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 Atsızelti
satsizelti22@ku.edu.tr
Koç University
Istanbul, Turkey

ABSTRACT

This article investigates how large language models (LLMs) reflect human preferences and exhibit biases influenced by the diversity and nature of their input data. We used survey data related to Turkish presidential elections alongside tweets to assess the predictive performance and bias manifestations of LLMs under three different data inclusion strategies: (1) using only demographic information, (2) integrating demographic information with tweets, and (3) relying solely on tweets. Our findings reveal that prompts enriched with tweets typically achieve higher F1 Macro scores. However, this trend differs significantly when examining classes individually. While user-generated content significantly improves performance in predictions related to Recep Tayyip Erdoğan, it does not show the same effect for Kemal Kılıçdaroğlu. This study shows that different models and prompting styles result in varied biases for each candidate, leading to mixed outcomes. These results underscore the importance of exploring how biases vary across different scenarios, models, and prompting strategies for each case.

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 Atsızelti. 2024. Assessing the Predictive Power of Social Media Data-Fed Large Language Models on Voter Behavior. In *16th ACM Web Science Conference (Websci Companion '24), May 21–24, 2024, Stuttgart, Germany.* ACM, New York, NY, USA, 3 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 prompting scenarios:

- **Scenario 1:** Only demographic data — The model was prompted using basic demographic information such as age, education, ethnicity, gender, and location.
- **Scenario 2:** Demographic data and random 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.
- **Scenario 3:** Only random tweets of a Twitter user — The model's input consisted solely of random tweets from the user's Twitter account.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Websci Companion '24, May 21–24, 2024, Stuttgart, Germany
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0453-6/24/05
https://doi.org/10.1145/3630744.3659831

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

Model ID	Actual Kılıçdaroğlu		Actual Erdoğan	
	Predicted Kılıçdaroğlu	Predicted Erdoğan	Predicted Kılıçdaroğlu	Predicted Erdoğan
Model: 3.5, Prompt: 1	651	57	72	8
Model: 3.5, Prompt: 2	644	64	56	24
Model: 3.5, Prompt: 3	648	60	59	21
Model: 4, Prompt: 1	577	131	56	24
Model: 4, Prompt: 2	600	108	34	46
Model: 4, Prompt: 3	567	141	29	51

The effectiveness of these scenarios was evaluated based on their success scores using two versions of the ChatGPT model: 3.5 and 4. 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 different models (ChatGPT 3.5 and 4). 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.

3 RESULTS

The general results indicate that the ChatGPT 4 model performed better than the 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 scenario, excluding the recall score of ChatGPT 4, 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 (Erdoğan: 164, Kılıçdaroğlu: 666), 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. The Figure 2 shows the precision, recall, and F1 scores for both groups separately. Though

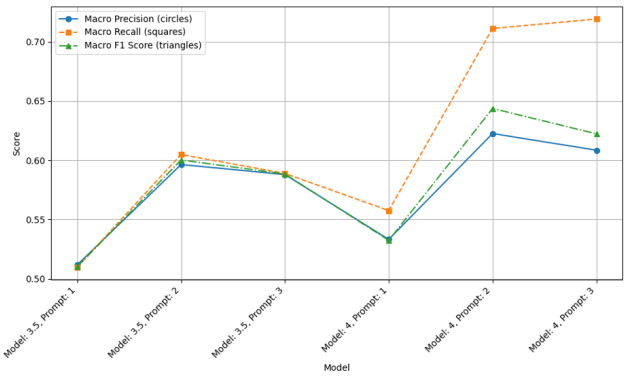


Figure 1: Performance scores of different models and prompting strategies.

ChatGPT 3.5 performs better compared to ChatGPT 4 in terms of recall and F1 scores for Kılıçdaroğlu predictions, both models show high precision, recall, and F1 scores compared to Erdoğan predictions. Erdoğan supporters generally reflect the pattern that observed in general scores, achieving the highest scores in the scenario combining demographic information with tweets (again, excluding the recall score of ChatGPT-4).

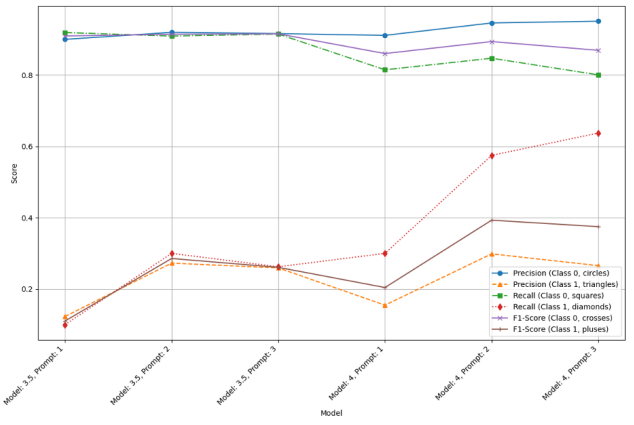


Figure 2: Performance scores for Kılıçdaroğlu and Erdoğan supporters.

When it comes to Erdoğan supporters, there is a notable tendency for models to misclassify them. Moreover, the inclusion of tweets has improved the performance of models in all aspects. One might expect that models would more successfully identify Erdoğan supporters in contexts where only tweets are presented compared to the prompts where both information are present due to the misleading nature of presenting demographic information. While the highest recall scores were achieved in the scenario using the ChatGPT 4 model and tweets without demographic information, the performance of the ChatGPT 3.5 model declined in this scenario.

Looking at recall, precision, and F1 scores alongside the ratio of users who assigned to a label to the users truly belong that label (i.e., (TP + FP)/(TP + FN)) can provide a broader perspective on bias in the models. Graph 3 shows these crude bias scores. Comparing outputs from models 3.5 and 4, there is a decrease in the tendency to label someone as Kılıçdaroğlu and an increase in labeling as Erdoğan. This graph should be interpreted with caution as the classes are quite unbalanced. Thinking along with confusion matrix can be more informative (see Table 1).

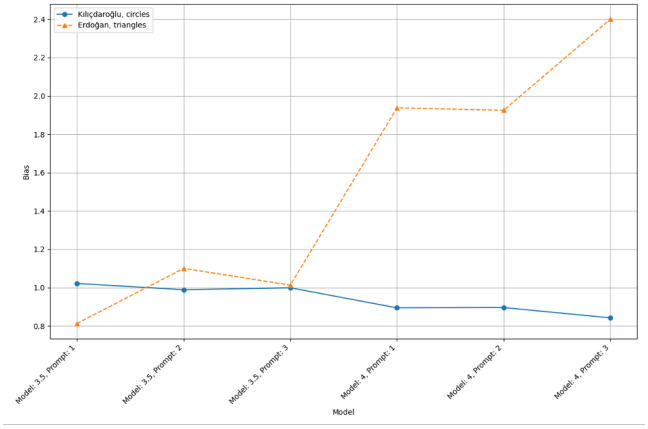


Figure 3: Class-specific bias comparison by model.

According to table 1, models mark actual Erdoğan supporters as Kılıçdaroğlu supporters less frequently as we move from model 3.5 to 4, 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 say important things about biases in the models. Unfortunately, data insufficiency makes it impossible to examine each level individually.

The findings offer mixed results regarding the three prompting techniques examined. Prompts involving tweets were found to be more successful in general. This study points out to the potential of adding tweet or tweet-like texts to survey participant simulations to counteract the misleading effects of providing solely demographic information. However, differing biases among models and prompting techniques between different cases should be considered. Changes in models can lead to different types of biases, so efforts to reduce bias in a particular model may not always be applicable to others. The results of this study should be read with caution as the data suffer from unbalanced classes. Future studies with a larger, more balanced sample and including demographic levels will provide significant insights.

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.