Predicting Political Affiliations of Social Media Users

Haoyu Fu
h6fu@ucsd.edu

Yixuan Zhang
yiz119@ucsd.edu

Haojian Jin
h7jin@ucsd.edu

Abstract

Traditional methods for political polling often focus on the political opinions of a society as either an integrity or some groups. We present a novel approach to predict the political positions in the scale of individuals, which helps the exploration on the latent logics of the political trends. By conducting in-depth sentiment analysis and manual labeling on a selected sample of Twitter users, this project aims to train a model to learn the public's stance on specific socio-political issues and predict responses to new survey questions. The research methods include applying natural language processing techniques to analyze user tweets, and utilizing pre-trained GPT-2 for generating semantic embeddings of questions, which leads to the design of a neural network model capable of accurately predicting user opinions. To enhance the model's generalization ability, strategies such as cross-layers and deep cross-network were employed to prevent overfitting.

Website: https://haoyufu2.github.io/Twitter_Analysis_Website/

Code: https://github.com/HaoyuFu2/DSC180B_Twitter_Political_Analysis

1 Introduction

1.1 Project Description

The core objective of this project is to select a sample of Twitter users and, through in-depth sentiment analysis and machine learning on manual labeling conducted by team members on these selected accounts, and simulate responses to political questions. By analyzing these users' Twitter feedback for training our model, it will learn to understand the public's stance on specific socio-political issues and predict these users' responses to new survey questions.

On the technical implementation front, we first collected and cleaned Twitter data to build an initial dataset. Using sentiment analysis and manual labeling, we created a small base dataset and performed logistic regression to cluster different groups of users. With this clustering settings, we applied the cluster to the whole dataset and train large language models with the dataset. Further, we generated several embeddings representations of dataset features using pre-trained GPT-2 and designed a neural network model that could integrate user Twitter information to enhance the accuracy of predicting user opinions. To prevent overfitting, the model used cross-layers and deep cross-network to enhance the ability to process feature interactions.Our initial model choice was GPT-2-small, and the accuracy was surprisingly well given the model's small training parameters.

1.2 Literature Review

We reviewed a study that proposed a novel framework that utilized large language models to predict responses and to impute missing values in social surveys. Our research was inspired by this paper's idea to make predictions by utilizing large language models. However, the dataset used by this study was in a question-answer survey format, with all the answers being binary results. We therefore made large effort to modify their framework so that it can work in our research scope which is to predict one's opinion based on natural language sentences. (Kim and Lee 2023)

2 Methods

2.1 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 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 time this tweet was posted, and the text content of the tweet. We also removed all the url links, @

and # in the tweets, because we don't want these being generated in later processing steps. The cleaned dataset contains 556,334 rows of tweets made by 6400 unique users.

2.2 Clustering

To perform fine tuning and training on the dataset, we want to find a well defined label on each user. However, it's difficult to programmatically extract one's political positions from natural language sentences, so we alternatively imputed them based on manual labelling. Firstly we have read the tweets of 50 unique users from both original datasets, manually determine whether they support Trump or Biden in the 2020 Presidential Election. Then we performed sentiment analysis on each tweet to calculate the scores of its subjectivity, sentiment, strong negation, excessive marks, and topic diversity. The manual labels and sentiment scores resulted in a sparse matrix, therefore we used principal component analysis to reduce the dimensions of the results. We used logistic regression to cluster the manual labelled dataset, and applied the clustering to the original dataset so that we have a well labelled for our later processes.

2.3 Fine Tuning with Large Language Model

With the above processing, we obtained a well labelled dataset we can use to train and predict. We made three embeddings on the datasets. The first embedding contains the users' id number to distinguish the tweets posted by individuals. The second embedding contains the text content of tweets. We fine-tuned large language models (such as GPT-2-small) with these text information in order to extract the latent semantic information. The third embedding contains the year which the tweet was posted. In the context of our research, the year is fixed to be in 2020, and we kept this embedding for the convenience of future study on dataset across years. The embeddings are processed with cross layers and dense layers and then finally delivered as individual-level predictions.

2.4 Evaluation

We would evaluate our approach in two parts. The first part is about the accuracy of the clustering. The second part is about the accuracy of predictions made by large language models. We made separate evaluation because the predictions from large language models can be largely impacted by the result of clustering.

3 Results

3.1 Clustering

By utilizing the above mentioned approach, we achieved an accuracy of approximately 82% on the validation set. While it is worth noting that most of the true labels are negative results, which means the high accuracy was majorly contributed by true negatives. This 82% accuracy allows us to continue on the next steps, as the predictions using large language models depend on labels made by this process, and a low accuracy on clustering would make the prediction meaningless even if it has a good performance.

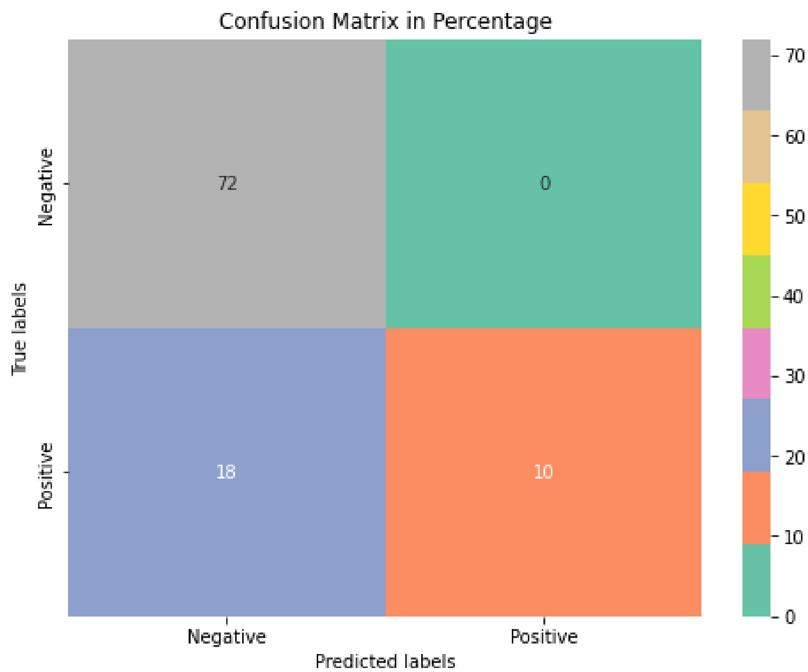

Figure 1: The confusion matrix of the clustering on validate set.

3.2 Prediction using Large Language Models

After training with the cluster-labelled dataset with large language models, the overall accuracy was approximately 72.7% on the validation set. The performance is not very high largely due to we are running the code with GPT-2-small, which is a pretty old language model, with small parameters compared to the modern models. We picked GPT-2-small mainly because it requires less computing resources. With the work we have currently done, we will attempt to improve this accuracy by using more advanced models with more parameters, such as Alpaca-7b.

4 Discussion

In reflecting upon the limitations of our project, several key aspects merit attention. Firstly, our analysis was confined to use manual labelling and it's challenging to validate our predictions with ground truth in larger scale beyond manual labelling. In future work, we may improve it by collecting our own surveys about political positions and respondents' Twitter/X account, so we have a better dataset which suits our research.

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 modified and used the prediction framework from our reference paper to make predictions that reflect Twitter users' political affiliations. We developed another approach different from the original methodologies, so that we can perform the predictions on the natural language sentences. This is fulfilled by learning (clustering and sentiment analysis) from human labeling on a small dataset, and apply to the whole dataset. This approach showed promising results in reflecting people's political affiliations solely based on their comments on social media, which is a potentially novel approach to reduce the high expense of traditional political polling.

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 to further improve the reliability of our model.

References

Kim, Junsol, and Byungkyu Lee. 2023. "Ai-augmented surveys: Leveraging large language models for opinion prediction in nationally representative surveys." *arXiv preprint arXiv:2305.09620*