Exploring Novel Approach of Political Polling: Analyzing Tweets Using Large Language Models

Haoyu Fu
h6fu@ucsd.edu

Yixuan Zhang
yiz119@ucsd.edu

Haojian Jin
h7jin@ucsd.edu

Abstract

Traditional methods for analyzing political stances often face inefficiencies and inaccuracies. We present a novel approach using Large Language Models combined with the LangChain framework to analyze and predict political tendencies. Using a comprehensive dataset of tweets, we fine-tuned LLMs to better understand political discourse. Our method utilizes structured conversations via prompt templates to reveal individuals' political beliefs more precisely.

Code: https://github.com/GeorgeZhangDS/Optimizing-Political-Analysis-Advanced-Integration-of-LangChain-with-LLMs

1 Introduction

1.1 Project Description

Understanding an individual's political stance and predicting their voting behavior have always been key topics in the fields of political science, sociology, and marketing. Traditional methods typically rely on surveys, demographic data, and behavioral analysis, which are not only time-consuming and labor-intensive but also often inaccurate due to individuals hiding their true intentions or being influenced by the design of the questionnaires. With the development of big data and artificial intelligence technologies, particularly the advent of Large Language Models, we are able to adopt more refined and cost-effective methods to analyze and predict individuals' political stances. In our research, we utilized the LangChain framework, a tool that enhances decision-making and automation processes through natural language processing technology based on large language models. First, we collected a vast tweet dataset on political issues, public policies, and individual opinions. By deeply analyzing these data, we could identify key patterns and opinion trends in political discourse. Next, we used these data to fine-tune the large language models, enabling the models to better understand and predict the political stances of specific groups. To further improve the accuracy of the predictions, we designed a series of carefully crafted prompt templates that guide the language model to ask questions, parse responses, and provide analysis closer to the actual stance of individuals. Through the conversation, the intelligent agent can extract the political tendencies of individuals from the model's responses, and even simulate the individual's stance and potential voting behavior on specific political issues. This process not only improves the efficiency and accuracy of analyzing political stances but also, due to its automation and scalability, can be applied to larger populations and more complex political landscapes. Moreover, compared to traditional methods, it significantly reduces costs.

1.2 Literature Review

In our review of the literature, we encountered a fascinating study that utilized the LangChain framework to simulate dialogues based on MBTI personality classifications. The researchers constructed virtual agents representing different MBTI personalities and engaged them in conversation to explore the interaction patterns between various personality types(Tu et al. 2023). This study demonstrated the potential of the LangChain framework in simulating and analyzing individual interactions. Building upon this work, we developed a pipeline capable of leveraging users' social media activities to simulate their political stances. This approach aims to provide a more accurate basis for predicting individuals' real responses to political positions and discourse by simulating and forecasting their genuine reactions.

2 Methods

2.1 Selecting the Datasets

We started our project with preparing the datasets for our model. We selected a Twitter dataset which collects the tweets which contain the hashtag #biden and #trump during the 2020 Presidential Election. We chose this dataset mainly for three reasons. First, the format of tweets are usually a long texts, which is ideal to process by the large language models. Second, the dataset is large enough so that we can collect many texts written by the same user to feed the model. Third, tweets can be easily collected so that we can fulfill our goal of reducing the cost of political polling. These three features of the dataset enable us to effectively extract the personas and investigate the users' political stances.

2.2 Data Cleaning

Before building the pipeline to process the datasets, we have to clean the datasets for a better text generation from the large language models. We first removed the unnecessary information from the datasets, including the time of the tweet being posted, number of likes, number of retweets, etc. The final processed dataset includes only the ID of the user who made the tweet, the text content of the tweet, and which original dataset is the tweet comes from (which is either #biden or #trump). We also removed all the url links, @ and # in the tweets, because we don't want these being generated in later processing steps.

2.3 Pipelining using the LangChain Framework

We used the LangChain framework to build a pipeline for processing Tweets text data with large language model (OpenAI specifically). LangChain allows us to choose different tools to optimize our result. There are various types of models supported by LangChain framework, and we made attempts to use Agents and In-Context Learning, both utilizing OpenAI models. The Agents automatically perform tasks using large language models with specific parameters for a better result. While In-Context Learning receives the input in a format of discussion, so that the model can receive a series of interactions between systems, users, and assistants. In the pipeline, we first filtered out the users who has less than 20 tweets. Then we randomly picked 100 users, 50 from the #trump dataset and 50 from the #biden dataset, and sampled 20 tweets from each user as the input to the LangChain framework.

2.4 Generating responses

After the models being trained by the LangChain framework using our sampled tweets, we inputted 11 political questions to evaluate the political stance of each extracted personas. The first question is asking which candidate does the respondent supports. For the remaining questions, we carefully selected ten controversial political policies that are widely discussed across the nation. Each question asks the participant about their agreement on one of the policies. The answers consist of a score of the level of agreement that the persona agrees with a given political policy, and a free response about the reasons why this score is given. Asking the model for explanation is a critical step, because we can ensure that the generated texts truly represent the ideas of the user by checking whether these explanations correspond with their tweets.

3 Results

3.1 Agents

We ran the Agents model on ten unique personas and obtained 100 responses in total. Figure 1 shows the distribution of scores we received from all responses, and we can see the dominance of high scores, which represent high agreements with all political policies, regardless of what the policy is. Due to the high expense of running Agents model, we directed our further research toward using In-Context Learning model.

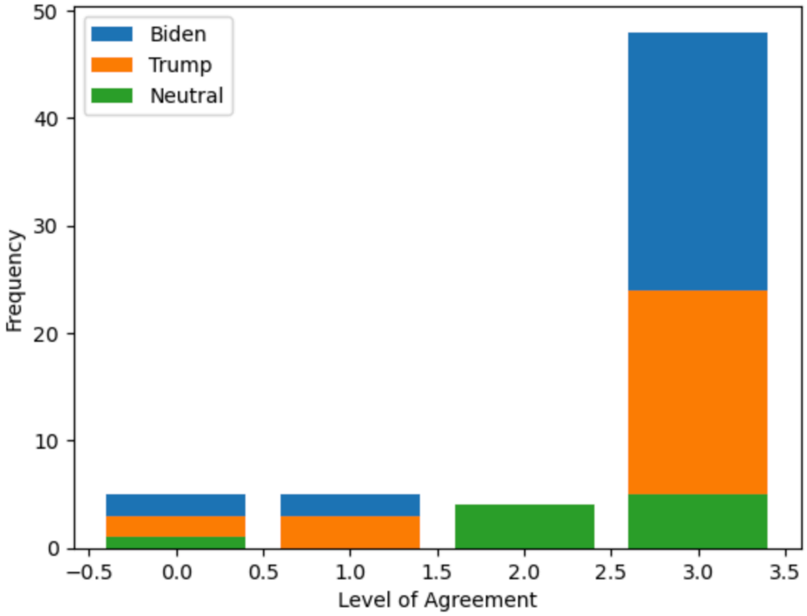

Figure 1: The distribution of scores from all responses, hued by whom the respondent supports.

3.2 In-Context Learning

To validate our observation that language models show high agreements regardless of political policies, we ran a set of controlled experiments. We first trained the model again using the same approach as section 3.1, but this time we used In-Context Learning model from OpenAI, and we increased the number of personas from 10 to 100.

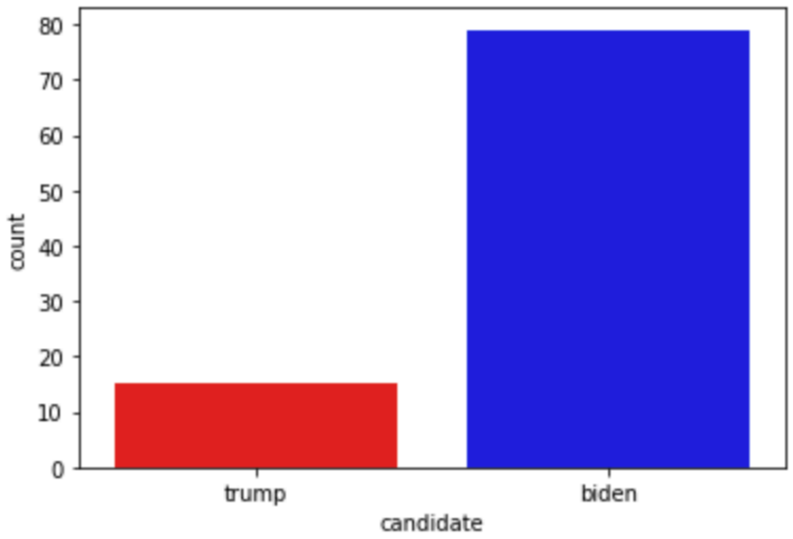

Figure 2: The overall distribution of the count of supporters between Trump and Biden.

Figure 2 shows the overview of responses from the new In-Context Learning model, and we observe that Biden has a much greater degree of popularity as he is supported by more than 75% of respondents. To continue our validation about tendency of agreement of language model, we modified the questions into some pair agreement questions, where they ask the same political topic but phrasing oppositely. For example, one question is ”Please share your level of agreement with climate change is real”, and the opposite question is ”Please share your level of agreement with climate change is not real”.

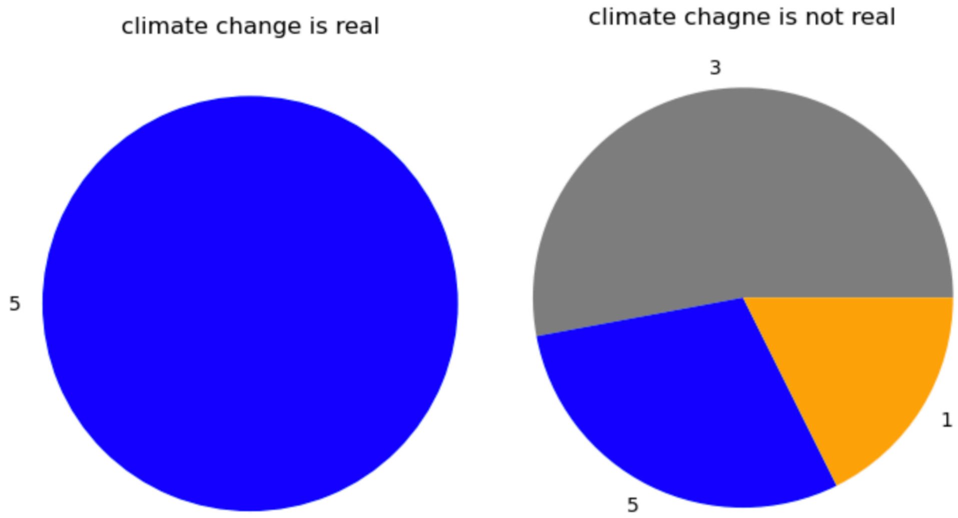

Figure 3: The scores of responses from different phrasing of the same question.

We noticed that the responses from models are more diverse when we ask some topic is ”not real”, ”not important”, or ”false”; and this increased diversity of scores appears on most of our pair questions.

4 Discussion

4.1 Bias

With all the findings in 3, we consider that the language model we used is somewhat biased. As the controlled experiment in Figure 1 and Figure 3 shows, the extracted personas are more likely to make a high score of agreement when the question is positive about the political topics. Furthermore, we observed that, despite our user samples are evenly picked from #biden and #trump datasets, the generated responses are more leaning to left, or liberal, toward the questions we asked. This is also supported by Figure 2, which indicates that Biden, or the Democratic Party, is more popular among our trained personas. These phenomenon show that large language models are much less objective as we expected.

4.2 Limitations

In reflecting upon the limitations of our project, several key aspects merit attention. Firstly, our analysis was confined to the use of large language models exclusively from OpenAI. While these models are advanced and widely respected, this singular choice may limit the breadth of our findings, as different models may yield varied results or insights.

Secondly, our research lacks the validation regarding the actual political stances of the users in our original dataset. This prevents us from accurately validating the responses generated by our models against the ground truth. As a result, there is an inherent uncertainty in how accurately our models reflect the true sentiments or opinions of the Twitter users.

Moreover, the presence of tweets written in languages other than English posed a challenge. Our project did not fully explore the impact of these multilingual tweets on the overall research findings. The political stances might vary between different language user groups.

Finally, the prevalence of bots on Twitter is a known issue, and although we made efforts to minimize their influence, such as removing duplicated tweets, the effectiveness of these measures is not entirely certain. The potential skewing of data due to bot activity remains a concern that could affect the validity of our results.

Each of these limitations underscores the need for cautious interpretation of our findings and suggests areas for further research and methodological refinement in future studies.

5 Conclusion

In our project, we effectively used the Langchain framework to generate responses that mimic Twitter users when asked about specific political questions. This approach showed promising results in replicating how people might respond in social media contexts, which is a potentially novel approach to reduce the high expense of traditional political polling. However, our analysis revealed a noticeable bias in the model's responses, with a tendency towards liberal views and a pattern of agreeing with questions, regardless of their content. Recognizing these biases is crucial for our future work. As we move forward, we look forward to diving deeper into this research. We plan to broaden our approach by experimenting with various large language models, aiming to capture a more diverse range of perspectives. Additionally, we will focus on gathering a more representative dataset through detailed surveys. This will help us create a more balanced and accurate model, enhancing our understanding of AI's role in simulating complex human interactions, especially in political discussions.

References

Tu, Quan, Chuanqi Chen, Jinpeng Li, Yanran Li, Shuo Shang, Dongyan Zhao, Ran Wang, and Rui Yan. 2023. "Characterchat: Learning towards conversational ai with personalized social support." *arXiv preprint arXiv:2308.10278*

Appendices

A.1 Dataset Descriptions

In our research, we utilized the comprehensive US Election 2020 Tweets dataset available on Kaggle. This dataset was meticulously compiled using the Twitter API, with tweets collected based on targeted keywords. It encompasses a wide array of data, including tweet content, usernames, and user IDs. With approximately 1,727,000 tweets, the dataset provides a rich tapestry of discourse, spanning from October 15, 2020, to November 8, 2020.

A.2 Training Details

The In-Context Learning model is based on the OpenAI ”gpt-3.5-turbo-1106” model. We set the parameter temperature to be 0.1, so that the generated responses are more deterministic for the convenience of reproducing the project.

A.3 Political Questions

We used the following 10 political topics to form our political questions to our extracted personas.

1. allowing transgender athletes
2. gun control
3. abortion legalization
4. the universities practicing affirmative action
5. raising fossil fuel taxes
6. helping undocumented immigrants
7. death penalty in the US
8. allowing women to serve in military combat roles
9. legalization of Marijuana
10. reducing interest rates for student loans

All the questions used the same question template below, and the topics are the only difference.

Please share your level of agreement with [topic], with a score from 1 to 5, where 1 indicating strongly disagree and 5 indicating strongly agree. The score should be an integer without any other words in the same line. In a new line, explain the reasons behind your level of agreement.