ChatGPT Bias

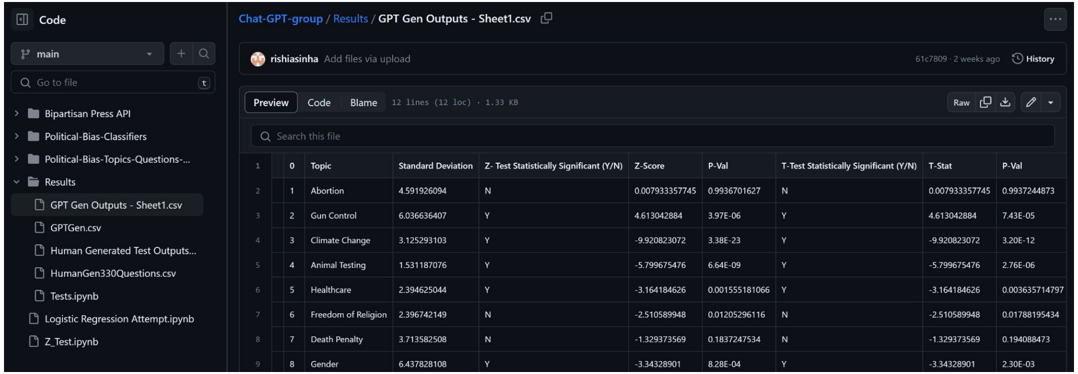
Hiresh Poosarla, Rishi Sinha, Hayden Fu, Alexander Suen, Aamod Gandhi, Vincent Lo, Harshaan Chugh, Lavi Avigad, Harish Senthilkumar, Karthik Subramanian

Mentor: Phil Mui

Introduction & Background

- What is ChatGPT and how does it work?
- What are some issues and problems with ChatGPT?
- How do we aim to analyze and quantify ChatGPT's (political) bias?

Methodology



Code

main

Go to file

- Bipartisan Press API
- Political-Bias-Classifiers
- Political-Bias-Topics-Questions-...
- Results
  - GPT Gen Outputs - Sheet1.csv
  - GPTGen.csv
  - Human Generated Test Outputs...
  - HumanGen330Questions.csv
  - Tests.ipynb
- Logistic Regression Attempt.ipynb
- Z_Test.ipynb

Chat-GPT-group / Results / GPT Gen Outputs - Sheet1.csv

rishiasinha Add files via upload 61c7809 · 2 weeks ago History

Preview Code Blame 12 lines (12 loc) · 1.33 KB

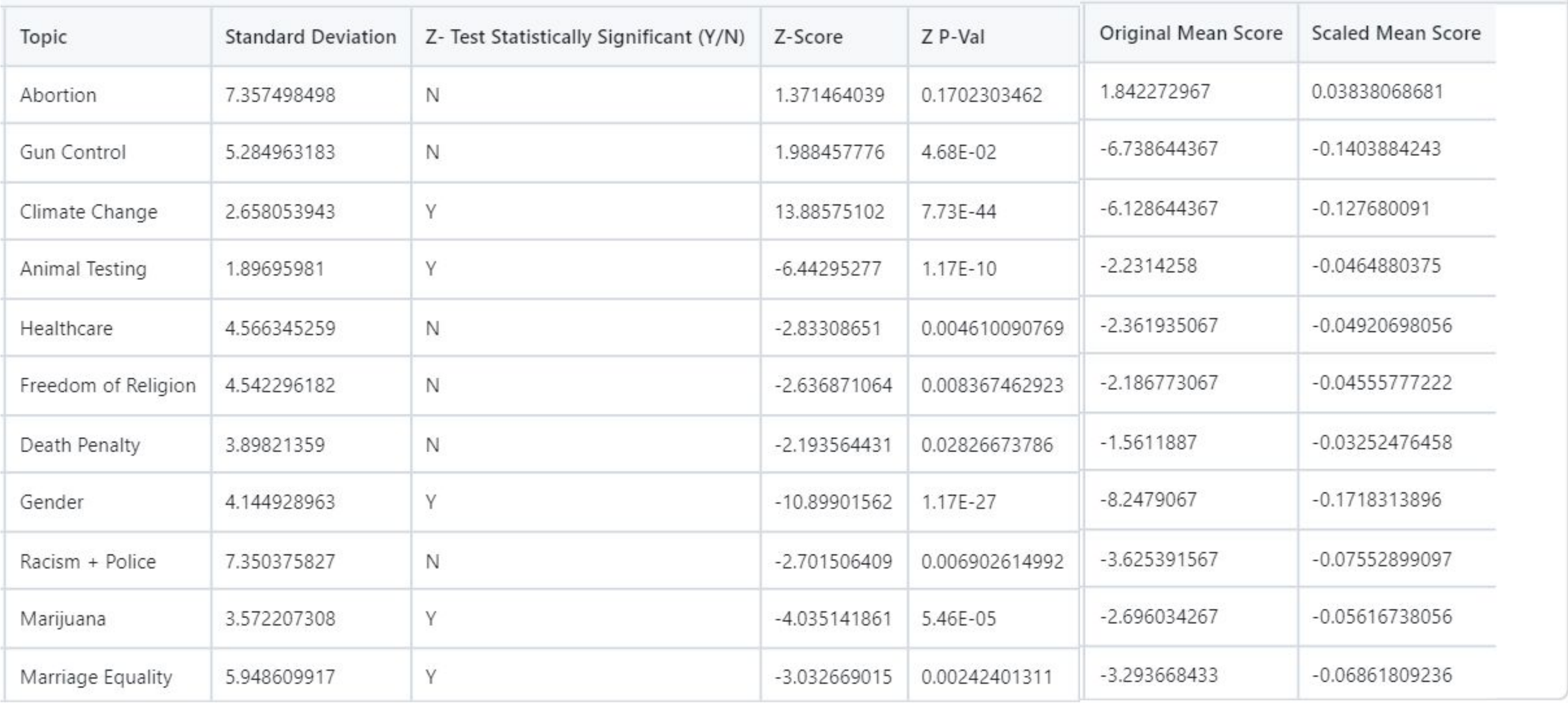
Search this file

0	Topic	Standard Deviation	Z- Test Statistically Significant (Y/N)	Z-Score	P-Val	T-Test Statistically Significant (Y/N)	T-Stat	P-Val
1	Abortion	4.591926094	N	0.007933357745	0.9936701627	N	0.007933357745	0.9937244873
2	Gun Control	6.036636407	Y	4.613042884	3.97E-06	Y	4.613042884	7.43E-05
3	Climate Change	3.125293103	Y	-9.920823072	3.38E-23	Y	-9.920823072	3.20E-12
4	Animal Testing	1.531187076	Y	-5.799675476	6.64E-09	Y	-5.799675476	2.76E-06
5	Healthcare	2.394625044	Y	-3.164184626	0.001555181066	Y	-3.164184626	0.003635714797
6	Freedom of Religion	2.396742149	N	-2.510589948	0.01205296116	N	-2.510589948	0.01788195434
7	Death Penalty	3.713582508	N	-1.329373569	0.1837247534	N	-1.329373569	0.194088473
8	Gender	6.437828108	Y	-3.34328901	8.28E-04	Y	-3.34328901	2.30E-03

The Bipartisan Press

- Human-Generated Questions
- GPT Questions
- Bipartisan Press API
- Statistical Z-Test
  - Z-score

Results



Topic	Standard Deviation	Z- Test Statistically Significant (Y/N)	Z-Score	Z P-Val	Original Mean Score	Scaled Mean Score
Abortion	7.357498498	N	1.371464039	0.1702303462	1.842272967	0.03838068681
Gun Control	5.284963183	N	1.988457776	4.68E-02	-6.738644367	-0.1403884243
Climate Change	2.658053943	Y	13.88575102	7.73E-44	-6.128644367	-0.127680091
Animal Testing	1.89695981	Y	-6.44295277	1.17E-10	-2.2314258	-0.0464880375
Healthcare	4.566345259	N	-2.83308651	0.004610090769	-2.361935067	-0.04920698056
Freedom of Religion	4.542296182	N	-2.636871064	0.008367462923	-2.186773067	-0.04555777222
Death Penalty	3.89821359	N	-2.193564431	0.02826673786	-1.5611887	-0.03252476458
Gender	4.144928963	Y	-10.89901562	1.17E-27	-8.2479067	-0.1718313896
Racism + Police	7.350375827	N	-2.701506409	0.006902614992	-3.625391567	-0.07552899097
Marijuana	3.572207308	Y	-4.035141861	5.46E-05	-2.696034267	-0.05616738056
Marriage Equality	5.948609917	Y	-3.032669015	0.00242401311	-3.293668433	-0.06861809236

Conclusion

We used the Bipartisan Press API to score the bias of responses from ChatGPT to a mix of human and GPT-generated questions.

ChatGPT is (relatively) unbiased!

- Most scores were negative (left-leaning)
- Overall, ChatGPT is not polarized in either direction

Future Works

- Prompt generation to reduce ChatGPT's bias
- Linguistic correlation for prompt <-> bias
  - Prompt generation models
- Eventually generalize this to other GPT models
  - And eventually all other future large-language models

Thank you!