RU-ADEPT:
Russian Anonymized Dataset with Eight Personality Traits


* University of Maryland, College Park, MD, U.S.A.
† Accenture, Washington, D.C., U.S.A

{crytting, vnovak, jhull1, vmfrank, jgwl, lmillers}@umd.edu, paul.rodrigues@accenture.com
Overview

RU-ADEPT – a dataset of Russian authors' personality trait scores, demographic information, with associated corpus of the authors' cross-contributions to (up to) four different social media platforms

Personality Survey Data

- OCEAN: Big Five Inventory (BFI-2)
- Dark Triad: Short Dark Triad (SD3)

Demographics Info

- Age, Gender, Urban/Rural

Social Media Text

- VKontakte (VK)
- LiveJournal
- Blogger
- Moi Mir (My World)
Motivation

• Train machine learned classifiers for the application of Big Five and Dark Triad labels on Russian authors.

• Understand the differences in individuals’ exhibited personality across social media platforms.

• Understand the impact text length and formality have on personality identification
Novel Contributions

- 1st anonymized dataset pairing demographic and personality trait data with Russian-language social media content.
- 1st paper to describe collection of Dark Triad scores across multiple Russian-language social media platforms.
- 1st dataset of personality traits with participant-generated content sampled across several different social media sites (for a subset of participants, n=149)
# Prior Russian Text Studies of Personality Traits

<table>
<thead>
<tr>
<th>Authors &amp; Year</th>
<th>Traits</th>
<th>Text Genre/Platform</th>
<th># of Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bogolyubova et al. (2018)</td>
<td>Dark Triad</td>
<td>Facebook</td>
<td>6,724</td>
</tr>
<tr>
<td>Stankevich et al. (2019)</td>
<td>Big Five</td>
<td>VKontakte</td>
<td>1,020</td>
</tr>
<tr>
<td>Litvinova et al. (2016)</td>
<td>Various (inc. Big Five)</td>
<td>Prompted picture descriptions, letters, essays, etc.</td>
<td>1,145 (BF for 192)</td>
</tr>
</tbody>
</table>
SocialShare – Social Media Collection Platform

3rd party survey

Recruitment
Demographics

SocialShare

OAuth login requested
(API retrieval)

username requested
(HTTP retrieval)

3rd party survey

BFI-2 Short Dark Triad

Discontinue
(flag data for deletion)
Five Factor Model ("Big Five")

- Five Factors of Personality (a.k.a. OCEAN)
  - Extraversion
  - Agreeableness
  - Conscientiousness
  - Open-Mindedness
  - Negative Emotionality (sometimes called Neuroticism).

We used the *Big Five Inventory-2 (BFI-2)* – Soto et al. (2017)
- 60-item self-report questionnaire; 5-point Likert scale
- Russian translation by Shchebetenko et al. (2020)
Big Five Inventory-2 (BFI-2) Results

Scores are similar to those reported in Shchebetenko et al. (2020).
Dark Triad

• Dark Triad (Paulhus and Williams, 2002)
  • Machiavellianism - essentially a manipulative personality
  • Narcissism - involves thoughts and behaviors that espouse entitlement, superiority, and dominance
  • Psychopathy - characterized by four subtraits: high impulsivity and thrill-seeking, combined with low anxiety, low empathy for others

We used the *Short Dark Triad* (SD3) (Jones, 2014)
• 27-item self-report questionnaire; 5-point Likert scale
• Russian translation by Egorova et al. (2016)
Short Dark Triad (SD3) Results
## RU-ADEPT Corpus Statistics

<table>
<thead>
<tr>
<th>Platform</th>
<th># of Participants</th>
<th># of Posts</th>
<th># of Word Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>VK Post</td>
<td>1,292</td>
<td>371,290</td>
<td>6,739,054</td>
</tr>
<tr>
<td>VK Repost</td>
<td>984</td>
<td>434,979</td>
<td>28,164,325</td>
</tr>
<tr>
<td>Live Journal</td>
<td>27</td>
<td>9,854</td>
<td>2,572,032</td>
</tr>
<tr>
<td>Blogger</td>
<td>17</td>
<td>888</td>
<td>11,590,630</td>
</tr>
<tr>
<td>Moi Mir (My World)</td>
<td>55</td>
<td>5,590</td>
<td>26,359</td>
</tr>
</tbody>
</table>
Lexical Differences (Extraversion)

Scattertext (Kessler, 2017) illustrates differences between VK text of low & high levels of extraversion
Lexical Differences (Extraversion)

Lowest Decile:
- уровень (‘grade’, ‘level’)
- axax (laughter)

Highest Decile:
- театр (‘theater’)
- спектакль (‘performance’)

45:13 per 1k posts

7:2 per 1k posts
Lexical Differences (Extraversion)

Lowest Decile:
- уровень (‘grade’, ‘level’)
- ахах (laughter)

Highest Decile:
- театр (‘theater’) 1:12 per 1k posts
- спектакль (‘performance’) ~0:5 per 1k posts
Discussion, Limitations, and Future Work

• Data Collection and Interpretation
  • User-authored vs. “auto-authored” text (see Hull et al. 2021)
  • Limitation: Verifying account IDs for Blogger, Live Journal

• Open Science and Data Privacy
  • Human data creates ethical issues for open science
  • Future work: Pseudonymization for more natural de-identification
Conclusions

• A greater diversity of datasets can make for better social science.
  Which textual correlates of personality traits are universal?
  Which vary by language, platform, time of collection, etc.?

• RU-ADEPT is a starting place for studying personality traits in a less-Western language.
  First large pairing of Big Five + Dark Triad in a Russian language dataset.
How to Get the RU-ADEPT Dataset

• RU-ADEPT is *not* available for direct download
  • Scholars must submit reuse requests for ethics review.
  • Send your request to research_support@umd.edu.

• Please include in the request:
  • institutional and/or funding information,
  • a description of your research plans,
  • what portion of the dataset you need, and
  • any relevant human subjects research review determinations from your internal review board (IRB) or equivalent.
Thank you!


* University of Maryland, College Park, MD, U.S.A.
† Accenture, Washington, D.C., U.S.A

{crytting, vnovak, jhull1, vmfrank, jgwl, lmillers}@umd.edu, paul.rodrigues@accenture.com

For requests or questions about the corpus, write:
research_support@umd.edu