Personality Trait Identification Using the Russian Feature Extraction Toolkit

James R. Hull[*], **Valerie Novak**[*], **C. Anton Rytting**[*],
Paul Rodrigues[†,*], **Victor M. Frank**[*], **and Matthew Swahn**[†]

[*]University of Maryland, College Park, MD, USA
[†]Accenture, Washington, DC, USA
{jhull1, vnovak, crytting, prr, vmfrank}@umd.edu, matthew.swahn@accenture.com

Abstract

Feature engineering is an important step in classical NLP pipelines, but machine learning engineers may not be aware of the signals to look for when processing foreign language text. The Russian Feature Extraction Toolkit (RFET) is a collection of feature extraction libraries bundled for ease of use by engineers who do not speak Russian. RFET's current feature set includes features applicable to social media genres of text and to computational social science tasks. We demonstrate the effectiveness of the tool by using it in a personality trait identification task. We compare the performance of Support Vector Machines (SVMs) trained with and without the features provided by RFET; we also compare it to a SVM with neural embedding features generated by Sentence-BERT.

1 Introduction

Data scientists and computer scientists working on natural language processing (NLP) problems are occasionally confronted with tasks in languages they do not know well. Technical linguistic issues that may be well known to researchers and language practitioners familiar with the language can be quite important to effective language engineering, and can take a lengthy career to master.

The Russian Feature Extraction Toolkit (RFET) unites a suite of tools for Russian language processing that would allow a programmer unfamiliar with the language the ability to get started quickly in common text classification tasks such as sociolinguistic factor classification, sentiment classification, and emotion analysis. Additionally, this toolkit provides those advanced in Russian NLP convenient accessibility of features and feature combinations to quickly iterate through multiple experiment scenarios.

Deep learning approaches are showing great promise, and feature extraction is less commonly utilized in these approaches. Still, there are existing systems in production which harness classical machine learning techniques. RFET, and tools like it for other languages, could improve performance of these systems without a complete redesign. Further, there are many languages in which deep learning approaches still lack pre-trained models or even a sufficient quantity of text examples to create them. The methodology presented in this paper could be used to design Feature Extraction Toolkits for these languages, in these cases.

This toolkit does not replace feature extraction functions in libraries like NLTK (Bird, 2006) or scikit-learn (Pedregosa et al., 2011). These systems can produce language-independent statistical measurements on text that could be used as features. It also does not seek to replace toolkits for basic NLP pipeline elements, such Natasha,[1] though there is some overlap between the tools. This toolkit focuses on uniting Russian language resources that provide linguistically-informed features that are distinctive of the Russian language or otherwise not adequately represented in features used in text analytics or corpus linguistics, with a focus on social media Russian. The toolkit includes 70 features.

In order to validate the toolkit's utility, we evaluate its efficacy in a challenging task with little prior work in Russian: namely, the identification of personality traits from user-generated social media text. This task is further described in Section 2.2.

[1] https://github.com/natasha/natasha

Proceedings of Recent Advances in Natural Language Processing, pages 583–592
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_066

Section 3 describes RFET's collected features in more detail; Section 4 describes the dataset and other methodological details of the task and our results. Other sections outline larger context and limitations of the study.

2 Related Work

2.1 Language-specific Feature Extraction

Structured Programming for Linguistic Cue Extraction (SPLICE) (Moffitt et al., 2012) is a tool for English language feature extraction. SPLICE offers (via API[2]) a variety of lexicons and other features relevant to credibility assessment and deception detection, including lexicons for deference, positive and negative self-evaluation, affect and sentiment (from SentiWordNet). SPLICE's features include part of speech, verb tense and passive voice (for immediacy), spoken word counts for hedging and disfluency, and a variety of readability scores. It offers a mechanism for users to submit their own lexicons.

The Arabic Data Science Toolkit (ADST) (Rodrigues et al., 2018) is a Python toolkit for analyzing Arabic, particularly social media Arabic. It addresses features both in Modern Standard Arabic and Egyptian Colloquial Arabic, and focuses on features that highlight emotion, such as laughter and emoji, and informal expressions of intensity, such as elongated words. It also includes several language-specific lexicons, such as honorifics, polite and pious expressions, abusive language, and transitional phrases. Its coverage of emotional language is somewhat incomplete, focusing more on positive emotions such as happiness and humor than negative emotions.

Linguistic Inquiry and Word Count (LIWC) (Pennebaker, 1993; Pennebaker et al., 2007, 2015) is a collection of vetted lexicons focused mainly on lexical features of psychological interest, though it also covers more general linguistic categories as well. All categories are vetted multiple times by human judges. The 2015 version has been updated and expanded to include "netspeak" language found in social media and SMS, including some common informal abbreviations and emoticons. Developed originally for English, it has been ported to 12 languages, including Russian. It is available only under paid license (even for non-commercial academic use) and (for Russian) only as a stand-alone program. The API only supports English.

Another tool, Empath (Fast et al., 2016), might be characterized as a partially automated extension of LIWC, with a framework for further extensibility. Empath uses a combination of human-generated seed words, semantic embedding-based term discovery to grow topic lexicons (categories) from these seed terms, and crowd-powered filtering to validate these categories. Unlike most embedding-based models, it is trained largely on fiction works in the public domain, which are claimed to offer more general coverage than other domains. It offers many more topics and categories than LIWC, but novel categories are largely unvetted. The pre-validated models currently available are (to the best of our knowledge) only available in English.

Natasha and DeepPavlov[3] (Burtsev et al., 2018) are NLP pipeline toolkits specifically built for Russian. They bring together NLP tools such as word token and sentence segmentation, word embeddings, morphological syntactic tools, and NER. Natasha adds fact extraction; DeepPavlov various conversational agent functions. As general purpose tools, they complement RFET, which has a more specific focus on social media text.

2.2 Personality Trait Identification

While several personality taxonomies exist, the most well researched set of personality traits is called the Five Factor Model, Big Five, or OCEAN (Openness, Conscientiousness, Extroversion, Agreeableness, and Negative Emotionality or "Neuroticism"), which is typically measured via a 60-item self-report questionnaire, the Big Five Inventory-2 (BFI-2) (Soto and John, 2017). The BFI-2 has been translated into Russian and validated with samples of students and internet users (Shchebetenko et al., 2020).

Golbeck et al. (2011a,b) were among the first to examine the efficacy of inferring personality from user-generated social media text (from Twitter and Facebook). Since that time, a number of studies have followed suit. Farnadi

[2] http://splice.cmi.arizona.edu/

[3] https://deeppavlov.ai/

et al. (2016) conducted a comparative analysis of computational methods on English social media text from three separate platforms: Facebook, YouTube, and Twitter. While most work in social media personality trait identification has focused on English, other European languages have also been examined. For example, the PAN research group organized a shared task in 2015 with Twitter data from four languages: English, Spanish, Italian, and Dutch (Rangel Pardo et al., 2015).

Only a few studies have attempted personality trait inference from Russian social media text. Stankevich et al. (2018) learned a three-way (low, medium, high) classification of the five factors from a dataset of 165 VKontakte profiles. However, due to the sparseness of usable, user-generated text in their dataset, they were unable to use lexical features, but only very basic features on the text (such as average numbers of words and sentences, use of punctuation and uppercase). Their reported F1 scores range from 36% for Conscientiousness to 53% for Agreeableness.

Similarly, Ignatiev et al. (2019) used both SVM and Random Forest approaches for a two-way (highest quartile, lowest quartile) classification of five factors traits from their dataset of 1,020 VKontakte profiles. They used lexical features, an aggression lexicon, user profile information, and a repost matrix, and reported F1 scores ranging from 61.75% on Openness to Experience to 73.75% on Extroversion.

Litvinova et al. (2015; 2016) and Vybornova et al. (2011) infer personality traits from other genres of Russian text (besides social media such as VKontakte). They propose several and test linguistic correlates to personality, such as content/function words, readability indices, lexical diversity and usage of first-person singular pronouns.

3 RFET Features

3.1 Morphological

Morphological features are generated using PyMorphy2, a Russian morphological analyzer and inflection engine (Korobov, 2015).[4] The morphological analyzer is able to detect the following morphological features: part of speech, animacy, verbal aspect, dictionary citation forms, case, gender, involvement, mood, number, person, tense, transitivity and voice. Because PyMorphy2 evaluates on the token level, without looking at the context, syntax level features are not captured by the toolkit.

The toolkit leverages the output of PyMorphy2 to calculate the frequencies of these features within the text, but we added extensions to produce additional linguistic features for RFET. We diverge from PyMorphy2 where the tool conflates logically different phenomena. For example, PyMorphy2 determines gender of nouns and adjectives by word ending; however, some borrowed nouns and acronyms are indeclinable and thus may not show any cues for gender. Not all such nouns are neuter, so assigning a default gender risks inaccuracies (Wang, 2014). Inevitably, some lexical items will be missing from PyMorphy2's lexicons–and thus have undetermined gender so far as PyMorphy2 is concerned.

Additionally, not all Russian parts of speech inflect for gender. The toolkit expands the tag set from 4 tags (masculine, feminine, neuter, null) to 5 tags (masculine, feminine, neuter, undetermined, non-gendered-pos) to differentiate between those words of specific parts of speech that are not gendered in Russian (i.e. conjunctions, comparatives, gerunds, adverbs, particle, infinitives, prepositions and predicatives) with those parts of speech that can have gender, but are not determined by PyMorphy2.

RFET also utilizes PyMorphy2's OpenCorpora Dictionary (Открытый корпус), to identify the ratio of words in the text that are not found in its lexicon. This ratio may indicate usage of new words, slang, typos, URLs and other Out of Vocabulary (OOV) items. The ratio of OOV items can be a useful feature for classification to sociolinguistic targets.

3.2 Laughter, Emoticons and Emoji

While laughter appears in informal text across languages, the characters and patterns used

[4] https://github.com/kmike/pymorphy2. At the time of RFET's design, PyMorphy2 seemed to be the most widely cited specifically Russian morphological parser and was integrated into DeepPavlov. UDpipe's Russian lemmatizer may have been a reasonable alternate choice. A Russian lemmatizer for SpaCy was not available until February 2021, as documented in the SpaCy blog. (https://explosion.ai/blog/spacy-v3). A comparison of lemmatizers is out of scope for this paper.

to represent laughter differ. RFET returns frequency features on Russian-specific laughter and emoticons found on social media, as well as language-agnostic emoji features. These features could be utilized in an author attribution system, sentiment analysis, or emotion classification system.

RFET returns frequency information on type, number of times used and length of laughter or emoticon within the text. RFET tracks the following types of laughter (in *featurename* (example) pairs) : *haa* (хааа), *haah* (хахаааах), *haahaa* (хаахаа), *ha-ha* (ха-ха), *hehe* (хехxe-ее), *hihi* (хихихии), *hi-hi* (хи-хи), *Lol* (лол), *Lolol* (лолол), *phaha* (пхаха), *HAHA* (ХА-ХА), *hoho* (хохо), *HIHI* (ХИХИ), *HEHE* (ХЕ-ХЕ), *HOHO* (ХОХО), *HA-HA* (ХА-ХА), *HI-HI* (ХИ-ХИ), as well as Russian happy face parenthesis (“))))”),[5] and sad face parenthesis (“(((”). These are implemented as templated regular expressions that match and report the length of variants.

RFET also has a feature which returns the frequency of emoji usage and leverages Cal Henderson’s emoji-data package.[6]

3.3 Sentiment and Emotion Features

Sentiment analysis is one of the most popular commercial applications of NLP, and RFET makes it easier for a nonnative speaker to implement a system in Russian. Currently RFET utilizes one emotion/sentiment lexicon and can easily be extended to allow for others.

The NRC Emotion Lexicon (Mohammad and Turney, 2013), also known as EmoLex, is a lexicon translated into 104 languages. Each lexical entry is coded for Positive and Negative sentiment and the emotions Fear, Anger, Sadness, Joy, Disgust, Surprise, Trust and Anticipation.[7] RFET reports the emotions present in the text by using PyMorphy2 to resolve the dictionary citation form of each token in the text and reporting the emotion(s) and sentiment orientation associated with that form in the lexicon, if available. If the token is present in the lexicon but is not coded as positive or negative, the token is coded as neutral. Similarly, RFET reports whether a token is not found in the lexicon. RFET uses these resources to return a dictionary of sentiment (positive, negative, neutral) and emotion (fear, sadness, joy, disgust, surprise, trust, and anticipation) or “token_not_in_lexicon” and the count of tokens representing these.

3.4 Lexical Diversity

According to Litvinova et al. (2016, 2017), a lack of lexical diversity was associated with individuals with a greater likelihood of self-destructive behavior, which may be useful to author profiling or personality identification systems.

Litvinova et al. (2017) described lexical diversity through a variety of features, including type to token ratio, an index of formality, an index of lexical density, the ratio of function words (particles, prepositions, conjunctions, etc) to total tokens, the ratio of content words (nouns, verbs, infinitives, adjectives, adverbs, etc.) to total tokens, the ratio of personal pronouns to total tokens, and the proportions of the 100 most frequent Russian words in the document to all tokens.

RFET implements these key lexical diversity features replicating Litvinova’s descriptions; in addition to extracting the proportion of the top 100 most frequent Russian words (unigrams), it also tracks usage of the top most frequent bigrams, trigrams, and 4-, 5-, and 6-grams in the Russian National Corpus (RNC).[8]

3.5 Other Lexical Features

Other Russian specific features that RFET extracts and quantifies are punctuation, digits, diacritics, other languages, and other scripts. The punctuation, diacritics, and quotation scripts include Russian specific unicode characters (i.e. «, » , „ and “) in addition to the punctuation that is shared across languages. The punctuation feature does overlap with the emoticon features.

[5] Garber, M. (2013, July 20). 55555, or, How to Laugh Online in Other Languages. Retrieved July 09, 2020, from https://www.theatlantic.com/technology/archive/2012 /12/55555-or-how-to-laugh-online-in-other-languages/266175/

Why do Russians use ’)’ as a smiley instead of ’:)’? - Quora. (n.d.). Retrieved July 9, 2020, from https://www.quora.com/Why-do-Russians-use-as-a-smiley-instead-of

[6] https://github.com/iamcal/emoji-data

[7] https://saifmohammad.com/WebDocs/Mohammad-Turney-NAACL10-EmotionWorkshop.pdf

[8] *N*-grams are found at e.g., https://ruscorpora.ru/old/1grams.top.html.

The ratio of code switching or borrowings from Western languages may be indicated by the ratio of Latin characters to total characters in the text, and this is reported by RFET as an additional feature. Another feature utilizes a language identification package[9] to determine the language of the text, in order to identify instances of non-Russian text within a corpus of presumed Russian documents, including languages such as Bulgarian, Macedonian, and Ukrainian that also use Cyrillic.

4 Inferring Personality Traits

4.1 Dataset

The dataset for evaluating the personality trait inference task was collected by the authors during 2020. It consists of 149K VKontakte posts from 288 consenting participants, with a total of 3.8M word tokens. (This corpus was filtered from a larger collection by excluding posts containing URLs, ASCII art, duplicate posts, and posts that appeared to be auto-generated. It includes only those participants with at least 1200 tokens in their VK posts after this filtering process.) Each of the 288 participants took the Russian version of the BFI-2 inventory (Soto and John, 2017; Shchebetenko et al., 2020). The labels (personality trait scores) were rescaled from the raw inventory scores to the interval $[-5, 5]$. The dataset was partitioned by author into train and test sets (with 80% of author accounts comprising 79% of the total word tokens in train and the rest in test).

4.2 Baselines

We created two sets of baseline models for personality ID, one using classical machine learning methods, and one using neural embeddings.

The first set of baseline models created for the five personality traits used a standard bag-of-words approach using term frequency–inverse document frequency ($tf{*}idf$) and included language-independent features of "lexical richness" such as Yule's (2014) K, Sichel's (1975) S, Honoré's (1979) Measure, Brunet's (1978) Measure, Maas's (1972) a^2, and Rubet's k (Dugast, 1979). None of these features require specific knowledge of the target language. Because of the size of the $tf{*}idf$ data, principal component analysis (Jolliffe and Cadima, 2016) was used to reduce the number of features. In total, 73 features were kept from the ($tf{*}idf$) data, which corresponded to keeping 85% of the total variance, as well as the six lexical richness features.

The second set of baseline models created for the five personality traits used Sentence-BERT (SBERT) (Reimers and Gurevych, 2019) in combination with Russian-specific language model Sentence RuBERT (rubert-base-cased-sentence) [10], a fine-tuned version of RuBERT (Kuratov and Arkhipov, 2019) for sentence embedding. Sentence-BERT produces fixed-sized embeddings by pooling the output of a language model. These fixed-sized embeddings are convenient to use in combination with classification algorithms for sentence classification tasks. We used Sentence-BERT's highest performing pooling strategy to produce these embeddings by taking the mean of all output language vectors and Sentence-BERT's default maximum sequence length of 128 WordPieces.[11]

4.3 Method

We used a standard implementation of a support vector machine (SVM) to train personality identification regression models from a corpus of Russian VKontakte text labeled for Big Five personality traits, varying the features supplied to each system. Each feature set we utilized was used to train and test against the five traits-Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Negative Emotionality (N) traits. A linear kernel was used for all traits, and the regularization parameter (C) was tuned separately for each trait.

The target data are scaled to the interval $[-5, 5]$ for each trait, and we report the root

[9] https://pypi.org/project/langdetect/, which is a Python port of Shuyo Nakatani's (2010-2014) language-detection Java language ID package.

[10] https://huggingface.co/DeepPavlov/rubert-base-cased-sentence

[11] Preliminary experiments suggested that using a maximum sequence length of 512 WordPieces did not meaningfully improve performance accuracy, while adding significant time requirements. The average post in our training corpus, when tokenized to the rubert-base-cased-sentence lexicon, was 50 WordPieces. There were 4719 posts longer than 128 WordPieces, about 4.5% of our training corpus, which means with a maxlength of 128, only 4.5% of our samples were clipped.

mean square error (RMSE). Across all experiments, the same training and testing splits were used on the data.

Our goal is to predict all five personality traits, so this problem can be treated as a multi-target regression problem. Since there is evidence in our dataset and other literature that personality traits are correlated (Gosling et al., 2003), the target trait of one model may be useful as a feature when predicting another trait. Because of this, predicted values for some traits were fed back in as features for other models, called "stacked single target" (SST) chains (Spyromitros-Xioufis et al., 2016). Similar work has been done for other multi-target regression problems with positive results (Melki et al., 2017). It is possible for this process to be optimized, standardized, or generally improved. We look to investigate this in future work. For the experiments reported in Table 1, the trait that was predicted the most accurately was the one chosen as a feature for the next round of models. Each trait was added in as a feature exactly one time, such that the final database consisted of all the raw data, plus five new features (one for each trait). Thus each of the five traits is predicted using the stacked single targets (i.e., trait value predictions from previous iterations of the model) for the four other traits as features.

4.4 Results

Table 1 shows root mean squared error (RMSE) (lower is better) and R^2 (higher is better) for our bag of words classifier with and without the RFET features. Each post was treated as a unique entry, and the data was fed into a support vector regression model. To prevent overfitting, both the test-train split and cross validation splits were done on the participant level. The model never saw any data from any participants in test or validation sets. The inclusion of RFET features produced a more accurate model based on both RMSE and R^2 values. We believe these features add significant value to modeling personality traits and likely other tasks as well.

Our Sentence-BERT neural baseline results for post-level predictions of a post author's personality traits, as well as comparable predictions for the BoW+RFET model, can be found in Table 2. Unlike our SVM baseline and SVM with RFET features results in Table 1, neither Sentence-BERT nor the SVM with RFET features shown here utilize SST as input to help improve the performance. We found that the SVM with SST and SVM with RFET features (with or without SST) outperformed the Sentence-BERT neural baseline.

Features	BoW w/o RFET		BoW+RFET	
Trait	RMSE	R^2	RMSE	R^2
O	1.67	0.18	1.43	0.38
C	1.76	0.22	1.47	0.46
E	1.79	0.18	1.58	0.37
A	1.54	0.28	1.46	0.37
N	1.89	0.20	1.67	0.38

Table 1: Psychological Trait Regression predicting per post with and without RFET features. Both versions predict traits iteratively, using SST for the other four traits as features. Values shown are Root Mean Squared Error (RMSE) and R^2 Scores.

Features	SBERT only	BoW+RFET
Trait	RMSE	RSME
O	2.49	2.04
C	2.83	1.92
E	2.53	2.06
A	3.03	1.68
N	2.59	1.65

Table 2: Psychological Trait Regression on SBERT Neural Baseline and bag-of-words + RFET, predicting per post. Unlike Table 1, both versions predict all five traits independently (in parallel). Values shown are Root Mean Squared Error (RMSE).

Comparing the two BoW+RFET columns in Tables 1 and 2, we can see that the SST chains do improve the BoW+RFET model's accuracy for all features but Negative Emotionality, but of course this comes at the cost of speed, as the five traits can no longer be trained or decoded in parallel (and training requires several more iterations).

5 Discussion

For the personality trait identification task on this social media dataset, we see that a classic SVM baseline using RFET features outperforms a similar SVM without the RFET features. We likewise see that the SVM with RFET features outperforms a model using a transformer model pre-trained on Russian text.

One trait where the advantage of the RFET features is particularly large is the Agreeableness trait. One possible explanation for this may be differences in participants' use of emoji, emoticons, and/or emotional words correlating with their Agreeableness trait values. RFET includes features specifically developed for emoji, emoticons, and emotional words, and even the SVM BoW model may be somewhat sensitive to them; the pretrained model, on the other hand, may be ignoring emoji and emoticons, since such "words" may not have appeared in its original training data.

Another advantage the standard machine learning methods have over neural models is interpretability. Since a SVM was used with a linear kernel, it is possible to extract feature importance from each model, and gain insights into where the strongest correlations lie. To do this, we employed the R package *e1071* (Meyer et al., 2021). Every model is trained independently, and so produces different feature weights, but in general, these features appeared as the most important for our machine learning models:

- content, function word to token ratios;
- NRC emotional lexicon tokens;
- number frequency (singular vs. plural)
- frequency of morphological features: grammatical number, animacy, case, verbal mood
- frequencies of top 100 (RNC) unigrams

6 Limitations and Future Work

Neural NLP models that follow the BERT architecture, like Sentence RuBERT, grow in memory requirements quadratic to the sequence length. Because of this, the models are limited to a sequence length (often 512 WordPieces, but 128 here) with the remaining WordPieces in a post ignored and left unprocessed. VK posts longer than 128 WordPieces are clipped in our Sentence-BERT experiments, while the full posts serve as input to the SVM bag of words and RFET feature extraction systems. Model architectures have been released catering to long form English text, such as Big Bird (Zaheer et al., 2020) and Longformer (Beltagy et al., 2020), but models using these architectures are not yet available for Russian. Lastly, RuBERT was trained on Wikipedia and books. Our experiments were on social media text, which RFET was designed to address. This genre mismatch (and RuBERT's resulting limited coverage for WordPieces specific to social media) may have limited Sentence RuBERT's effectiveness for this task.

One possible weakness in the RFET model (as trained on these training data) is sparsity of certain features. Certain RFET features may be overfitting to particular participants. Idiosyncratic uses of words from a few people with unusual personality trait values or a large volume of posts may be inappropriately generalized as signals for those trait values.

One possible mitigation, of course, is the collection of data from a much larger set of individuals. In these experiments, we apply a more local mitigation strategy: combining different kinds of laughter into a single feature. One strength of RFET is the flexibility of combining or splitting apart features according to the need of the task. For example, personality trait (or other attribute) detection may benefit from combining features; author identification or verification may benefit from keeping very specific features split apart.

Future feature extractors include frequency of Russian diminutives through the usage of suffixes and infixes, ratio of other language scripts being used in the text and usage and frequency of filler words and phrases. Additional information about sentiment and prevalence of emojis will be incorporated from the Emoji Sentiment Ranking (Kralj Novak et al., 2015),[12] which provides usage statistics and associated sentiment for each emoji. An expanded list of emoticons will also be added to reflect the frequency and usage of more emoticons.

Future iterations of RFET may utilize RuSentiLex (Loukachevitch and Levchik, 2016), a Russian emotion lexicon with 16,057 entries. Each entry includes a description of its syntactic category, a lemmatized version, the sentiment valence, and source of the valence (opinion, feeling, or fact). Ambiguous entries in RuSentiLex (with more than one possible value for valence or source) are also elaborated with examples. We anticipate these features will provide better coverage of sentiment than the

[12] http://kt.ijs.si/data/Emoji_sentiment_ranking/

Russian translation of the EmoLex dictionary.

7 Conclusion

This paper introduces the Russian Feature Extraction Toolkit, an API for feature extraction on the Russian language. Each feature in the toolkit utilizes linguistic knowledge of the Russian language. It is designed to get a Russian non-speaker up and running quickly on Russian NLP tasks, and to speed up the workflow of Russian speaking NLP programmers. We have shown that it improves performance on a Big Five personality trait inference task relative to a SVM baseline with only language-independent features and, more surprisingly, to a pre-trained transformer baseline using Sentence RuBERT. This suggests that RFET's features specific to social media can be very useful for enhancing state-of-the-art methods for certain genres or domains.

Licensing: We plan to release RFET for non-commercial research and education. A public API will be made available for demonstration purposes. For commercial licenses, contact the University of Maryland's Office of Technology Commercialization.[13]

Ethical Considerations

No novel data collection was done specifically for developing RFET; RFET features depend on pre-collected corpora such as the Russian National Corpus (RNC). Data collection for the evaluation of RFET's efficacy for personality trait estimation (as described in section 4.1) was conducted with the approval of the University of Maryland Institutional Review Board (IRB). During the consent process, potential respondents were informed of the research purposes, and that the researchers would remove, anonymize, or pseudonymize names of entities in the collected social media data deemed to risk personally identifying the participant prior to any sharing of the data with those outside the IRB protocol. Any additional potential risks to confidentiality have been minimized by keeping all data on a secure Amazon Web Service (AWS) server, to which only authorized researchers with the requisite IRB training have access. Non-anonymized data will be destroyed upon the close of the IRB protocol; only data which has undergone our de-identification process would be retained.

Although we set collection targets by gender and age bracket to obtain representative samples to encourage greater equity of representation by gender and age, the use case evaluation was naturally biased towards individuals who write and post lots of text on social media (for which women and younger writers were over-represented in our sample). An evaluation which used equal amounts of text for each individual would avoid this bias, at the cost of leaving unused large portions of the corpus.

Since RFET is a toolkit to assist researchers in improving their own NLP applications, the primary beneficiaries are NLP researchers and developers, particularly those working with social media. Likewise, the main source of potential harm lies with what researchers and developers decide to do with the RFET tool. (Personality trait inference is just one example of potential downstream applications and its own ethics of use depend on where and for what purpose it is applied.) Biases may exist in the older texts used (e.g., some of those in the RNC) but since the features used here are largely based on grammatical categories and lists of keywords, the toolkit is arguably more transparent than tools based on semantic embeddings and such bias easier to identify and address.

Acknowledgments

The authors thank Nadiya Klymenko and Ali Bhatti for work on early versions of RFET, Ewa Golonka for Russian language expertise, Jarrett Lee for technical support, and Mike Bunting, Laurel Miller-Sims, Tom Conners, Michelle Morrison, Aric Bills, and several others for assistance in the data collection. We also thank anonymous reviewers whose reviews improved the presentation of the work.

[13] www.otc.umd.edu.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Steven Bird. 2006. NLTK: The Natural Language Toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72, Sydney, Australia. Association for Computational Linguistics.

E. Brunet. 1978. *Vocabulaire de Jean Giraudoux: Structure et Evolution*. Slatkine.

Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, Alexey Litinsky, Varvara Logacheva, Alexey Lymar, Valentin Malykh, Maxim Petrov, Vadim Polulyakh, Leonid Pugachev, Alexey Sorokin, Maria Vikhreva, and Marat Zaynutdinov. 2018. DeepPavlov: Open-source library for dialogue systems. In *Proceedings of ACL 2018, System Demonstrations*, pages 122–127, Melbourne, Australia. Association for Computational Linguistics.

D. Dugast. 1979. *Vocabulaire et stylistique I. Théâtre et dialogue, travaux de linguistique quantitativ*. Slatkine.

Golnoosh Farnadi, Geetha Sitaraman, Shanu Sushmita, Fabio Celli, Michal Kosinski, David Stillwell, Sergio Davalos, Marie-Francine Moens, and Martine De Cock. 2016. Computational personality recognition in social media. *User modeling and user-adapted interaction*, 26(2-3):109–142.

Ethan Fast, Binbin Chen, and Michael S Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4647–4657.

Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. 2011a. Predicting personality from twitter. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pages 149–156. IEEE.

Jennifer Golbeck, Cristina Robles, and Karen Turner. 2011b. 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. Association for Computing Machinery.

Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. 2003. A very brief measure of the big-five personality domains. *Journal of Research in personality*, 37(6):504–528.

A. Honore. 1979. Some simple measures of richness of vocabulary. *Association for Literary and Linguistic Computing Bulletin*, (7):172–177.

N.A. Ignatiev, Maksim Alekseevi$\hat{c}$ Stankevi$\hat{c}$, N.V. Kisel'nikova, and O.G. Grigoriev. 2019. Opredelenie li$\hat{c}$nostn'yx $\hat{c}$ert u pol'zovatelej VKontakte na osnove analiza izobra$\hat{z}$enij [determination of personality traits of VKontakte users based on image analysis]. *Iskusstvennij Intellekt i Prinyatie Re$\hat{s}$enij [Artificial Intelligence and Decision Making]*, (4):29–36.

I. T. Jolliffe and J. Cadima. 2016. Principal component analysis: a review and recent developments. *Philos Trans A Math Phys Eng Sci*, (20150202).

Mikhail Korobov. 2015. Morphological analyzer and generator for Russian and Ukrainian languages. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 320–332. Springer.

Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. 2015. Sentiment of emojis. *PLoS ONE*, 10(12):e0144296.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language.

Tatiana Litvinova, Olga Litvinlova, Olga Zagorovskaya, Pavel Seredin, Aleksandr Sboev, and Olga Romanchenko. 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.

Tatiana Litvinova, Pavel Seredin, Olga Litvinova, and Olga Zagorovskaya. 2017. Differences in type-token ratio and part-of-speech frequencies in male and female Russian written texts. In *Proceedings of the Workshop on Stylistic Variation*, pages 69–73, Copenhagen, Denmark. Association for Computational Linguistics.

Tatiana Litvinova, P.V. Seredin, and O.A. Litvinova. 2015. Using Part-of-Speech Sequences Frequencies in a Text to Predict Author Personality: a Corpus Study. *Indian Journal of Science and Technology*, 8.

Natalia Loukachevitch and Anatolii Levchik. 2016. Creating a General Russian Sentiment Lexicon. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1171–1176, Portorož, Slovenia. European Language Resources Association (ELRA).

H. D. Maas. 1972. Zusammenhang zwischen wortschatzumfang und länge eines textes. *Zeitschrift für Literaturwissenschaft und Linguistik*, 8(2):73–79.

Gabriella Melki, Alberto Cano, Vojislav Kecman, and Sebastián Ventura. 2017. Multi-target support vector regression via correlation regressor chains. *Information Sciences*, 415:53–69.

David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. 2021. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.7-6.

K.C. Moffitt, J.S. Giboney, E. Ehrhardt, J.K. Burgoon, and J.F. Nunamaker. 2012. Structured programming for linguistic cue extraction (SPLICE). In *Proceedings of the HICSS-45 Rapid Screening Technologies, Deception Detection, and Credibility Assessment Symposium*, pages 103–108.

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

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

James W Pennebaker. 1993. Putting stress into words: Health, linguistic, and therapeutic implications. *Behaviour research and therapy*, 31(6):539–548.

James W Pennebaker, Roger J Booth, and Martha E Francis. 2007. Linguistic inquiry and word count: Liwc [computer software]. *Austin, TX: liwc. net*, 135.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report, University of Texas at Austin, Austin, TX.

Francisco Manuel Rangel Pardo, Fabio Celli, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. 2015. Overview of the 3rd Author Profiling Task at PAN 2015. In *CLEF 2015 Evaluation Labs and Workshop Working Notes Papers*, pages 1–8.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Paul Rodrigues, Valerie Novak, C Anton Rytting, Julie Yelle, and Jennifer Boutz. 2018. Arabic data science toolkit: An api for arabic language feature extraction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

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

H. S. Sichel. 1975. On a distribution law for word frequencies. *Journal of the American Statistical Association*, 70(351):542–547.

Christopher J Soto and Oliver P John. 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.

Eleftherios Spyromitros-Xioufis, Grigorios Tsoumakas, William Groves, and Ioannis Vlahavas. 2016. Multi-target regression via input space expansion: treating targets as inputs. *Machine Learning*, 104(1):55–98.

Maxim Stankevich, Ivan Smirnov, Nikolay Ignatiev, Oleg Grigoryev, and Natalia Kiselnikova. 2018. Analysis of big five personality traits by processing of social media users activity features. In *DAMDID/RCDL*, pages 162–166.

Olga Vybornova, Ivan Smirnov, Ilya Sochenkov, Alexander Kiselyov, Ilya Tikhomirov, Natalya Chudova, Yulia Kuznetsova, and Gennady Osipov. 2011. Social tension detection and intention recognition using natural language semantic analysis: On the material of russian-speaking social networks and web forums. In *2011 European Intelligence and Security Informatics Conference*, pages 277–281. IEEE.

Qiang Wang. 2014. Gender Assignment of Russian Indeclinable Nouns. Master's thesis, University of Oregon, September. Accepted: 2014-09-29T17:49:26Z.

C Udny Yule. 2014. *The statistical study of literary vocabulary*. Cambridge University Press.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33.