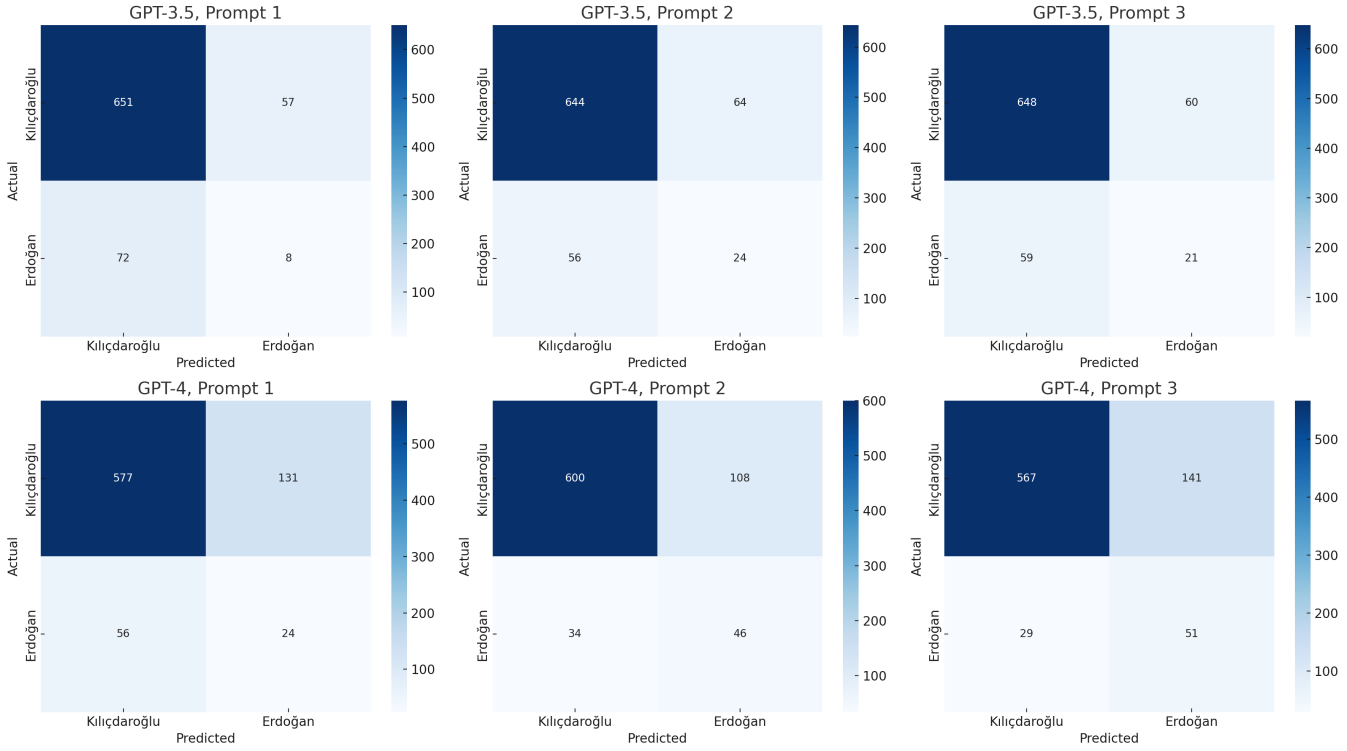

Figure 1: Confusion matrix of the different models and prompting strategies

The effectiveness of these scenarios was evaluated using precision, recall, and F1 scores for two versions of the models: GPT-3.5 [1] and GPT-4-turbo [2]. The results were then compared with the political preferences indicated by the users in a survey we conducted among Twitter users.

2 METHOD

This study is based on social media and survey data collected as part of the Politus project. The survey data was gathered between the first and second rounds of the 2023 Turkish presidential elections. The survey was conducted using Twitter ads and asked participants' consent to link their responses with their Twitter accounts, providing their usernames. Approximately 2,000 individuals completed the survey and half of these respondents' survey answers successfully linked with their Twitter accounts. Tweets from these linked profiles were then collected.

Prompts containing three different types of information were prepared and presented to two OpenAI models (GPT-3.5 and GPT-4-turbo). The first prompt included only demographic information (age, education, ethnicity, gender, and location) gathered from the survey. The second prompt included these demographic details plus a random selection of 30 tweets (including favorites and retweets) from the relevant user, while the third prompt contained only the random tweets. These prompts were used to predict which candidate the users likely voted for during the second round of the 2023 presidential elections. Responses indicating "I did not vote" in the survey were ignored in the analysis.

2.1 Dataset

The dataset comprises demographic data and linked Twitter data. Below are some examples from the demographic data:

Table 1: Examples from the demographic dataset

Username	Age	Education	Ethnicity	City
user1	59	university graduated	Turkish	Istanbul
user2	37	PhD holder	Turkish	Ankara
user3	21	high-school graduated	Turkish	Çanakkale
user4	28	master's degree	kurdish	Mardin
user5	62	university graduated	Arab	Adana

Table 2: Political Support Distribution

Candidate	Number of Supporters
Kemal Kılıçdaroğlu	708
Recep Tayyip Erdoğan	80

[1] gpt-3.5-turbo-0125
[2] gpt-4-turbo-2024-04-09

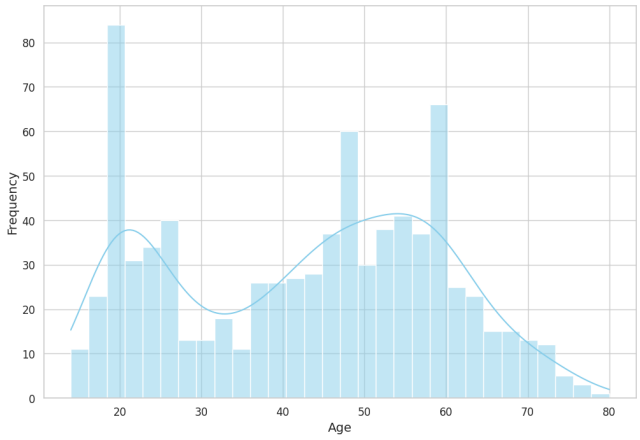

Figure 2: Age distribution of users in the dataset.

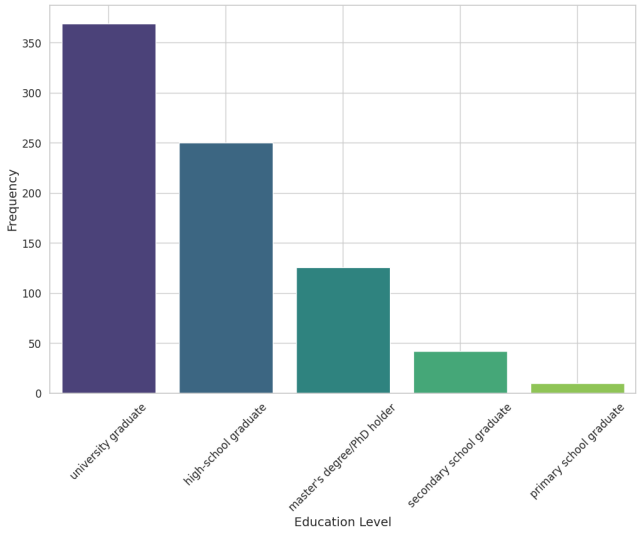

Figure 3: Education distribution of users in the dataset.

3 RESULTS AND DISCUSSIONS

The general results indicate that the GPT-4-turbo model performed better than the GPT-3.5 model across all prompt types. Additionally, prompts that included social media data were more successful than those containing only demographic information. The most successful strategy, excluding the recall score of GPT-4-turbo, was when demographic information and tweets were presented together. Presenting only tweets was found to be more advantageous than presenting only demographic information.

Due to a significant imbalance between the two candidates' supporters among survey respondents, these numbers might be misleading. It is crucial to consider the supporters of each candidate separately to better assess the biases of different prompts and models. Figure 5 shows the precision, recall, and F1 scores for both groups separately. The scores for Kılıçdaroğlu supporters do not vary significantly across models, showing high precision, recall,

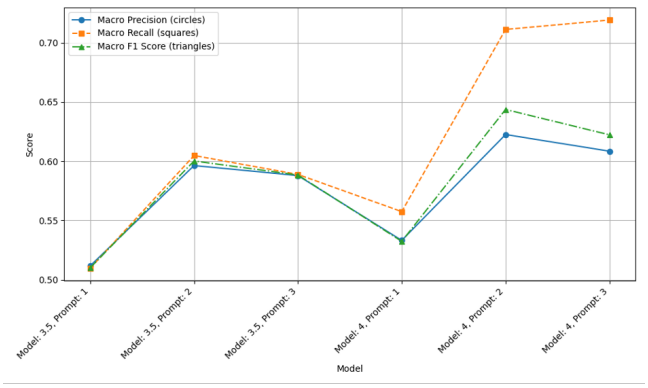

Figure 4: Performance scores of different models and prompting strategies.

and F1 scores. However, while Erdoğan supporters generally reflect the pattern observed in general scores, achieving the highest scores in the strategy combining demographic information with tweets (again, excluding the recall score of GPT-4-turbo), the scores for Kılıçdaroğlu supporters tend to decrease in the GPT-4-turbo model.

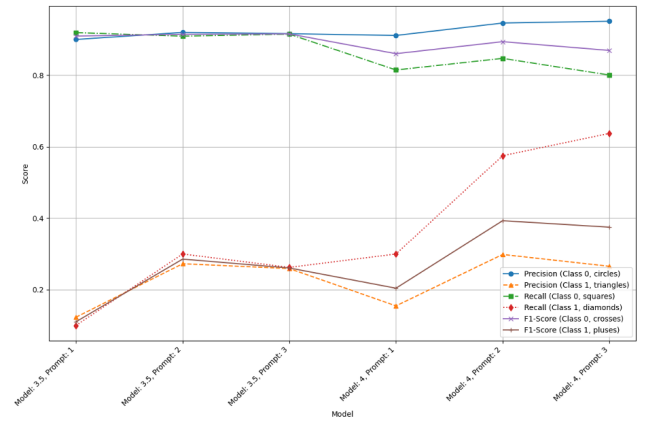

Figure 5: Performance scores for Kılıçdaroğlu and Erdoğan supporters.

To address the imbalance issue, we analyzed a balanced sample of 180 respondents (80 Erdoğan supporters and 80 Kılıçdaroğlu supporters). The results are shown in Figure 6. The balanced dataset analysis indicates that the overall trends remain consistent, with the GPT-4-turbo model outperforming across all prompting strategies. However, precision, recall, and F1 scores for both groups are more comparable, offering a clearer view of model performance without sample size imbalance.

For Erdoğan supporters, models often misclassify them. The inclusion of tweets has enhanced model performance across all metrics. One might expect better identification of Erdoğan supporters when only tweets are used, as demographic information can be misleading. However, the highest recall scores were achieved with the GPT-4-turbo model using tweets without demographic data, while GPT-3.5 performance declined in this scenario.

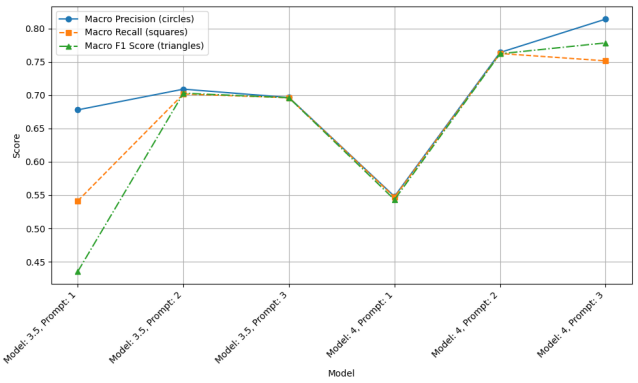

Figure 6: Performance scores of different models and prompting strategies with a balanced sample size (80 Erdoğan, 80 Kılıçdaroğlu).

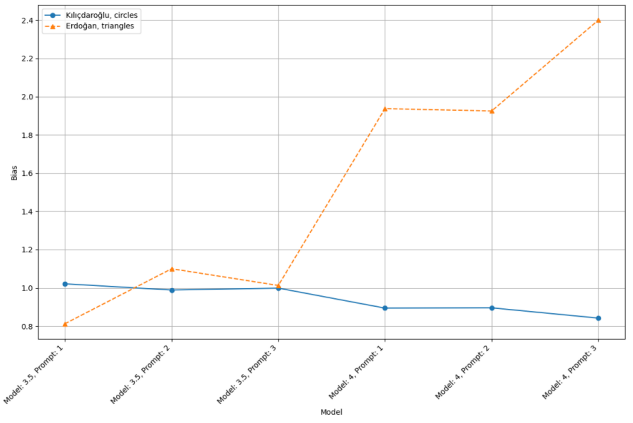

Figure 7: Class-specific bias comparison by model.

According to Figure 1, models mark actual Erdoğan supporters as Kılıçdaroğlu supporters less frequently as we move from GPT-3.5 to GPT-4-turbo, and from prompts without tweets to those with tweets. Additionally, the tendency of models to misclassify Kılıçdaroğlu supporters as Erdoğan supporters has similarly increased. Beyond the differences between models, the inclusion of tweets also seems to have led models to better identify and overpredict Erdoğan. Examining results on individual levels can provide important insights about biases in the models. Unfortunately, data insufficiency makes it impossible to examine each level individually.

4 CONCLUSION

This study provided a comparative analysis of Large Language Models (LLMs) in terms of their ability to predict voter behavior using different types of data inputs, specifically demographic information versus social media content. Our findings demonstrate that LLMs incorporating a mix of demographic and tweet data deliver the most accurate predictions. This superior performance suggests that the integration of real-time, self-generated user data can significantly enhance the model's understanding of complex human behaviors and preferences, reducing the occurrence of biases typically seen in models trained solely on demographic data.

Moreover, the results highlight the importance of model version updates, as GPT-4-turbo outperformed its predecessor in all scenarios. This underscores the continuous improvements in model architectures and training methodologies, contributing to more sophisticated data processing capabilities.

4.1 Limitations

This study has several limitations. The dataset, primarily consisting of Twitter users, may not be fully representative of the general population, introducing a potential bias toward more tech-savvy individuals. Additionally, the imbalance in the number of supporters for each candidate could have influenced the model's performance. The focus on Twitter data alone may not capture the full spectrum of an individual's online behavior and preferences.

4.2 Future Work

Future research should expand datasets to include more diverse demographic groups and social media interactions. Exploring other social media platforms like Facebook, Instagram, and LinkedIn, as well as incorporating multi-modal data sources such as images and videos, could provide further insights. Investigating different data preprocessing techniques and model fine-tuning strategies could optimize performance. Longitudinal studies tracking voter behavior over time with continuous data collection methods could offer deeper insights into voter preference dynamics.

REFERENCES

[1] William Agnew, A Stevie Bergman, Jennifer Chien, Mark Díaz, Seliem El-Sayed, Jaylen Pittman, Shakir Mohamed, and Kevin R McKee. 2024. The illusion of artificial inclusion. *arXiv preprint arXiv:2401.08572* (2024).

[2] Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*. PMLR, 337–371.

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

[4] Étienne Ollion, Rubing Shen, Ana Macanovic, and Arnault Chatelain. 2024. The dangers of using proprietary LLMs for research. *Nature Machine Intelligence* (2024), 1–2.

[5] Angelina Wang, Jamie Morgenstern, and John P Dickerson. 2024. Large language models cannot replace human participants because they cannot portray identity groups. *arXiv e-prints* (2024), arXiv–2402.