Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 109–118
Marseille, 20-25 June 2022

RU-ADEPT:

Russian Anonymized Dataset with Eight Personality Traits

C. Anton Rytting*, Valerie Novak*, James R. Hull*,
Victor M. Frank*, Paul R. Rodrigues†, Jarrett G.W. Lee*, Laurel G. Miller-Sims*
*University of Maryland
College Park, MD, U.S.A.
{crytting, vnovak, jhull1, vmfrank, jgwl, lmillers}@umd.edu

†Accenture
Washington, D.C., U.S.A
paul.rodrigues@accenture.com

Abstract

Social media has provided a platform for many individuals to easily express themselves naturally and publicly, and researchers have had the opportunity to utilize large quantities of this data to improve author trait analysis techniques and to improve author trait profiling systems. The majority of the work in this area, however, has been narrowly spent on English and other Western European languages, and generally focuses on a single social network at a time, despite the large quantity of data now available across languages and differences that have been found across platforms. This paper introduces RU-ADEPT, a dataset of Russian authors' personality trait scores—Big Five and Dark Triad, demographic information (e.g. age, gender), with associated corpus of the authors' cross-contributions to (up to) four different social media platforms—VKontakte (VK), LiveJournal, Blogger, and Moi Mir. We believe this to be the first publicly-available dataset associating demographic and personality trait data with Russian-language social media content, the first paper to describe the collection of Dark Triad scores with texts across multiple Russian-language social media platforms, and to a limited extent, the first publicly-available dataset of personality traits to author content across several different social media sites.

Keywords: personality, Big Five, Dark Triad, Russian language, author profiling

1. Introduction

There has been longstanding interest in understanding the manner in which people's personality traits may be manifested in what they say or write. With the advent of widespread online written communication through the Internet, including social media, it is now feasible to collect user-generated text paired with information on personality traits and demographical information at much larger scales than a decade ago. It is thus not surprising that many researchers have examined social media text to study how personality traits are manifest in one's expression of their language. Many—probably most—of these early studies looked exclusively at English (Golbeck et al., 2011; Farnadi et al., 2016, inter alia) though a few datasets exist in other (Western European) languages as well (Rangel Pardo et al., 2015).

While numerous ways of conceptualizing and measuring personality differences have been proposed, two commonly studied frameworks for studying personality in online media are the Five Factor Model (McCrae and John, 1992) and the "Dark Triad" (Paulhus and Williams, 2002). The former posits general traits of application to many aspects of life; the latter focuses on traits encompassing a propensity for malevolence or antisocial behavior. Each is briefly described below.

1.1. The Big Five Inventory-2 (BFI-2)

The Five Factor Model or "Big Five" utilizes five broad factors of personality: *Extraversion, Agreeableness, Conscientiousness, Open-Mindedness*, and *Negative Emotionality* (sometimes called *Neuroticism*). Several different instruments have been developed to measure the Big Five personality factors or traits. One recent instrument, the Big Five Inventory-2 (BFI-2) is a 60-item self-report questionnaire measuring the five personality traits as well as associated facets (Soto and John, 2017). This questionnaire is a recent update of the original BFI measure and has been used in hundreds of personality studies. It was designed to be short enough for participants to complete in less than ten minutes. Attested translations in ten languages are available from the author's website, with preliminary translations for 30 additional languages.[1]

[1]https://www.colby.edu/psych/personality-lab/

1.2. The Short Dark Triad (SD3) Scale

The concept of the Dark Triad was introduced by Paulhus and Williams (2002), who proposed to view three traditionally studied personality traits—Machiavellianism, narcissism, and psychopathy—as one triad of overlapping but distinct constructs (with some evidence that these are sub-clinical personality traits rather than psychological disorders). Paulhus and Williams describe Machiavellianism as essentially a manipulative personality. Narcissism involves thoughts and behaviors that espouse entitlement, superiority, and dominance. Psychopathy is characterized by four subtraits: high impulsivity and thrill-seeking, combined with low anxiety and empathy for others. Jones and Paulhus (2014) developed a 27-question questionnaire for these three traits called the "Short Dark Triad" or SD3.

As there is ever-increasing concern about various types of social ills from social media, including divisive, manipulative, or misinformative language, there is increased interest in how dark personality traits are expressed online, or influence online behavior (Moor and Anderson, 2019, provides a recent systematic review). For example, in a study of 6,724 Russian-speaking Facebook users, Bogolyubova et al. (2018) found gender differences and a main effect for one Dark Triad trait (psychopathy), with men more likely to send insulting messages and post aggressive comments than women, and men and individuals with high scores on psychopathy to be more likely to engage in harmful online behaviors.

1.3. Contributions of This Work

Within the computational social science community, there is a growing awareness for the benefits of a larger body of diverse datasets covering multiple languages and platforms, rather than assuming that datasets on highly resourced languages like English (or easily shared platforms like Twitter) are representative of human language as a whole (Bender, 2011; Tufekci, 2014, inter alia).

In this paper we address the need for greater diversity in text corpora for psychological traits by introducing RU-ADEPT: a corpus of individuals' traits from Russian-language social media. De-identified Russian text from four social media platforms (VK, LiveJournal, Blogger, and Moi Mir) are associated with basic demographic information (age, gender) and eight self-reported personality trait scores from two instruments (BFI-2 and SD3). This may be the first publicly-available dataset associating demographic and personality trait data with Russian-language social media content, the first paper to analyze Dark Triad scores across multiple Russian-language social media platforms, and while limited, the first publicly-available dataset of personality traits to author content across several different social media sites.

The remainder of the paper describes related resources and research (Section 2), brief descriptions of the data collection methodology and the resulting dataset (Sections 3-4), followed by a discussion of considerations of open science and data privacy, as well as anticipated contributions and known limitations of the dataset (Section 5).

2. Related Work

Studies of online language and personality traits are often conducted with some specific applied research question in mind, such as studying antisocial online behavior (Moor and Anderson, 2019). For obvious privacy and data sensitivity concerns, the social media data collected is seldom, if ever, made generally available to the research community. Within the computational social science research community, there is interest in training and evaluating automatic methods for inferring traits from text posted online. Comparing different algorithms and models requires the existence of common datasets for comparison; still, researchers must balance human subject privacy concerns with the researcher need for common evaluation corpora. One emerging solution is automatic de-identification of the text by masking named entities or replacing them with pseudonyms (Eder et al., 2019) as in the CoDE Alltag 2.0 Corpus (Eder et al., 2020). Another is sequestering sensitive text in secure environments for research purposes (data enclaves), such as the NORC data enclave (Lane and Shipp, 2008) or the UMD/NORC Mental Health Data Enclave (MacAvaney et al., 2021).

Although measures for the Big Five factors and the Dark Triad have both been translated into the Russian language—cf. Shchebetenko et al. (2020) for a translation of the BFI-2, as well as Egorova et al. (2016) for a translation of the SD3—there are, so far as the authors are aware, few corpora of social media text in Russian with personality trait labels (and none shared publicly). Bogolyubova et al. (2018) collected Facebook posts from 6,724 Russian users, in addition to the SD3 scale and data about harmful online behaviors. Stankevich et al. (2018) reports a small pilot dataset of 165 VKontakte (VK) profiles. However, due to the sparseness of usable, user-generated text in their dataset, they were unable to use the text to extract any useful lexical features, but had to perform personality trait inference using only very basic features on the text (such as average numbers of words and sentences, use of punctuation and uppercase). Ignatiev et al. (2019) and Stankevich et al. (2019) report the collection of a larger dataset of 1,020 VKontakte profiles including text of original posts,

repost information, and user profile information. Subsequent work by Ignatiev focuses on image data rather than text (Ignatiev et al., 2020). Unfortunately, Boglyubova, Stankevich, and Ignatiev do not explicitly mention how their datasets would be shared with other researchers.

One research group, the RuProfiling Lab, describes three corpora which include personality features in Russian such as the Big Five associated with written texts, though not from social media. The Ruspersonality Corpus contains 1,850 written Russian texts contributed as response to experimental stimuli (e.g., description of a picture, letter to a friend, essay on a given topic) from 1,145 authors, of which 192 also took the Big Five Personality Test (Litvinova et al., 2016). A related resource, RusNeuroPsych (Litvinova and Ryzhkova, 2018), associates picture descriptions and letters to friends with profiles of lateral organization of brain functions. It consists of two subcorpora: one from school-age children (252 texts by 246 respondents) and one from adults (392 texts by 209 participants). The adult subcorpus includes Big Five scores. RusIdiolect, a resource primarily for authorship attribution, also may be relevant to personality trait identification (Litvinova, 2020).

Datasets associating social media text with Dark Triad measures are even less commonly shared or reused (if at all). While a few studies correlate social media behavior with Dark Triad traits in Russian (Moskvichev et al., 2017; Bogolyubova et al., 2018, inter alia), the publications do not mention sharing of de-identified datasets. To our knowledge, no prior work has examined Dark Triad traits in VK or in multiple Russian-language social media platforms.

3. Dataset Collection Methodology

3.1. Participant Selection and Recruitment

Participant data was collected using an online survey hosted by an organization with deep expertise in survey design and deployment. Participants were native Russian speakers recruited by the survey organization, which maintains a large pool of potential participants for various survey panels; thus the participants in this study come from a self-selected group that has already indicated a general willingness to participate in surveys. Participants had the option to decline to participate, or request additional information; those who opted to participate had the option of ending participation at any time during the project.

Inclusion criteria required participants to:

- Give (and not withdraw) informed consent, as explained in the project's Institutional Review Board (IRB) human subjects research protocol and participant consent form;
- Complete all required surveys, including basic demographic information, BFI-2, and SD3;[2] and
- Indicate that they have a VK account and are willing to provide access to it. (Additional social media accounts could be provided.)

3.2. Initial Dataset Collection

After providing demographic information within the third-party survey web interface, participants were asked to share their social media pages via a web application written by the research team, which we called SocialShare. Although publicly accessible, the SocialShare site required pass-through parameters from the third-party platform in order to log in, limiting users to the participants who had already completed the initial surveys.

Continuing participants were required to allow SocialShare to access their VK data, which they approved by logging into VK directly and confirming the permission; this performed the task of retrieving an access token from VK's API. They were then asked for Blogger, Twitter, and LiveJournal usernames. If supplied, SocialShare confirmed the existence of a publicly available account on each platform using HTTP retrieval or API calls. If the account did not exist or its content was not made available publicly, a message explained the issue and allowed the participant to retry. Participants were also asked to log into Moi Mir to confirm access by SocialShare to their data, again using platform APIs to obtain access tokens. Finally, a freeform entry page was shown, allowing participants to enter information about any other social media platform they wished to share. The API credential and username information for each social media platform was stored in a message queue for further processing. Finally, the SocialShare site rerouted the participant back to the third-party web survey system where they were asked to complete the demographics survey (age, gender, employment category, rural/urban), the BFI-2 survey, and the SD3 survey. Respondents are able to receive their BFI-2 scores and explanation before they leave the survey site.

Participants were able to stop data collection at any time by clicking a link, supplied on each page, and confirming their discontinuation, which flagged any data for subsequent deletion and notified the third-party survey workflow of their discontinuation. Participants were also able to skip the submission of platforms other than VK by indicating that they did not use the specified site. In all cases, participant passwords were neither

[2]The initial collection protocol included two other tasks, which were discontinued in a second round of collection.

passed through SocialShare nor stored; any platform which required a login by the participant did so using the platform's API and OAuth procedures. For VK and Moi Mir, the obtained access tokens were used to retrieve the corresponding records via the corresponding APIs. All others sites used retrieval of data over HTTP, with the exception of the freeform entry page, which would be manually reviewed.

Only text data was retained and used for this study; image and video data was ignored due to the anticipated challenges of de-identifying these media. A full automated approach to anonymzing audio-video and images would need speech recognition, OCR, and facial recognition to identify candidates for anonymization, and audio clipping and blur to anonymize text and faces.

3.3. Quality Assurance with Revised Protocol

Data collection was conducted in two phases, with additional quality assurance measures introduced after initial quality control checks. See Tables 1-3 for statistics on the social media data collected in each phase (Table 1 and Table 2) and the aggregated social media data from both phases (Table 3). While the participant selection process largely remained the same, we introduced changes to ensure participants were active social media users and to seek more truthful survey response. These changes are described below.

3.3.1. Strengthened Requirements for Quantity of Social Media Content

Following the determination in an interim evaluation using the Russian Feature Extraction Tool, or RFET (Hull et al., 2021) that much more social media content, specifically text content that was (likely) generated by the participant, was required to perform the personality trait inferencing trials, we introduced a minimum required quantity of social media content in order to qualify for the study: at least 1,200 words of user content on at least one social media platform (after a filtering process to count only those posts that were likely to be written by the user). As it would be excessively burdensome for respondents to wait while we downloaded and parsed their content on the fly, we created proxies of quantity using post counts by measuring against the extracted data from the prior round. This gave us an estimate of at least 360 VK posts, or 500 Twitter posts, or 5 posts in either LiveJournal or Blogger. We added functionality in SocialShare to extract post counts at the point of participant submission for each platform; VK and LiveJournal post counts were read from labels in the public profile of the individuals, and Twitter and Blogger counts were retrieved via Web API. Partway through our collection, we found that certain VK accounts (most likely those marked as private) and all Moi Mir accounts did not provide public post counts, lowering eligibility opportunities. In response to this, we amended (with IRB approval) our survey to ask participants to estimate, for each social media platform they reported using, the number of months in which they actively created new posts. This provided an alternate method for VK and Moi Mir users to qualify based on their self-reported number of months of active posting: participants reporting at least 60 months of active VK posting or 36 months of active Moi Mir posting were also included in the survey. Respondents lacking a sufficient number of posts in any social media platform were disqualified from the study.

Platform	Participants	Posts	Tokens
Blogger	11	67	6,877
Live Journal	9	4,870	516,274
VK Post	996	265,676	5,030,679
VK Repost	855	342,342	21,713,401
Twitter	64	164,215	1,740,363
Total	1,005	777,170	29,007,594

Table 1: Descriptive statistics for data collected during Round 1

Platform	Participants	Posts	Tokens
Blogger	6	821	11,583,753
Live Journal	18	4,984	2,055,758
Moi Mir	55	5,590	26,359
VK Post	296	105,614	1,708,375
VK Repost	129	92,637	6,450,924
Twitter	32	258,163	2,866,371
Total	306	467,809	24,691,540

Table 2: Descriptive statistics for data collected during round 2

Platform	Participants	Posts	Tokens
Blogger	17	888	11,590,630
Live Journal	27	9,854	2,572,032
Moi Mir	55	5,590	26,359
VK Post	1,292	371,290	6,739,054
VK Repost	984	434,979	28,164,325
Twitter	96	422,378	4,606,734
Total	1,311	1,244,979	53,699,134

Table 3: Descriptive statistics for the full dataset

3.3.2. Protocol to Detect Invalid Survey Completion

In Round 1 of data collection we discovered that 28 participants had selected the same response type on the Likert scale for over 90% of their responses

for either personality instrument. Accordingly we implemented code to detect such careless behavior directly in the data collection process. Non-compliant respondents were offered an opportunity to re-take the survey, and were disqualified from the study if the behavior recurred. Five participants were automatically excluded from the study during the Round 2 collection for this behavior.

4. Description of the Dataset

The dataset consists of two separate parts: social media text corpus in JSON Lines (jsonl) format and the participant data in csv format. The key variable between the datasets is the participant_id. There are 1,311 participants in the datasets. Data was collected in two rounds. In Round 1, we collected social media data from the following platforms: VK, Twitter[3], Blogger and LiveJournal. In round 2, we expanded to collect data from Moi Mir, as well as the initial platforms. The original VK data included both users' original content (posts) and reposts, which were not necessarily created by the user. The VK data is split into VK_post and VK_repost. We excluded VK_repost from the analyses as the repost data could not provide reliable information about the participants' personality traits. The participant data includes demographic information provided by the participant (age, location, gender, etc.) and scores on the BFI-2 and SD3.

4.1. Social Media Text Corpus

The social media text corpus is in jsonl format. Each file contains data from only one platform; each line of the file represents a post on that platform. The social media text is anonymized in three ways (explained in section 4.1.1), and contains the metadata about the social media post, i.e., platform, participant_id, and the post_id (assigned by the UMD research team).

4.1.1. Anonymization of Social Media Data

Anonymization is the process of removing information that identifies an individual. This process focused on removing names of persons, locations, organizations, urls, personally identifying "contact information" such as telephone numbers, usernames, email and physical addresses, dates and numbers over four digits long from the text. Anonymization in our corpus takes three formats: removing personally identifying information (PII) from the text, replacing PII from the text with a tag (i.e., PER, ORG, HYP, etc.), and replacing PII from the text with a pseudonym (i.e., он 'he', она 'she', оно 'it', здесь 'here', etc.) Each social media post is available in all three anonymized formats, so that researchers can choose which is most relevant to their downstream tasks.

In developing our anonymization methodology, we consulted privacy and anonymization standards used for similar social media datasets (Rangel Pardo et al., 2015; Eder et al., 2020, inter alia). Further details of how de-identification of participants was done will be explained in a separate paper. Table 4 shows examples of content from three platforms after the de-identification process preceding pseudonymization.

Pipeline	Anonymized Text
VK	д. **HYP0077** С Днём рождения! Всего самого - самого...:)
Live Journal	**PER0155** философеет на глазах - в смылсе, раньше она философела себе потихоньку дома, а теперь вот на наших глазах.
Moi Mir	**ORG0007** а особенно **ORG0008** активно работает над фальсификацией истории не ведитесь на истинные открытия о Руси и русских, это запланированная акция имеет корни уже **NUM0002** -е десятилетие!

Table 4: Examples of content from social media collection after de-identification

4.2. Demographic and Personality Data

In order to create a corpus generally reflective of the Russian speaking population, we aimed for roughly equal gender balance and an age distribution (over 18) tracking the age distributions reported for VK users.[4] We also aimed for 75% of the participants to live outside of the two most populous metropolitan areas in order to have greater geographic diversity. Table 5 and Table 6 shows the age and location distributions of survey participants across both rounds of data collection.

We also include here distributions of participant self-report scores for personality traits measured by the BFI-2 and the Dark Triad (SD3). Figure 1 shows the raw scores for the five traits measured by BFI-2: Open-mindedness, Conscientiousness, Extraversion, Agreeableness, and Negative Emotionality (or Neuroticism). Possible scores range from 12 to 60. Figure 2 shows the raw scores for the three traits measured by SD3: Machiavellianism, Narcissism and Psychopathy. Possible scores range from 9 to 45 for each trait.

[3]While we collected Twitter data from users, we chose not to distribute the Twitter data for reasons explained in Section 5.1.

[4]We used a recent study by Brand Analytics.

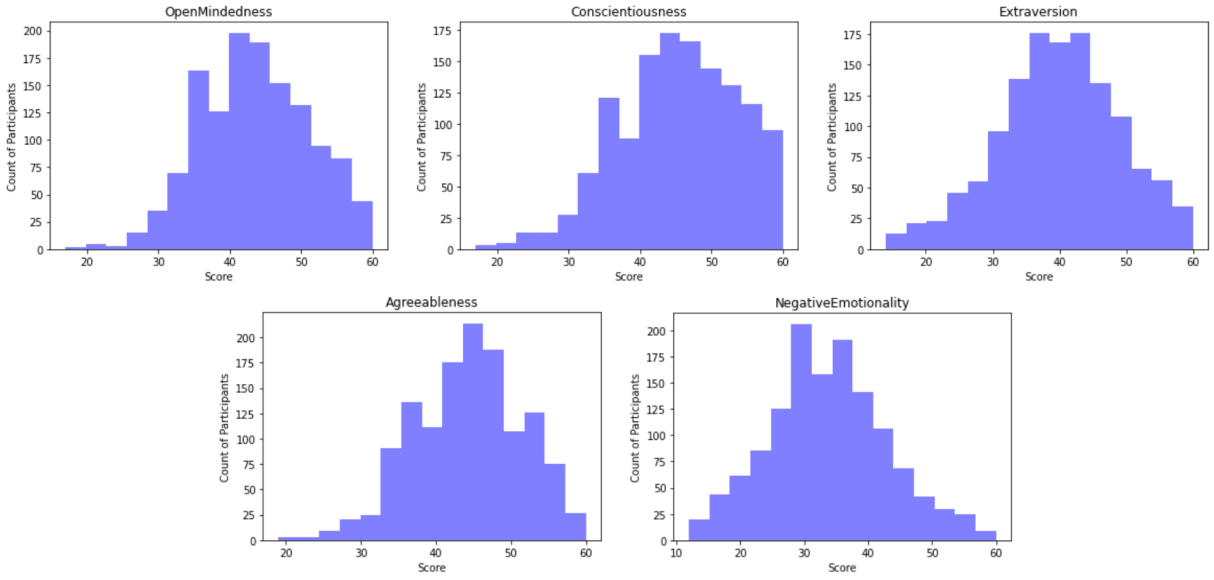

Figure 1: Distributions of raw self-report scores for the five factors across the participants in the RU-ADEPT dataset. Open-mindedness: mean = 43.6, standard deviation = 7.6. Conscientiousness: mean = 45.3, std. dev. = 8.3. Extraversion: mean = 39.9, std. dev. = 8.9. Agreeableness: mean = 44.3, std. dev. = 7.2. Negative Emotionality: mean = 33.7, std. dev. = 9.2.

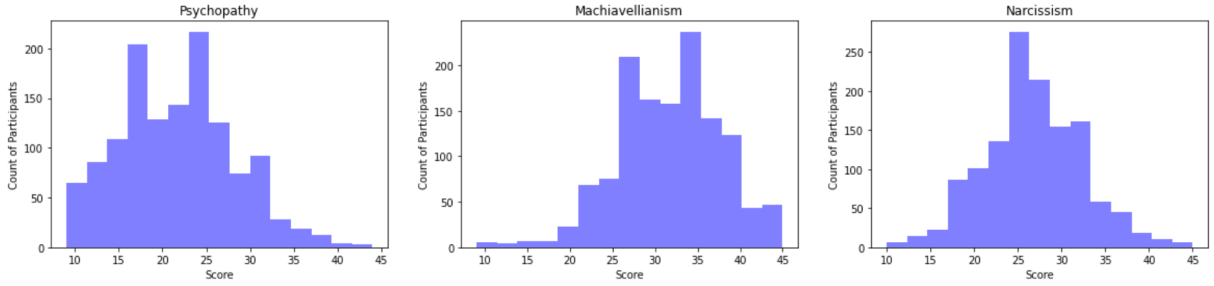

Figure 2: Distribution of Dark Triad scores across participants. Machiavellianism: mean = 31.6, standard deviation = 6.1. Narcissism: mean = 26.7, std. dev. = 5.6. Psychopathy: mean = 21.5, std. dev. = 6.5

Age Breakdown		
Age Group	Participants	Percentage
18-24	261	19.9
25-34	446	34.0
35-44	369	28.2
45-54	159	12.1
55-65	65	5.0
65 and Over	11	0.8
Total	1311	100

Table 5: Age distribution of survey participants

Location Breakdown		
Region	Participants	Percentage
Metropolis #1	245	18.7
Metropolis #2	122	9.3
Other Large City	683	52.1
Small Town	182	13.9
Rural Areas	79	6.0
Total	1311	100

Table 6: Location distribution of survey participants

4.3. Lexical Associations with Personality Traits

We believe that this corpus has great potential for exploratory analyses for generating new ideas and hypotheses for relations between language and personality, as well as for replicating and testing the universality of previous findings.

Detailed lexical and other linguistic analyses of this corpus are beyond the scope of this paper. As a small, illustrative example of the potential for such analysis we see in the corpus, we show a visualization of lexical items associated with the low and high ends of one dimension of personality, extraversion, using a freely available visualization package for highlighting differences between corpora (Kessler, 2017). We restrict this illustra-

tive example to a subset of the corpus: the portion taken from one social media platform (VK), excluding reposts, with limited pre-processing to the text and using SpaCy's Russian lemmatizer. The labeled scatterplot displayed in Figure 3 represents the relative frequencies of words used by participants who scored in the top decile for extroversion (score = 51.0 or higher, n=153, with 697,478 total tokens) and the bottom decile for extroversion (score of 28.0 or lower, n=131, with 365,042 total tokens). The x-axis shows the frequency in the bottom-decile corpus; the y-axis the frequency in the top-decile corpus.

We include also two sample posts from the corpus, with the corresponding output for emotional valence using the crowdsourced Russian version of the NRC Emotion Lexicon (Mohammad and Turney, 2013) as retrieved from RFET: (Hull et al., 2021).

a. text: Так хорошо отдохнул в деревне! response: 'neutral': 2, 'anticipation': 1, 'joy': 1, 'trust': 1, 'surprise': 1, 'positive': 1, 'token not in lexicon': 2

b. text: Повышаете возраст,платите зарплату и пенсию, как платят в Европе и Америке. response: 'fear': 1, 'negative': 1, 'neutral': 2, 'trust': 3, 'positive': 3, 'anticipation': 3, 'joy': 3, 'token not in lexicon': 6

5. Discussion

As mentioned above, there is growing societal concern about harmful effects of social media, some of which (such as cyberbullying) may correlate with dark personality traits, and so be better understood (and detected) as techniques for detecting dark traits improve (Balakrishnan et al., 2019). Within the scientific community, a perhaps more fundamental concern is growing about the dangers of basing our understanding of social phenomena on just a few studies, and a growing realization of the need for replication of studies in a variety of contexts. This applies not only to experiments in the lab, but just as much (if not more so) to studies based on collections of social media data (Liang and Fu, 2015; Olteanu et al., 2019) .

5.1. Open Science and Data Privacy

While a corpus of the kind described in this paper is clearly relevant to several research communities of interest, including computational and corpus linguistics, computational social science, and (cross-cultural) psychology of personality, to name a few, the interests and concerns of these communities differ. On the one hand, we believe that a greater diversity of available corpora can add to the goals of open science and help combat issues of replication in the social sciences. On another hand, given the ease of searching social media and other publicly available electronic text, there are serious ethical considerations to releasing any dataset based on social media, even one that has been altered with masking or other anonymization procedures, which must not be taken lightly. These considerations increase in seriousness with the sensitivity of associated data. At one extreme, suicidality and mental health issues obviously need a great deal of protection such as that afforded by a data enclave (MacAvaney et al., 2021). Personality traits such as big five may require relatively little; dark traits, while sub-clinical, are arguably more sensitive.

Likewise, the considerations differ from platform to platform based on terms of service and ease of searchability. (Private posts, while potentially more personal and sensitive for individuals, have the benefit of being less searchable.) Sets of associated data from multiple sources may be more sensitive than single datasets alone.

While our collection included Twitter data (for a small subset of participants who opted-in and shared it with us), we could not include it in our dataset for distribution to other researchers, and therefore chose to not discuss the results of our Twitter analysis in this paper. The standard approach in sharing Twitter research involves distributing the TweetIds with any annotation columns. Researchers who receive the corpus can re-download the metadata and text content from the shared TweetId and merge this information with the annotation columns supplied by the research team. While Twitter's terms of service may reduce the amount of user-contributed data distributed outside of the Twitter network, it would not be appropriate in our case. Distributing TweetIds would associate our evaluated personality traits with a participant's online persona directly and would be a violation of our IRB protocols. Distributing text associated with each TweetId would conform to our IRB, but would not conform to the Twitter Terms of Service. Distributing TweetId would conform to the Twitter Terms of Service, but not to our IRB. As such, we chose not to describe our Twitter data collection in this paper.

5.2. Data Collection Issues

Olteanu et al. (2019) catalogue some of the many biases and pitfalls commonly found in social media collection and analysis. Even cursory inspections of a dataset will reveal issues that could interfere with certain objectives if not properly addressed. This dataset is no exception. In prior work using the VK portion of the dataset for inference of Big Five traits (Hull et al., 2021), we found large numbers of repeated posts apparently machine-generated. For that study we excluded such posts and performed other selection

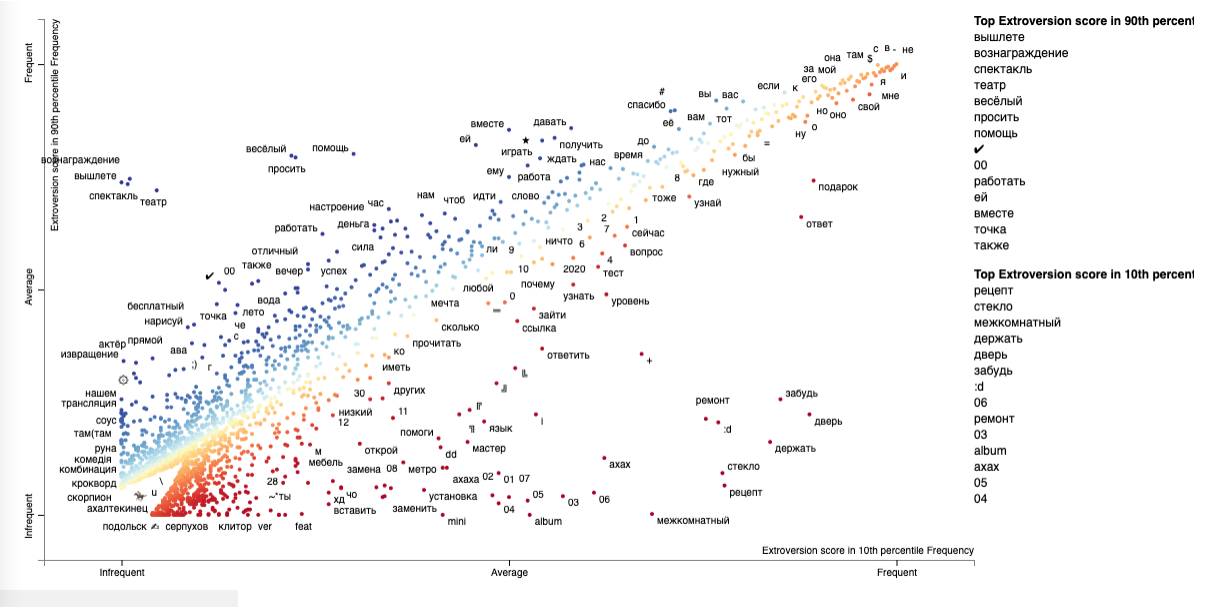

Figure 3: Comparison of term frequencies used by participants in the bottom decile (x-axis) vs. the top decile (y-axis) along the Extraversion dimension of the BFI-2. Visualization created using the Scattertext package (Kessler, 2017)

and post-processing deemed necessary for the integrity of that study. However, as future studies may have different data cleaning needs, the RU-ADEPT dataset leaves such decisions to those who use it. Other than the necessary masking for de-identification, pre-processing is left to the researchers. Similar caveats apply to the survey data: Although some efforts were made to exclude careless responses to the surveys, researchers may wish to inspect the answers carefully and do their own additional exclusions as they deem necessary. Finally, we recognize that revising data collection protocols during these processes are not ideal because the datasets could be inconsistent. In this case, we judged that the benefits to these revisions outweighed the risks, but we have marked the dataset by round so that researchers can choose to use one portion or the other as needed (or test for effects of collection time for their study).

6. Conclusion

As mentioned above, a greater diversity of datasets can make for better social science. The RU-ADEPT dataset provides a starting place for studying personality traits of continuing relevance in a language—and at least one social media platform—that has gotten relatively little treatment in the North American or Western European computational social science community. In addition to providing social media text associated with scores for a common standard instrument for the Big Five factors, it likewise provides scores for the Dark Triad of narcissism, Machavellianism, and psychopathy—perhaps the first shared dataset to do so in the Russian language.

As more such datasets are made available in the community, a better understanding of what textual correlates of personality traits are universal and which are dependent on language, platform, or other factors (e.g., time of collection) should emerge. The dataset seeks to balance the utility of such data to researchers with the protection of participants' privacy. As such, this corpus is not available for direct download; we ask participants to submit requests for reuse for ethical review.

In order to obtain a copy of those portions of the dataset needed for non-commercial research, please send a 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 or equivalent.

7. Acknowledgements

The authors thank Ali Bhatti, Dhanvee Ivaturi, Xiuwei Li, Kevin Ngo, Samara Orellana, and Daniel Smolyak for their role in developing the surveys and collecting the corpus; Ewa Golonka for Russian language expertise; Aric Bills and Adam Liter for assistance with anonymization; and Mike Bunting, Tom Conners, Michelle Morrison, and several others for other types of assistance in the data collection. We also thank anonymous reviewers whose reviews improved the presentation of the work.

Bibliographical References

Balakrishnan, V., Khan, S., Fernandez, T., and Arabnia, H. R. (2019). Cyberbullying detection on twitter using big five and dark triad features. *Personality and Individual Differences*, 141:252–257.

Bender, E. M. (2011). On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6(3):1–26.

Bogolyubova, O., Panicheva, P., Tikhonov, R., Ivanov, V., and Ledovaya, Y. (2018). Dark personalities on facebook: Harmful online behaviors and language. *Computers in human Behavior*, 78:151–159.

Eder, E., Krieg-Holz, U., and Hahn, U. (2019). De-identification of emails: Pseudonymizing privacy-sensitive data in a german email corpus. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 259–269.

Eder, E., Krieg-Holz, U., and Hahn, U. (2020). Code alltag 2.0—a pseudonymized german-language email corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4466–4477.

Egorova, M. S., Parshikova, O. V., and Sitnikova, M. A. (2016). The structure of the short dark triad questionnaire on russian population. *Personality and Individual Differences*, 100(101):475–476.

Farnadi, G., Sitaraman, G., Sushmita, S., Celli, F., Kosinski, M., Stillwell, D., Davalos, S., Moens, M.-F., and De Cock, M. (2016). Computational personality recognition in social media. *User modeling and user-adapted interaction*, 26(2-3):109–142.

Golbeck, J., Robles, C., and Turner, K. (2011). Predicting personality with social media. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '11, pages 253–262, Vancouver, BC, Canada, May. Association for Computing Machinery.

Hull, J. R., Novak, V., Rytting, C. A., Rodrigues, P., Frank, V. M., and Swahn, M. (2021). Personality trait identification using the Russian feature extraction toolkit. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 583–592, Held Online, September. INCOMA Ltd.

Ignatiev, N., Stankeviĉ, M. A., Kisel'nikova, N., and Grigoriev, O. (2019). Opredelenie liĉnostn'yx ĉert u pol'zovatelej VKontakte na osnove analiza izobraẑenij [determination of personality traits of VKontakte users based on image analysis]. *Iskusstvennij Intellekt i Prinyatie Reŝenij [Artificial Intelligence and Decision Making]*, (4):29–36.

Ignatiev, N. A., Stankevich, M. A., Smirnov, I. V., Kiselnikova, N. V., and Grigoriev, O. G. (2020). Predicting personal traits from vkontakte images. *Scientific and Technical Information Processing*, 47(6):383–388.

Jones, D. N. and Paulhus, D. L. (2014). Introducing the short dark triad (sd3) a brief measure of dark personality traits. *Assessment*, 21(1):28–41.

Kessler, J. (2017). Scattertext: a browser-based tool for visualizing how corpora differ. In *Proceedings of ACL 2017, System Demonstrations*, pages 85–90.

Lane, J. and Shipp, S. (2008). Using a remote access data enclave for data dissemination. *International Journal of Digital Curation*, 2(1).

Liang, H. and Fu, K.-w. (2015). Testing propositions derived from twitter studies: Generalization and replication in computational social science. *PloS one*, 10(8):e0134270.

Litvinova, T. and Ryzhkova, E. (2018). RusNeuroPsych: Open Corpus for Study Relations between Author Demographic, Personality Traits, Lateral Preferences and Affect in Text. *International Journal of Open Information Technologies*, 6(3):32–36.

Litvinova, T., Litvinlova, O., Zagorovskaya, O., Seredin, P., Sboev, A., and Romanchenko, O. (2016). "Ruspersonality": A Russian corpus for authorship profiling and deception detection. In *2016 International FRUCT Conference on Intelligence, Social Media and Web (ISMW FRUCT)*, pages 1–7, August.

Litvinova, T. (2020). RusIdiolect: A new resource for authorship studies. In Tatiana Antipova, editor, *International Conference on Comprehensible Science*, pages 14–23. Springer.

MacAvaney, S., Mittu, A., Coppersmith, G., Leintz, J., and Resnik, P. (2021). Community-level research on suicidality prediction in a secure environment: Overview of the CLPsych 2021 shared task. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 70–80, Online, June. Association for Computational Linguistics.

McCrae, R. R. and John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215.

Mohammad, S. M. and Turney, P. D. (2013). Crowdsourcing a Word–Emotion Association Lexicon. *Computational Intelligence*, 29(3):436–465.

Moor, L. and Anderson, J. R. (2019). A systematic literature review of the relationship between dark personality traits and antisocial online behaviours. *Personality and Individual Differences*, 144:40–55.

Moskvichev, A., Dubova, M., Menshov, S., and Filchenkov, A. (2017). Using linguistic activity in social networks to predict and interpret dark psychological traits. In *Conference on Artificial Intelligence and Natural Language*, pages 16–26. Springer.

Olteanu, A., Castillo, C., Diaz, F., and Kıcıman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13.

Paulhus, D. L. and Williams, K. M. (2002). The dark triad of personality: Narcissism, machiavellianism, and psychopathy. *Journal of research in personality*, 36(6):556–563.

Rangel Pardo, F. M., Celli, F., Rosso, P., Potthast, M., Stein, B., and Daelemans, W. (2015). Overview of the 3rd Author Profiling Task at PAN 2015. In *CLEF 2015 Evaluation Labs and Workshop Working Notes Papers*, pages 1–8.

Shchebetenko, S., Kalugin, A. Y., Mishkevich, A. M., Soto, C. J., and John, O. P. (2020). Measurement invariance and sex and age differences of the big five inventory–2: Evidence from the russian version. *Assessment*, 27(3):472–486.

Soto, C. J. and John, O. P. (2017). The next big five inventory (bfi-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of personality and social psychology*, 113(1):117.

Stankevich, M., Smirnov, I., Ignatiev, N., Grigoryev, O., and Kiselnikova, N. (2018). Analysis of big five personality traits by processing of social media users activity features. In *DAMDID/RCDL*, pages 162–166.

Stankevich, M., Ignatiev, N., Smirnov, I., and Kiselnikova, N. (2019). Personality traits prediction from vkontakte social media. *Voprosy kiberbezopasnosti*, (4(32)):80–87.

Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Eighth international AAAI conference on weblogs and social media*.