Psychological Microtargeting in Online Environments

Adam Sutton – NEWS project
Acknowledgments

Dr Almog Simchon
Dr Matthew Edwards
Prof Stephan Lewandowsky
How does Psychological Microtargeting work in political campaigns?

- Target undecided voters

- Derive personality makeup for *each* individual

- Construct a persuasive message to sway individuals for political action or inaction based on ideology AND personality
ABSTRACT

A social networking system obtains linguistic data from a user’s text communications on the social networking system. For example, occurrences of words in various types of communications by the user in the social networking system are determined. The linguistic data and non-linguistic data associated with the user are used in a trained model to predict one or more personality characteristics for the user. The inferred personality characteristics are stored in connection with the user’s profile, and may be used for targeting, ranking, selecting versions of products, and various other purposes.
Personality-based Microtargeting

- **Does it work?**
  - Evidence suggests it does (Matz et al., 2017; *PNAS*)

- **Is it a problem?**
  - Depend on who you ask: unacceptable in Germany but passes in the US
  - Personalization for political campaigning is unacceptable across the board
  (Kozyreva et al, 2021; *Humanit. Soc. Sci. Commun*)
Solution

• Boosting: empowering individuals to make **informed** decisions
  • Letting people know psychological microtargeting exists + information about their personality leads to accurate identification of such attempts (Lorenz-Spreen et al., 2021; *Sci Rep*)

• How can we boost individuals in online environments?
• How can we know when people are being microtargeted?
Uncover the algorithms in action
The Current Project

• Population: Reddit users of fiction-writing communities
• Text-based models:
  • Model 1: predict stable psychological characteristics based on the text people produce
  • Model 2: predict stable psychological characteristics based on the text people consume
• Find if indeed psychologically concordant messages are more persuasive
• Apply in the real world

bristol.ac.uk
Model 1 & Model 2

**Model 1**
- Sample of Reddit participants
- Personality inventory
- Comments (text)
- Machine learning

**Model 2**
- Sample of Reddit users
- Comments (text)
- Posts (text)
- Predicted personality
Model 1: Method

- VW Fiction collection
  - Communities of Fiction writing
  - 32,344 Potential Participants
  - 9,244 sent PMs
  - 1,100 participants
  - 650,000 comments

- NEWS - collection
  - Communities of Political News
  - 18,293 Potential Participants
  - 1,063 sent PMs
  - 123 participants
  - 290,000 comments

- Measures
  - BFI-2 (Soto & John, 2017; JPSP)
  - SVS-PVQ (Schwartz, 1992; 2012)
Model 1: Text Modeling

Input Features

- Reddit Comment
  - BERT
  - Regression Head
  - Predictions: 0.3, 0.2, 0.9, 0.7, 0.6

- Reddit Comment
  - BERT
  - Regression Head
  - Predictions: 0.3, 0.1, 0.2, 0.7, 0.3

- Reddit Comment
  - BERT
  - Regression Head
  - Predictions: 0.2, 0.3, 0.5, 0.9, 0.7

Output Prediction

Average Predictions

Person #1: 0.3, 0.2, 0.8, 0.8, 0.7

Adapted from Jay Alammar
https://jalammar.github.io/illustrated-bert/
Model 1: Results for NEWS

- **Training set:** 100% - Fiction
- **Test set:** News Set (N = 123)
  - Average performance:
    - Pearson's $r = 0.24$
  - Correlation between the model predictions and ground truth

- Small test set limited to difficulties with reddit sampling

<table>
<thead>
<tr>
<th>Personality Dimension</th>
<th>Pearson’s $r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extraversion</td>
<td>0.22</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>0.23</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>0.25</td>
</tr>
<tr>
<td>Neuroticism</td>
<td>0.27</td>
</tr>
<tr>
<td>Openness to Experience</td>
<td>0.23</td>
</tr>
</tbody>
</table>
Model 1: Results for Fiction

- Training set: 80% Fiction
- Test set: 20% Fiction Set (N = 215)
  - Average performance:
  - Pearson's $r = 0.33$
  - Correlation between the model predictions and ground truth
- Performance within meta-analytic estimates (SOTA; Eichstaedt et al., 2021; Psych Methods)

<table>
<thead>
<tr>
<th>Personality Dimension</th>
<th>Pearson’s $r$ [95% CI]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extraversion</td>
<td>0.26 [0.13, 0.38]</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>0.35 [0.23, 0.46]</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>0.37 [0.25, 0.28]</td>
</tr>
<tr>
<td>Neuroticism</td>
<td>0.28 [0.15, 0.40]</td>
</tr>
<tr>
<td>Openness to Experience</td>
<td>0.39 [0.28, 0.50]</td>
</tr>
</tbody>
</table>
What linguistic features best predict personality?

<table>
<thead>
<tr>
<th>Agreeableness</th>
<th>Conscientiousness</th>
<th>Extraversion</th>
</tr>
</thead>
<tbody>
<tr>
<td>my_husband</td>
<td>my_not_the_fuck</td>
<td>had</td>
</tr>
<tr>
<td>her</td>
<td>im_not_the_fuck</td>
<td>kill</td>
</tr>
<tr>
<td>I've</td>
<td>don't</td>
<td>kill</td>
</tr>
<tr>
<td>she's</td>
<td>kinda</td>
<td>so</td>
</tr>
<tr>
<td>kinda</td>
<td>so</td>
<td>much</td>
</tr>
<tr>
<td>you</td>
<td>i</td>
<td>don't</td>
</tr>
<tr>
<td>husband</td>
<td>and</td>
<td>i</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Neuroticism</th>
<th>Openness</th>
</tr>
</thead>
<tbody>
<tr>
<td>talk</td>
<td>I know</td>
</tr>
<tr>
<td>was</td>
<td>sorry</td>
</tr>
<tr>
<td>oftentimes</td>
<td>i just</td>
</tr>
<tr>
<td>guides</td>
<td>so</td>
</tr>
<tr>
<td>said</td>
<td>with</td>
</tr>
<tr>
<td>level</td>
<td>the</td>
</tr>
<tr>
<td>transmutation</td>
<td>my</td>
</tr>
<tr>
<td>connected_with</td>
<td>kidneys</td>
</tr>
<tr>
<td>you</td>
<td>can</td>
</tr>
<tr>
<td>i</td>
<td>feel</td>
</tr>
<tr>
<td>i'm</td>
<td>weight_in</td>
</tr>
<tr>
<td>and</td>
<td>i</td>
</tr>
<tr>
<td>but</td>
<td>i</td>
</tr>
</tbody>
</table>

Pearson's r

[Heatmap showing correlation between linguistic features and personality traits]
Model 2 – Currently Fiction Only

Model 1
Sample of Reddit participants
- Personality inventory
- Comments (text)
  
  Machine learning

Model 2
Sample of Reddit users
- Comments (text)
- Posts (text)

Predicted personality

1.

Number 2
Model 2: Method for Fiction

- Reddit collection
  - N = 4,466 participants
  - 1,756,819 comments – Used for Model 1 predictions.
  - 4,466 unique pieces of fiction – Used for Model 2 training.
Model 2: Consumed Fiction Predictions

Input Features

Reddit Posts

BERT

Elastic Net Regression on Layers 11-12

Output Prediction

Person #1

Neuroticism 0.3
Conscientiousness 0.2
Agreeableness 0.8
Openness 0.8
Extraversion 0.7
Model 2: Results

- 5-fold Cross Validation
  Average Performance: Pearson’s $r = 0.106$

- Ground Truth
  – N = 689

<table>
<thead>
<tr>
<th>Personality Dimension</th>
<th>Pearson’s $r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extraversion</td>
<td>0.07</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>0.11</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>0.11</td>
</tr>
<tr>
<td>Neuroticism</td>
<td>0.13</td>
</tr>
<tr>
<td>Openness to Experience</td>
<td>0.11</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Personality Dimension</th>
<th>Pearson’s $r$ [95% CI]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extraversion</td>
<td>0.06 [-0.02, 0.13]</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>0.12 [0.05, 0.20]</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>0.09 [0.02, 0.16]</td>
</tr>
<tr>
<td>Neuroticism</td>
<td>0.08 [0.01, 0.16]</td>
</tr>
<tr>
<td>Openness to Experience</td>
<td>0.08 [0.01, 0.15]</td>
</tr>
</tbody>
</table>
Size Matters

- Isn’t an effect size of $r = 0.1$ negligible?
  - Yes, for a particular event, not at the aggregate or at scale

- Real world examples:
  - The effect of antihistamines on runny nose and sneezing: $r = 0.11$
  - The effect of ibuprofen on pain relief: $r = .14$
  - The correlation between extraversion and spent on holiday shopping: $r = .09$
What linguistic features best predict personality?

Agreeableness  Conscientiousness  Extraversion

Neuroticism  Openness

bristol.ac.uk
What can we say about Model 2?

- Content vs. Style
  - Seems to be more sensitive to “how” in contrast to “what”
  - Which can be consistent with “linguistic markers”

- The linguistic features seem to overlap
  - Evidence for the General Factor of Personality? (Musek, 2007; *J Res Pers*)
    - Five personality dimensions can be reduced to two dimensions (Neuroticism and Agr/Con/Ext/Opn)
The current project

- Population: Reddit users of fiction-writing and political news communities
- Text-based models:
  - Model 1: predict stable psychological characteristics based on the text people produce
  - Model 2: predict stable psychological characteristics based on the text people consume
- Find if indeed psychologically concordant messages are more persuasive
- Apply in the real world

bristol.ac.uk
Next steps and challenges

• Behavioral studies based on the applicability of Model 2
  • Are personality-congruent political ads rated as more persuasive? (different sample; experimental setting)

• Real-world application
  • How can we harness the science of boosting in developing interventions “in the wild”?
To learn more about REPHRAIN, our future plans and how to get involved:

[Website] www.rephrain.ac.uk

[Twitter] @REPHRAIN1

[Email] rephrain-centre@bristol.ac.uk

We would love to hear from you. Thank you!

TeDCog group: sks.to/tedcog