Identifying and Manipulating Personality Characteristics in Pretrained Language Models

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Graham Caron

UNC NLP

Research Question

- There is extensive work in psychology and cognitive science investigating and characterizing personality traits in humans.
- Humans are increasingly interacting with language models like BERT and GPT2.
- This work analyzes personality traits in GPT2 and BERT by applying language-based questionnaires.
- The study looks at whether these models mirror personality in context and thus could be influenced predictably.
- Implications:
  - Communication efficiency/accuracy in text suggestions settings
  - Mirroring of traits in dialog systems

Background

Transformer Architecture

- Transformers are deep learning models designed for natural language data
- They can be set up for a wide range of natural language processing (NLP) tasks including language modeling.

Big Five Dimensions of Personality

- Extroversion (E)
- Agreeableness (A)
- Conscientiousness (C)
- Emotional Stability (ES)
- Openness to Experience (OE)

Assessment Procedure

- A 50 Item personality assessment from the International Personality Item Pool (IPIP) was used to measure the Big Five dimensions.
  - Subjects to respond with one of 5 answer choices.
- The assessment was adjusted from a question answering format to fill in the blank (BERT & GPT2 not pretrained for closed book question answering).
- Post-adjustment item example:
  - "I am {BLANK} the life of the party."
- Post-adjustment answer choices:
  - never(1)
  - rarely(2)
  - sometimes(3)
  - often(4)
  - always(5)
- 10 items were assigned to each dimension. Using the numerical values associated with responses to these items and a standard scoring procedure from the IPIP, a score out of 40 was calculated for each dimension.

Evaluating Dimension Scores

- Base dimension scores (X_{base}) were evaluated for each model using the previously outlined assessment procedure.
- A z-test was performed against the human population distribution for each dimension and percentiles of the model scores were calculated.

Dimension	X_{base}	z-val, p-val	P (%)
E	18	-0.42, 0.67	34
A	27	0.061, 0.95	52
C	25	-0.071, 0.94	47
ES	22	0.27, 0.78	61
OE	25	-0.73, 0.46	23

Table 3: BERT Base Model Evaluation Results & Analysis

Dimension	X_{base}	z-val, p-val	P (%)
E	21	-0.093, 0.93	46
A	26	-0.078, 0.94	47
C	29	0.48, 0.63	68
ES	25	0.62, 0.54	73
OE	28	-0.27, 0.78	39

Table 4: GPT2 Base Model Evaluation Results & Analysis

The z-test results in table 3 & 4 suggest that no base scores differed significantly from the population means.

Manipulating Dimension Scores

- The models were evaluated with assessment items and answer choices (modifiers) serving as context (see paper for more details).
- Differences between base scores and scores with context were calculated as Δ_{cm}.
- The relative expected change in score (expected behavior) for a context item/modifier pair was calculated as the context/modifier rating (r_{cm}).
- Δ_{cm} and r_{cm} were plotted against each other.
- Figure 2 shows the mean and median Δ_{cm} at each r_{cm} for both models.
- If the model mirrored the personality in the context and thus behaved as expected there would be a positive linear correlation between Δ_{cm} and r_{cm}.

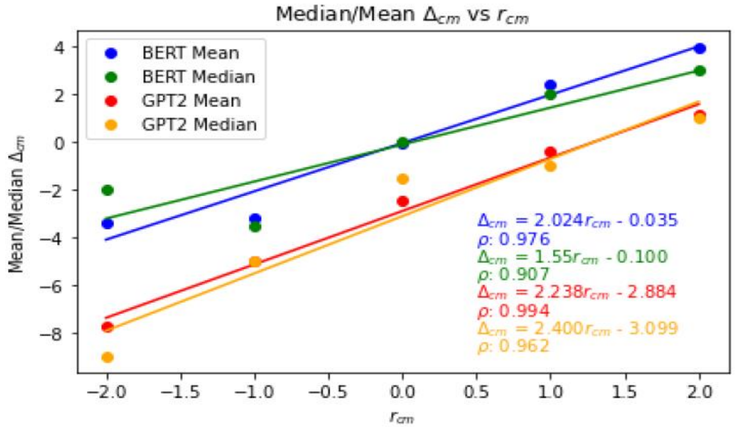

Figure 2: Mean & Median Δ_{cm} vs r_{cm}

The relatively strong correlation coefficients (ρ) from figure 2 suggest that the models reacted to context as expected.

Manipulating Dimension Scores Cont.

- Scores for different dimensions and context may not be directly comparable.
- To account for this, data was split such that the scores calculated using each context item were plotted against their context/modifier ratings.
- Correlation coefficients for these graphs were plotted in the figure 3 histograms.
- Note: the paper further breaks down the data to draw conclusions about each dimension.

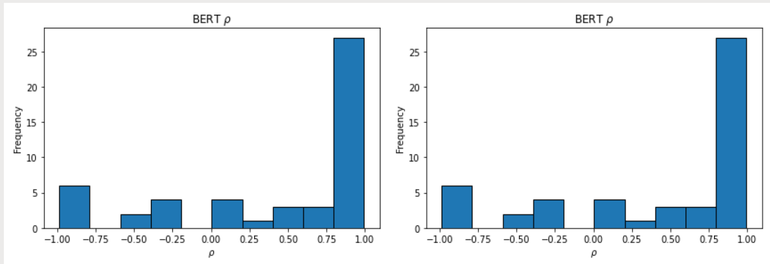

Figure 3: Histograms of all ρ

The results from figure 3 show large groupings of correlation coefficients around 1, providing further evidence that that models mirror personality in context.

Other Experiments

- Context w/o modifiers: A similar experiment was performed using slightly adjusted context (see paper for more details).
- Reddit context: Models evaluated on context from reddit threads on personality.
  - A logistic regression model was trained on n-gram count representations of the context (x) and the corresponding dimension scores (y).
  - The strongest feature weights were mapped to phrases in the context, indicating sequences that had the largest effect on the dimensions scores.
  - A qualitative analysis indicated that phrases with the largest weights generally caused the expected behavior.

Discussion

- Conclusions
  - BERT & GPT2 have identifiable personality traits.
  - The models reflect personality in context they see.
  - Context can be used to predictably manipulate the models' personality.
- Further Work
  - Evaluate personality of models with context provided by survey responses & compare against personality assessments completed by subjects.
  - Analyze how models reflect personality in a dialog setting

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL