Primenjena psihologija
Vol. 16, No. 4, pp. 475-519, 2023

Research Article

Personality Adjectives in Serbian Tweets: An Opening

Petar Čolović [1], Marija Bojanić ✉[1], Anastazia Žunić [2] and Alexandre José de Souza Peres [3]

[1] *Faculty of Philosophy, University of Novi Sad, Serbia*
[2] *Mathematical Institute of the Serbian Academy of Sciences and Arts, Serbia*
[3] *Federal University of Mato Grosso do Sul (UFMS), Brazil*

ABSTRACT

There has been a great interest in investigating relations between personality and language use on the web or social media. Most of the recent studies are based on mining the users' information available online and then using machine learning algorithms to predict their personality characteristics. On the other hand, a few studies relied on the traditional lexical hypothesis when exploring personality under the assumption that personality-related attributes could be obtained from dictionaries. However, little is known about personality structure from Twitter/X - do data strictly reflect personality structure as represented by personality models, or as unique personality semantic patterns. The aim of the study was to assess and interpret the personality adjective-based structure contained in tweets. The data were collected from an open-access „Tweet-sr" Serbian Twitter linguistic corpus (Ljubešić & Klubička, 2014). Latent Dirichlet Allocation, a topic modeling technique, was conducted to extract topics and cosine similarity was used as a measure to determine topic similarities, as well as similarities between the topics and personality dimensions. The results showed that the optimal solution comprised four non-overlapping topics reflecting specific semantic structures. Topics did not replicate trait constructs but were modestly related to them. The largest similarities were found with Extraversion and Agreeableness, pointing out the conceptual importance of these traits when describing interpersonal behavior. Also, no inter-topic differences in word category distributions were found, with the evaluation

terms being the second most frequent in three topics. Although tweets are short-form text messages, they have the potential to communicate socially relevant information through personality descriptors.

Keywords: personality structure, personality descriptors, topic modeling, Twitter.

UDC: 159.923:316.472.4(497.11)
DOI: 10.19090/pp.v16i4.2514
Received: 12.10.2023.
Revised: 07.12.2023.
Accepted: 12.12.2023.

✉ Corresponding email: marija.bojanic@ff.uns.ac.rs

Introduction

In recent years, particularly in the last decade, a growing body of literature has highlighted the importance of exploring relations between natural language and various social, behavioral, and psychological phenomena (Boyd & Pennebaker, 2017; Kosinski et al., 2013; Pennebaker et al., 2003). One of the benefits of using language-based measures in personality research is that data available on the web or social media channels reflect a more realistic representation of personality characteristics from the language people use on a daily basis, compared to self-report measures (Boyd & Pennebaker, 2017). Therefore, how people use words and express themselves online has initiated researchers' interest in finding linguistic phenomena (words and word patterns, such as sentiments and topics) as correlates or functions of personality attributes (i.e., personality traits). This approach is primarily based on mining the users' information gathered on social media platforms and using machine learning algorithms to predict their personalities (i.e., personality prediction framework).

Twitter, personality, and lexical hypothesis

Personality has been a consistent point of interest in natural language processing (NLP) and Twitter-related studies. Personality information has been chiefly derived from the web and social media platforms (e.g., Facebook, Twitter- currently rebranded as X) by employing text-mining techniques (Carducci et al., 2018; Golbeck et al., 2011; Qiu et al., 2012; Quercia et al., 2011; Schwartz et al., 2013; Yarkoni, 2010; Zhao et al., 2020). Personality studies on social media platforms have primarily been based on mining the users' information using machine learning algorithms to predict their personality features. On the other hand, little is known about personality structure within the social media context without relying solely on the predictive modeling paradigm.

Few studies have addressed the issue of personality cues in social media from the perspective of the traditional lexical hypothesis. Assuming that

personality-related attributes (i.e., personality traits) are embedded within natural language and extractable from dictionaries (De Raad & Mlačić, 2020; Goldberg, 1981, 1990, this paradigm has yielded several methodological strategies for gathering personality-relevant words. Influential personality models have stemmed from psycholexical studies, such as the Big Five (Hofstee et al., 1992), Big Six / HEXACO (Lee & Ashton, 2004), Big Seven (Almagor et al., 1995; Benet-Martínez & Waller, 2002), and Cattell's Sixteen personality factors (Cattell & Kline, 1977). The informally termed „Dutch" and „German" methodological frameworks are usually considered „classic approaches" in the field. Both focus primarily on adjectives, though nouns and verbs are steadily gaining more attention from researchers (De Raad et al., 1988; De Raad & Ostendorf, 1996; Henss, 1995; Paulsen, 2011; Saucier, 2003). Both advocate using comprehensive word lists extracted from dictionaries (instead of using descriptor samples). The „Dutch" methodology assumes that a personality descriptor is relevant if it fits in the phrase „I am…" (or „She/he is…", „They are…") and does not pose any additional restrictions regarding word category or function (Hofstee, 1990). German studies were focused on thirteen word categories, based on Warren Norman's English descriptors' fifteen-category classification (Angleitner et al., 1990; Norman, 1967). In the third Serbian psycholexical study (De Raad et al., 2018), nine descriptor categories appeared: temperament and character traits; abilities, talents, or their absence; emotions, moods, and cognitions; states and activities; roles and relationships; social effects – reactions of others; pure evaluation; social status, and value orientations. The first two categories fall into the broader class termed „dispositions, " the following two into „temporary conditions, " and the next four into „social and reputational aspects." The third prominent methodology, proposed by Tellegen and Waller (Almagor et al., 1995), suggests sampling personality-relevant words from dictionaries, imposing no restrictions, and not relying on comprehensive descriptors' lists. This approach has highlighted the importance of evaluative terms, which constitute two personality dimensions – positive valence and negative valence. Out of three psycholexical studies in the Serbian language, two have applied Tellegen and Waller's methodology, yielding results comparable to Big Seven dimensions (Čolović et al., 2014;

Smederevac et al., 2007) but also hinted at the possibility of Big Five replication (Colovic et al., 2005). The third study, utilizing word categories, has demonstrated that the trait descriptor structures change according to the word categories included; dispositional terms result in dimensions similar to the Big Five, while the introduction of evaluative terms leads to solutions comparable to the Big Six or HEXACO (De Raad et al., 2018). Thus the relevance of methodological factors, particularly word categories (Barelds & Raad, 2015) in lexical studies and their impact on the results have once again been demonstrated. At the same time, statistical integration of the results of three Serbian psycholexical studies (De Raad et al., 2018) pointed to five dimensions as, so far, most plausible approximations of top-tier personality dimensions in Serbian language: Agreeableness, Conscientiousness, Extraversion, Negative Valence, and a Neuroticism-related factor.

At the same time, traditional psycholexical studies' have so far almost exclusively utilized the data gathered by self-report or peer-report questionnaires, with a few exceptions across several decades (Cutler & Condon, 2022; Čolović & Filipović Đurđević, 2019; Fischer et al., 2020; Passakos & De Raad, 2009; Oljača et al., 2018; Peres, 2018; Roivainen, 2015b, 2015a). One may argue that the Twitter format is a challenge for personality researchers due to its specific features: brevity, extensive use of colloquial terms and slang, vast diversity of topics, frequent dialogue or polylogue form, richness of production, and others. Due to all these idiosyncracies, one may wonder whether the Twitter form reflects the „known" personality trait structures, as represented by models of personality, or personality-related semantic patterns that we know little of.

Recognizing Twitter as a valuable source of personality information, Peres (2018) has conducted a study on Brazilian Portuguese self-reporting tweets, applying the methodology adherent to traditional lexical hypothesis and using Latent Dirichlet Allocation (LDA, a topic modeling technique) as the primary analytic tool. Brazilian Portuguese adjective list was assembled and used along the descriptors embedded in the Big Five, HEXACO and Cattell's models. Despite the semantic coherence of seven- and fourteen-topic solutions, their contents did not substantially overlap with Big Five, Big Six, or Cattell's model.

Within the seven-topic solution, which was deemed to be one of the two most plausible, three topics were predominantly related to Agreeableness.

Investigating Openness to Experience adjective descriptors as modifiers of person-related nouns in Google Books and Tweets, Roivainen (2015a) pointed to smaller linguistic diversity in Tweets than in books, whereby a small set of terms dominated the Twitter discourse. The same author (Roivainen, 2015b), in a similarly designed but more comprehensive study, emphasized the lack of replicability of established personality models but demonstrated substantial positive correlations regarding the use of personality modifiers of the nouns „man" and „woman" in English and French languages.

Current study

According to the results obtained so far, apparently there is a substantial amount of semantically coherent personality information on Twitter, but it does not appear to straightforwardly represent the structure of personality dimensions from lexically-derived personality models (Peres, 2018; Roivainen, 2015a, 2015b). In a study of the frequencies of Openness adjective markers, Roivainen (2015b) pointed to flawedness in laypersons' personality assessment skills. On the other hand, the prediction of participants' Big Five traits based on Tweets has yielded successful results (Christian et al., 2021; Jaimes Moreno et al., 2019; Kern et al., 2019; Mavis et al., 2021). Hence the crucial question arises: if tweets carry conceptually relevant personality information, whereby it does not seem to reflect lexical personality models directly, how can we assess and interpret the personality descriptor structure contained in them?

We opted to apply traditional psycholexical study methods to self-referencing tweets, adopting elements of both Dutch and German approaches (Angleitner et al., 1990; Hofstee, 1990), as they have been used in the third psycholexical study in the Serbian language (De Raad et al., 2018). An analogous methodological strategy was first employed by Peres (2018), though without using adjective categorization, which did not exist in Brazilian Portuguese at the time. Hence, in a sense, one may regard this study as a tentative replication of the pioneer work in the field (Peres, 2018), though in a different language and

cultural context. We adopted Latent Dirichlet Allocation technique of topic modeling as the analysis applied in the referential study (Peres, 2018), whereby its advantages have been outlined in prediction research (Jaimes Moreno et al., 2019).

Projecting the traditional personality psychology procedure on specific social media output, based on the results of previous studies, our tentative hypotheses may be as follows. Namely, we expect to find a substantial presence of personality descriptive adjectives in Serbian Tweets, though we expect the set of terms to be smaller than in psycholexical studies, as suggested by Roivainen (2015a). Secondly, we expect the topics extracted to contain semantically coherent adjective combinations, but we expect modest or moderate similarities to established trait structures such as Goldberg's personality adjectives (Goldberg, 1981, 1990) and the overall five-factor structure based on the merged results of three Serbian psycholexical studies (De Raad et al., 2018). This assumption is based on the results of previous studies, particularly Peres (2018). As for adjective categories, we expect all of them to appear in Serbian tweets.

However, the assumptions regarding categories' structure and distribution across topics are more challenging to articulate, since so far, they have not been used in Twitter-related studies. However, the implications of two possible outcomes can be provisionally outlined. If the results show no deviations of the word categories' distributions in Tweets from the distributions in lexical studies, one could assume that lexical word categories may be valid across discourses, and not only applicable to questionnaire-gathered data. If the deviations are found, it would suggest that Twitter discourse may be specific regarding the use of personality descriptor categories.

An additional incentive for this study regards the data sources used in previous studies. Most of the studies cited in this paper, including the referential ones, did not utilize fully open-access data sources. This state of affairs may be due to the limited accessibility of the sources such as Twitter/X and GoogleBooks. In the current study, we have chosen to capitalize on the open accessibility of the Serbian Twitter data collected between 2008 and 2014,

assembled in the *Tweet-sr* linguistic corpus (Ljubešić & Klubička, 2014) available in *noSketch* and *KonText* services. We believe that the use of an open, fully tagged linguistic corpus as a source of Twitter/X archival data will both contribute to the understanding of Serbian Tweeter discourse and encourage future replication studies in Serbian and other languages contained in similar corpora accessible in previously mentioned locations.

Method

Procedure

The methodological procedure we used in this study to extract and process the Twitter data was based on the methodology described in Peres (2018). This procedure is in line with the core methodological principles of the classical approaches in psycholexical studies, as described in Hofstee, whereby descriptor categorization was included as described in Angleitner et al. (1990) and applied in the third Serbian psycholexical study (De Raad et al., 2018). The procedure included the following steps:

Open data extraction:

1. Using an open-access linguistic repository, „Tweet-sr (Serbian Tweets)" (Ljubešić & Klubička, 2014) (containing 174,235,555 words), tweets in Serbian language were extracted containing the phrase „I am". Only Tweets in Latin alphabet were used. The tweets were extracted in the lemmatized form provided in the repository.
2. Tweets containing at least one adjective from the Serbian 383 personality-descriptive adjectives list (De Raad et al., 2018) were retained for further analysis. 2a. A total of 268 Serbian descriptors were found in 109759 Tweets containing the phrase "I am".
3. The retained adjectives' category was determined using the categorization from the third Serbian psycholexical study and described in detail in De Raad et al. (2018).
4. A document-feature matrix (Benoit et al., 2018), was formed using tweets as documents and the retained adjectives as features.

5. For validation purposes, steps 2 and 4 were applied using Goldberg's list of 100 personality-descriptive adjectives (Goldberg, 1981). We intended to use these results to estimate similarities between Serbian topics' contents and the Big Five personality traits. Word categorization was not available for Goldberg's adjective list and thus was not applied. To control for possible translation effects on results, two versions of Goldberg's 100 were used: the original English one, and the Serbian translated by the authors. We kept the version containing original Serbian and translated Goldberg terms, since it contains the original Serbian adjective descriptors and provides a more conservative estimation of topic similarities.

Data processing

6. Latent Dirichlet Allocation as a method of topic modeling was applied separately on Serbian adjectives and Goldberg descriptive adjectives. For both sets of terms, the analytic procedure included the following:

a. Determining the optimal number of topics, based on four coefficients and visual inspection of the topics' distances, explained in more detail in the Data analysis section.

b. Terms with the largest term-topic probabilities, i.e., the terms with the highest likelihoods of belonging to a particular topic and simultaneously the smallest likelihoods of belonging to other topics, were extracted as optimal topic descriptors. To obtain the broadest range of topic indicators and ensure optimum reliability, we decided to impose the maximum upper limit of the number of indicators per topic, i.e., indicator number divided by topic number, as enabled by the software used (Watanabe et al., 2023).

c. Topic content similarities were calculated using the document similarity function as implemented in the *quanteda* package (Benoit et al., 2018) This step was conducted for:

d. Serbian Twitter topics and topics based on Goldberg's Big Five descriptor list;

ii. Serbian 383-based topics and five overall factors from the third psycholexical study, as outlined in De Raad et al. (2018);
iii. Goldberg 100 adjectives-based topics and the contents of the five original Goldberg scales, measuring the lexical Big Five dimensions (Emotional stability, Extraversion, Intellect, Agreeableness, and Conscientiousness) (Goldberg, 1981).

7. For the Serbian 383 adjectives-based topics, frequencies of word categories were calculated. The relative frequencies of the categories appearing in tweet topics were compared to those in the third Serbian psycholexical study (De Raad et al., 2018). Categories' frequencies across topics were presented as a contingency table.

Data analysis

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is an unsupervised algorithm that groups the documents with respect to the topics (i.e., a topic modeling technique). The high-level idea of LDA is that each document is described with a set of topics. Whereas, each topic is represented with a group of words, more specifically a probability distribution of words is given for each topic. Within this work each personality trait is described with a set of adjectives, therefore the words we are interested in among the topics are the adjectives themselves. In order to see how the adjectives that belong to certain personality traits are distributed among the topics we performed LDA.

For LDA the most important parameter that needs to be defined is the number of topics N. If the number of topics is set to a small value the model will be focused around general topics, whereas if the number of topics is set to a large value, the model will create topics that overlap. LDA was conducted in R (Ponweiser, 2012; R Core Team, 2023) by using the packages *ldatuning* for LDA (Nikita, 2020) and *seededlda* for topics' term extraction (Watanabe et al., 2023). To determine the optimal number of topics for the LDA model we used four metrics from the *ldatuning* package (Nikita, 2020). The optimal number of topics show low values for CaoJuan and Arun metrics (Arun et al., 2010; Cao et al., 2009), and high values for Griffiths and Deveaud metrics (Deveaud et al., 2014; Griffiths

& Steyvers, 2004). Two smoothing hyperparameters, *alpha* and *beta*, whose combination determines the distributional features (Celard et al., 2020), were set to the values of one. We made this decision given that the hyperparameters' zero values would imply the smoothest distribution of words across topics, while the values of one would allow for the most scattered distribution. To enable the full range of possible distributional features, and not exclusively the smoothest one, we opted for the latter.

Similarly to Peres (2018) and according to the default parameters contained in the software solution we used (Nikita, 2020), we conducted the initial analyses including two to fifteen-topics solutions, among which we selected the one with the best coefficients' values. To complement the values of the raw coefficients, we calculated the standardized differences between the maximum- and minimum-values aimed coefficients and thus attempted to determine the optimum solution. Simultaneously, we plotted the initial fifteen-topic solution using multidimensional scaling based on topics' Euclidean distances, to visually asses topic overlap and choose the least-overlapping solution.

Topic similarities, as well as topic-personality dimensions' similarities, were calculated using cosine similarity between the sets of terms constituting each topic, as implemented in the *quanteda* package in R (Benoit et al., 2018). In LDA cosine similarity is a measure used for calculating a distance between two term frequency vectors with values ranging from 0 to 1, and larger values indicating higher similarity (Sidorov et al., 2014), while values closer to zero indicate orthogonality.

Results

Approximately 69.97% of the Serbian personality adjectives lexicon described in De Raad et al. (2018) were found in our study. The frequencies of the retained adjectives within the dataset used in this study are shown in Supplementary materials, Table 1.

To complement the information regarding word frequency, we have compared word frequencies of the retained 268 adjectives and the remaining

115 adjectives from the reduced Serbian list (De Raad et al, 2018). We used two data sources for the comparison: the Tweet-sr corpus as described previously in this paper, and the srWAC corpus of Serbian language, assembled using the available web resources and accessible within the NoSketch resources (Ljubešić & Klubička, 2016). The results show that the retained terms are significantly more frequent than the omitted ones both in Tweet-sr ($M_{retained}$ = 2276.429 (5592.151), $M_{omitted}$ = 51.243 (66.867), t (381) = -4.264, p < .001, Mann-Whitney = 2145.00, p <.001) and srWAC ($M_{retained}$ = 10506.332 (23437.028), $M_{omitted}$ = 630.835 (787.135), t (381) = -4.514, p < .001, Mann-Whitney = 5929.00, p <.001).

Determining topic numbers

Topic number: Serbian 383 adjectives–based topics

The results of four metrics for 2-15 topics are plotted in Figure 1. Topics' positions for the 2-15 solutions are shown in Figure 2. The positions were determined using multidimensional scaling (MDS) space based on topic Euclidean distances. Four topics-solution was chosen as optimal due to topical parsimony, as shown in Figure 3. Additional measures of topics' standardized differences are shown in Supplementary materials, Table 2.

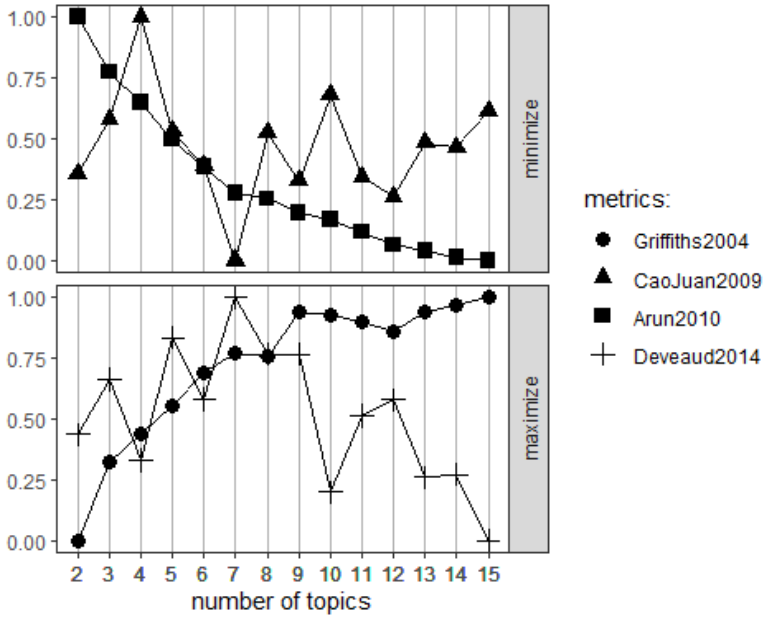

Figure 1. Fifteen topics based on Serbian 383 adjectives: coefficients

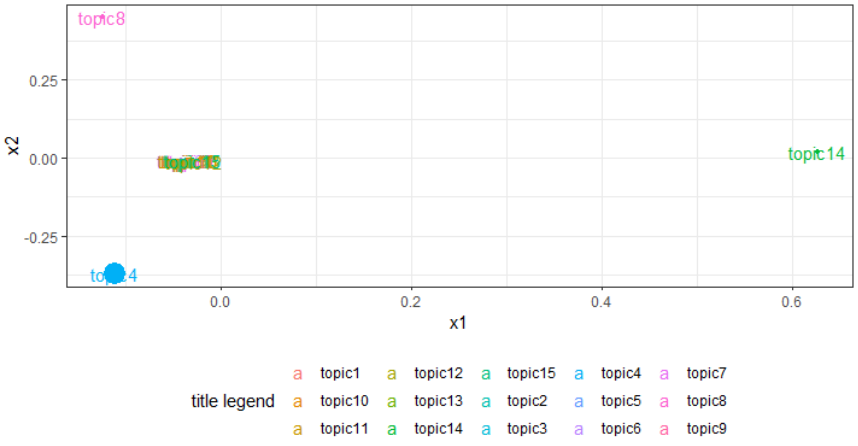

Figure 2. Fifteen topics based on Serbian 383 adjectives: Topic overlap (MSA)

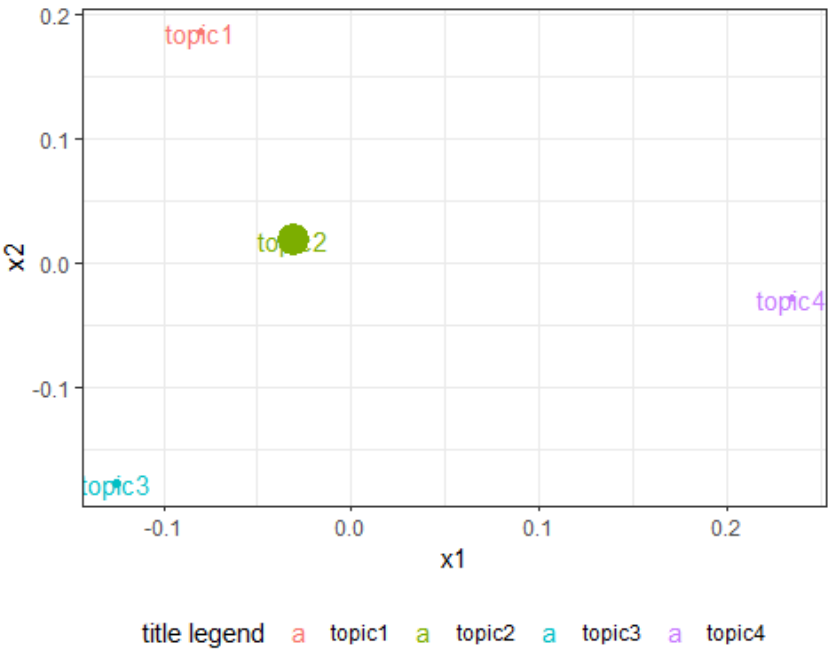

Figure 3. Four topics based on Serbian 383 adjectives: overlap

Topic number: Goldberg's 100 adjectives-based topics

The steps described in the previous section were also followed when the LDA model is based on Goldberg's 100 adjectives. According to the results of four metrics depicted in Figure 4, and the MDS - estimated topic positions (Figure 5) four topics were chosen as the optimal solution, with no visible overlaps among the four topics (Figure 6). Standardized differences based on minimum- and maximum-value aimed coefficients are shown in Supplementary materials, Table 3.

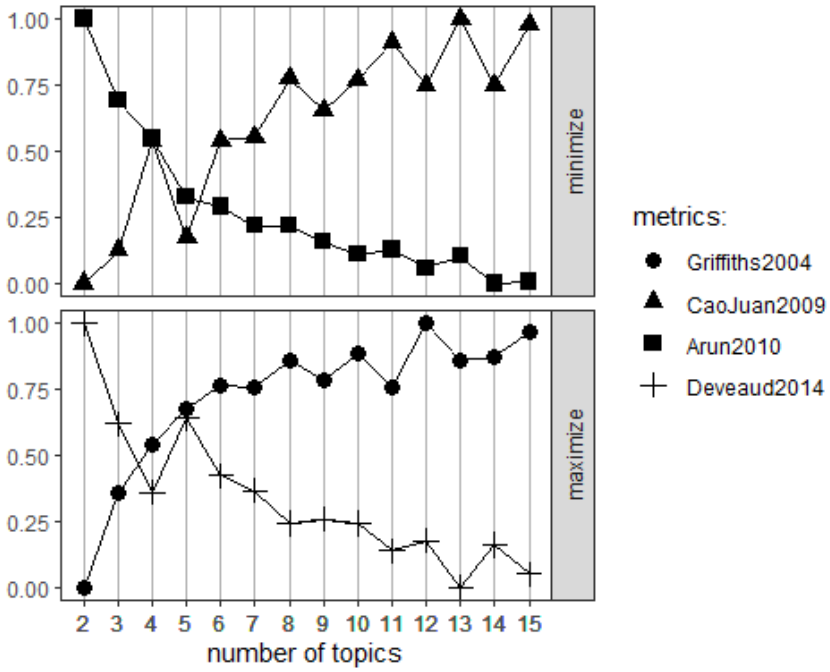

Figure 4. Fifteen topics based on Goldberg's 100 adjectives: coefficients

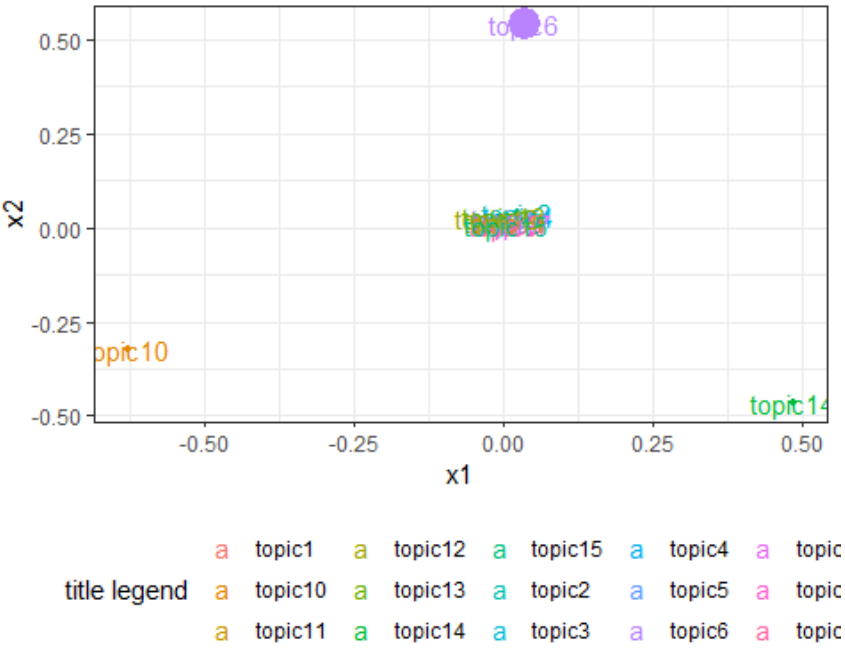

Figure 5. Fifteen topics based on Goldberg's 100 adjectives: Topic overlap (MSA)

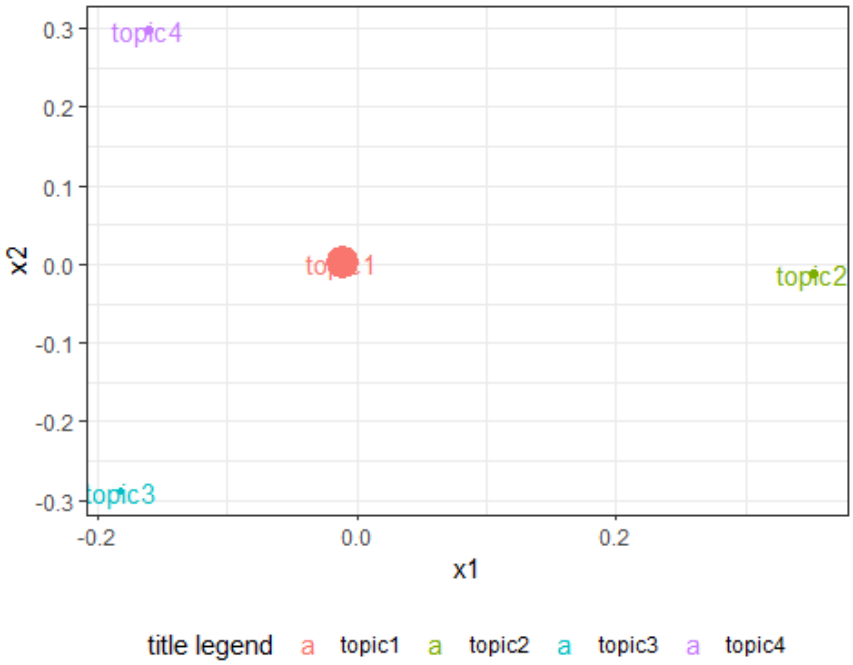

Figure 6. Four topics based on Goldberg's 100 adjectives: overlap

Topics contents

Serbian 383 adjectives–based topics

Table 1 outlines the 20 most frequent terms (adjectives) per every topic for Serbian adjectives and Goldberg adjective descriptors. The topics, both in Serbian Tweets' adjectives and in Goldberg adjective descriptors, are heterogeneous regarding the markers of personality traits that constitute their content. Given that topics are blends of personality trait markers, we opt to offer a more detailed interpretation at the end of this section, through a summary of the topics' similarity to personality traits and category distributions within topics.

Table 1

Serbian and Goldberg topics: Distribution of top 20 terms across each topic

Serbian topic 1	Serbian topic 2	Serbian topic 3	Serbian topic 4	Goldberg topic 1	Goldberg topic 2	Goldberg topic 3	Goldberg topic 4
crazy	satisfied	guilty	normal	nervous	cold	agreeable	jealous
mellow	interesting	boring	frank	deep	emotional	creative	relaxed
nervous	sad	modest	ordinary	kind	quiet	anxious	simple
proud	beloved	realistic	pert	unexcitable	pleasant	touchy	active
jealous	cold	hardworking	witty	careful	shy	helpful	artistic
important	cultured	weak	brilliant	selfish	uninquisitive	shallow	reserved
emotional	natural	weird	simple	organized	fretful	disorganized	conscientious
stubborn	depressive	romantic	lonely	envious	rude	intellectual	demanding
honest	naive	slow	capable	insecure	withdrawn	complex	generous
amusing	complicated	well-mannered	dependent	warm	introverted	neat	uncharitable
furious	quiet	smiling	creative	practical	harsh	distrustful	daring
responsible	captive	pleasant	active	efficient	unemotional	unrestrained	considerate
perverse	different	original	busy	thorough	inhibited	negligent	undemanding
powerful	crooked	concerned	unhappy	unintelligent	nervous	quiet	innovative
intriguing	shy	tolerant	decorous	untalkative	warm	imaginative	steady
self-supporting	intelligent	brutal	successful	vigorous	sloppy	haphazard	bright
contemporary	susceptible	insensitive	moral	bright	uncharitable	undependable	helpful
stupid	desperate	rugged	kind	unadventurous	helpful	talkative	uninquisitive
wise	advanced	careful	aggressive	temperamental	neat	trustful	imaginative
violent	subtle	spontaneous	dark	uncreative	distrustful	imperturbable	practical

Similarities

Serbian topics and Serbian lexically-derived personality dimensions

When examining similarities between Serbian topics and lexical markers of higher-order traits from De Raad et al. (2018), results show lower to average cosine similarities values (Figure 7). For the Serbian topic one, the maximum similarity was found for Extraversion, (theta = 0.16), while minimum cosine distance was found for Negative Valence , (theta = 0.03). For topic two, the similarities ranged from (theta = 0.03) for Neuroticism-related to (theta = 0.26) for Extraversion. Topic three was most similar to Agreeableness (theta = 0.21), and least similar to Negative Valence (theta = 0.08). Topic four showed the largest cosine similarity to Extraversion (theta = 0.18), and the smallest to Neuroticism-related (theta = 0).

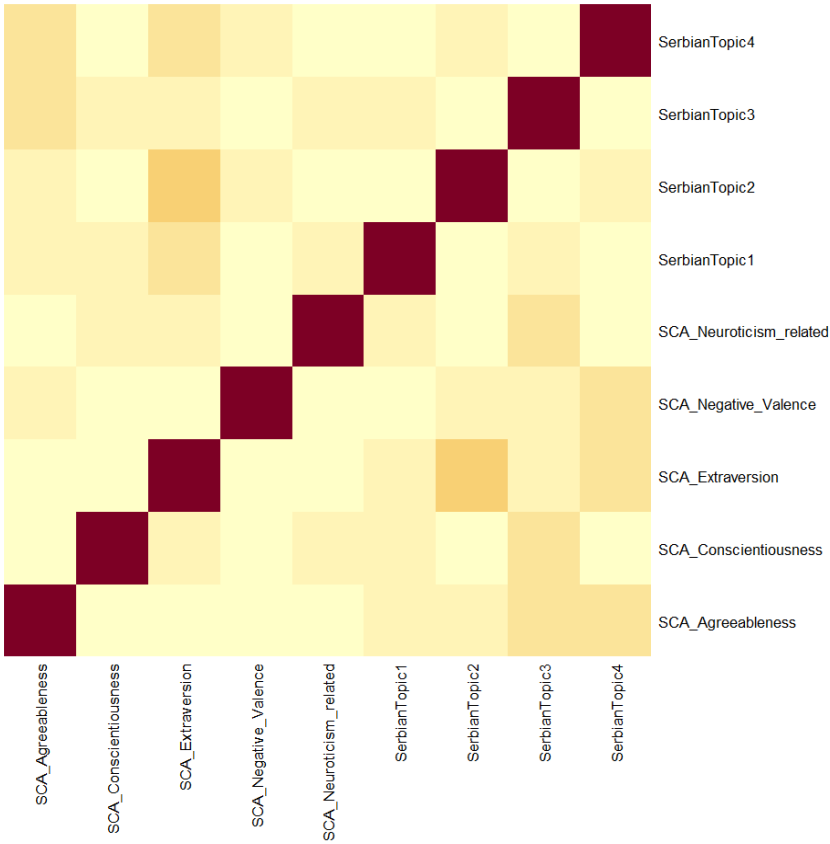

Figure 7. Four topics based on Serbian 383 adjectives: Word - topic probabilities

Note. SCA Agreeableness – SCA Neuroticism-related: personality dimensions subsuming the findings of the three Serbian psycholexical studies (De Raad et. al, 2018); SerbianTopic1 – Serbian Topic4: topics extracted in Serbian tweets gathered from the "Tweet-sr" corpus. Darker colors indicate larger cosine similarities.

Serbian topics and Big Five dimensions (Goldberg)

Topic one extracted from Serbian Tweets (Figure 8) is most similar to Emotional Stability (theta = 0.11) and least similar to Agreeableness (theta = 0.03). Topic two cosine similarities span from theta = 0 for Intellect to theta = 0.11 for Agreeableness. Topic three is most similar to Conscientiousness (theta = 0.08) and least similar to Agreeableness (theta = 0.03). Topic four has the largest

cosine similarity to Intellect (theta = 0.08) and the smallest to Agreeableness (theta = 0.03).

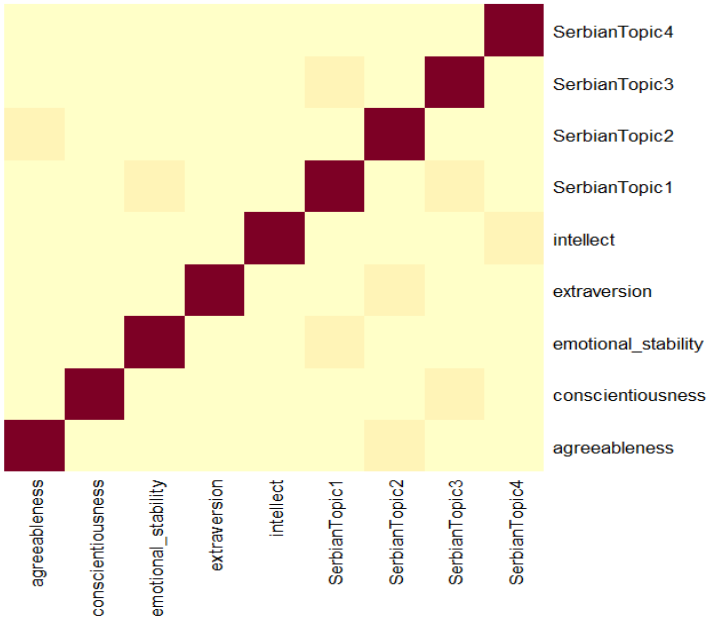

Figure 8. Serbian topics (original) and Big Five (Goldberg) dimensions (translated to Serbian) - cosine similarities

Note. Serbian Topic 1 – Serbian Topic 4: topics extracted in Serbian tweets gathered from the "Tweet-sr" corpus; intellect, extraversion, emotional stability, conscientiousness, agreeableness – Big Five dimensions measured using Goldberg's set of personality-descriptive adjectives (Goldberg, 1981; Goldberg, 1990). Darker colors indicate larger cosine similarities.

Serbian and Big Five-based (Goldberg) topics

Results point out lower to average similarities between topics based on Serbian 383 adjectives and topics based on Goldberg's 100 adjectives (Figure 9).

Serbian topic one is most similar to Goldberg Topic 2, (theta = 0.1) and least similar to Goldberg Topic 3, (theta = 0.02). For the second topic, the largest

similarity was with Goldberg Topic 2, (theta = 0.15) and the smallest with Goldberg Topic 3, (theta = 0.02). The largest cosine distance for Serbian topic three was with Goldberg Topic 2, (theta = 0.07) and the smallest with Goldberg Topic 1, (theta = 0.05). Cosine similarities for topic four spanned from theta = 0.05 for Goldberg Topic 1, to theta = 0.05 for Goldberg Topic 1.

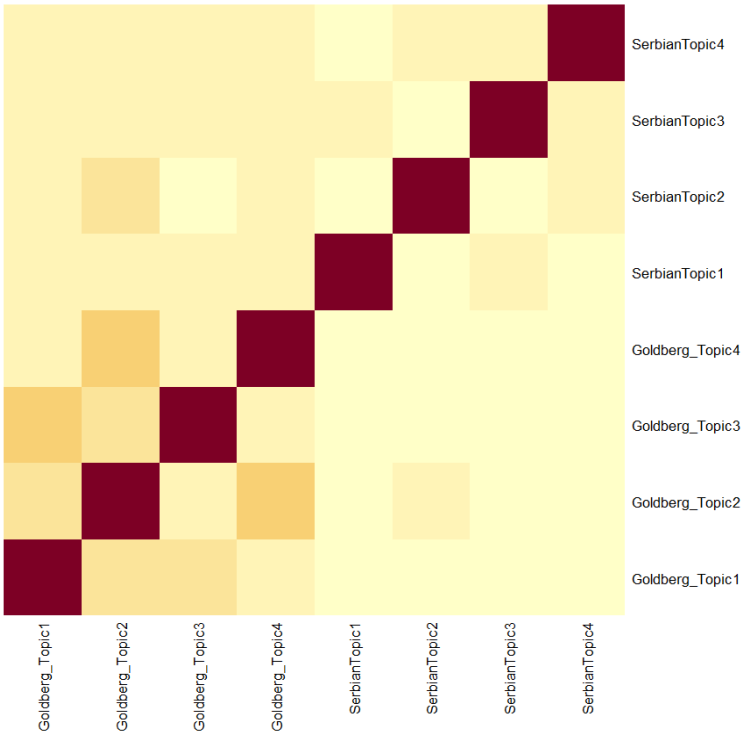

Figure 9. Serbian topics and Goldberg topics - cosine similarities

Note. SerbianTopic1 – Serbian Topic 4: topics extracted in Serbian tweets gathered from the "Tweet-sr" corpus; Goldberg Topic 1 – Goldberg Topic 4: topics extracted in Serbian tweets based on the Big Five dimensions' markers as conceptualized by Goldberg (Goldberg, 1981; Goldberg, 1990). Darker colors indicate larger cosine similarities.

Serbian topics - word categories

Supplementary materials, Table 4 displays word categories frequencies for Serbian 383 adjectives-based topics.

Word categories' frequencies in Serbian Tweets (Supplementary materials, Table 5) were compared to category proportions in the third Serbian psycholexical study (De Raad et al., 2018). The results showed no differences $\chi 2$ (8) = 7.79; p =0.45 in category distributions. Tweets' categories' distributions across topics (Supplementary materials, Table 5) did not reveal substantial differences in category patterns within topics $\chi 2$ (24) = 29.99; p =0.19.

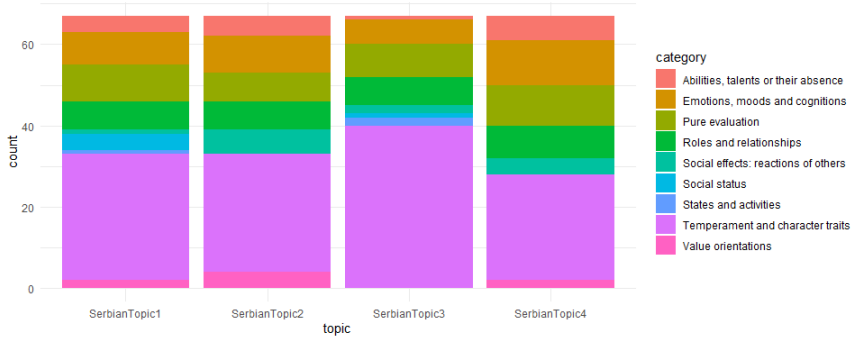

Figure 10. Distribution of word categories across Serbian topics

Note. Serbian Topic 1 – Serbian Topic 4: topics extracted in Serbian tweets gathered from the "Tweet-sr" corpus; categories – adjective categories as presented in the third Serbian psycholexical study (De Raad et al., 2018)

A summary of the topics' features is shown in Table 2. Three of four Serbian Twitter topics show most pronounced, though modest, cosine similarities to Serbian top-tier Extraversion dimension, while one is most similar to Agreeableness. Negative Valence and Neuroticism are the dimensions to which most of the topic vectors are orthogonal. Among all topics, markers of Temperament and character traits category are most frequent, with Pure evaluation in the second place for the first three topics, and Emotions, moods and cognitions for the fourth. Among the Big Five (Goldberg) dimensions, similarity of Twitter topics is more diverse, with extraversion not being among the most similar dimensions for any of the topics. However, one should bear in

mind the differences in conceptualizations of broad personality traits between Goldberg's and Serbian "emic" studies. Additionally, the similarities with Serbian traits are substantially larger than with Goldberg's, which could emphasize the relevance of language issues and cultural context.

A tentative interpretation of the topics' contents could suggest that all of them involve mostly self-descriptions of stable traits, with pure evaluation and emotions/moods as secondary saturators. Provisionally, only by referring to the twenty highest-loading indicators, topic one appears to contain descriptions pointing to social dominance and overt representation, the second one to emotional aspects of social presentation, the third points to activities and socially desirable behaviors, while the fourth appears to capture the terms that would constitute representations of one as ordinary and non-exceptional. If we approached the topics from the viewpoint of self-representation biases as conceptualized by Paulhus and John (1998), we could argue that topics one and four are more in line with "egoistic biases", while two and three are more in line with "moralistic biases" (Paulhus & John, 1998; Pedović, 2021).

Table 2

Serbian topic profiles based on similarities and category frequencies

Serbian topic	Goldberg - most similar	Goldberg - least similar	Serbian lexical - most similar	Serbian lexical - least similar	Descriptor category (minus stable) - most frequent	Descriptor category (minus stable) - least frequent
Serbian Topic 1	Emotional Stability	Agreeableness	Extraversion	Negative Valence	Pure evaluation	Social effects: reactions of others
Serbian Topic 2	Agreeableness	Intellect	Extraversion	Neuroticism-related	Pure evaluation	Social effects: reactions of others
Serbian Topic 3	Conscientiousness	Agreeableness	Agreeableness	Negative Valence	Pure evaluation	Value orientations
Serbian Topic 4	Intellect	Agreeableness	Extraversion	Neuroticism-related	Emotions, moods and cognitions	Social status

Discussion

In this study, we attempted to gain insight into the semantic structure of self-referent Tweets in Serbian language. To accomplish that, we approached the Twitter material using a methodological procedure applied in classic psycholexical studies, combined with a widely applied NLP technique (LDA topic modeling), and building on a single contemporary study conducted on Twitter material so far in a similar fashion. Despite its roots in classic and current studies, we tend to see this study as an exploratory one, primarily because it is, to our knowledge, the first personality study using Tweets in Serbian language. The results provide answers to the questions we posed, but, perhaps more importantly, open new ones to be addressed in future studies.

The number of extracted terms revealed that approximately 70% of the Serbian trait lexicon appeared in Tweets. This result is congruent with Roivainen (2015a) and speaks in favor of the findings suggesting that Twitter personality vocabulary is "smaller" than the one comprised in standard language. One possible account for this result could take into account the Tweets' brevity i.e., the pre-imposed restriction on a maximum number of words allowed. Shorter messages probably involve semantically condensed terms of specific connotation, which is an issue that should be addressed in future studies. The adjectives found within the Tweet-sr corpus and consequently analyzed are substantially more frequent than the remaining one hundred and fifteen adjectives that were not found in Tweets. This result is in line with the expectations that a communication "device" such as Twitter would rely on more common words. Nevertheless, it poses a more specific question of the impact of personality descriptors' frequencies on their use in various contexts. Such a question has recently been addressed by Condon et al (2022) and Condon & McDougald, (2022), and in Serbian language by Čolović et al. (2012). However, we believe that, due to its complexity, it should be a highly relevant topic for future studies in a range of languages.
We extracted four distinct topics in Serbian Tweets, which appear to reflect specific semantic structures.

This result is also in line with previous studies' results, which did not find conclusive links between Tweet topics and personality traits (Peres, 2018). Although topics do not replicate trait constructs, they are modestly related to them. Focusing on Serbian topics, we found the largest similarities (though still modest to moderate, according to standard interpretation) with Extraversion and Agreeableness. According to well-established conceptions in personality psychology, such as the Interpersonal circumplex (Gurtman, 2009), Extraversion and Agreeableness are perceived as the traits most relevant for interpersonal behavior. Hence the explanation of their similarity to Tweet topics may have sound conceptual foundations. As means of informal, brief written communication, Tweets are intuitively expected to convey socially relevant information that can best be carried through personality markers from the dimensions mentioned above. Hence we believe that, in future studies, more attention should be given to interpersonal circumplex concepts and their structure within the Twitter discourse. Topic categories are equally distributed across topics, and their distribution is equal to the distribution described in the third Serbian psycholexical study. This may be taken as a result in favor of the validity and applicability of personality adjectives' categories in Twitter discourse. However, there are no substantial inter-topic differences in category distributions. While this result can also be seen as a tentative confirmation of topic categories' validity, it limits the possibilities for topic distinction. Nevertheless, when stable trait terms are excluded, the less frequent categories' distributions apparently, though not largely, differ among topics. Pure evaluation is present as the second most frequent category in three topics, while in one of them emotions are most frequent. While evaluative personality dimensions are virtually orthogonal to topics' vectors, evaluation is still present within the predominantly socially themed topics. That may mean that, when communicating socially relevant self-referent information, Serbian Tweeters may be using evaluatively profiled (desirable or undesirable) terms either to facilitate the comprehensibility of communication or to establish more transparent impressions of themselves with their co-communicators. At the same time, communicating situation-specific emotions and moods may be one

of the most important functions of self-talk in Tweets and, as such, deserves more careful consideration in future studies.

Limitations and future directions

One major conceptual (no less methodological) limitation of this study is the exclusive use of adjectives as personality descriptors. We made this decision to ensure compliance with previous psycholexical studies, where adjectives have been the most frequently used word type. However, given the idiosyncrasies of Twitter discourse (or slang), one may wonder whether nouns (as more efficient "type" descriptors) and verbs (as more accurate regarding behavioral cues) should be included. The use of adjectives in self-describing tweets may have overlooked the effect of other word types, and even syntactic variables (tweet length, sentence length, etc.) and hence obscured their relevance for the current results. Including other word types is certainly one of the crucially important tasks for future studies.

Methodologically, we have made several decisions whose implications could be termed as either overly liberal or overly conservative. Having no prior knowledge of words' distributions within the Serbian Tweets, we opted for the least restrictive setting for topic formation, allowing for any distributional features in the final outcome (i.e., topics.) This way we obtained maximally distinctive topics, having no information on the implications of such distinctiveness. Additionally, we have used the tweets in a single (Serbian) language, which limits the possibility of full validation. Open accessibility of Twitter resources in similar (Croatian, Bosnian) and less similar languages (English, Japanese) can enable a good starting point for the validation of these results and possible replication of the current study in different language settings. Finally, self-themed Tweets are only one piece of the personality-tweeting puzzle. Addressing the issues of tweeting about others may greatly help us understand the structure and specific features of tweet topics.

To conclude, in this study we have applied classic psycholexical methodology to study self-referencing tweets. While the results show that personality trait content is present in the extracted topics, it suggests that

personality adjectives or adjective-based traits are most likely not sufficient to provide the full account on personality descriptors' use in this specific medium. Hence future studies are warranted to address the issues of word types, syntactic features, self-talk or talk of others, and other important and still open questions.

Conflicts of Interest

The authors declare no conflicts of interest with respect to the authorship or the publication of this article.

Data availability statement

Primary data used in this study are available upon a reasonable request.

References

Almagor, M., Tellegen, A., & Waller, N. G. (1995). The Big Seven model: A cross-cultural replication and further exploration of the basic dimensions of natural language trait descriptors. *Journal of Personality and Social Psychology*, *69*(2), 300. https://doi.org/10.1037/0022-3514.69.2.300

Angleitner, A., Ostendorf, F., & John, O. P. (1990). Towards a taxonomy of personality descriptors in german: A psycho-lexical study. *European Journal of Personality*, *4*(2), 89–118. https://doi.org/10.1002/per.2410040204

Arun, R., Suresh, V., Veni Madhavan, C. E., & Narasimha Murthy, M. N. (2010). On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. In M. J. Zaki, J. X. Yu, B. Ravindran, & V. Pudi (Eds.), *Advances in Knowledge Discovery and Data Mining* (Vol. 6118, pp. 391–402). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-13657-3_43

Barelds, D. P. H., & Raad, B. D. (2015). The role of word-categories in trait-taxonomy: Evidence from the Dutch personality taxonomy. *International Journal of Personality Psychology*, *1*, 15–25. https://ijpp.rug.nl/article/view/19316

Benet-Martínez, V., & Waller, N. G. (2002). From adorable to worthless: Implicit and self-report structure of highly evaluative personality descriptors. *European Journal of Personality*, *16*(1), 1–41. https://doi.org/10.1002/per.431

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). Quanteda: An r package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774–774. https://doi.org/10.21105/joss.00774

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf?ref=https://githubhelp.com

Boyd, R. L., & Pennebaker, J. W. (2017). Language-based personality: A new approach to personality in a digital world. *Current Opinion in Behavioral Sciences*, *18*, 63–68. https://doi.org/10.1016/j.cobeha.2017.07.017

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, *72*(7–9), 1775–1781. https://doi.org/10.1016/j.neucom.2008.06.011

Carducci, G., Rizzo, G., Monti, D., Palumbo, E., & Morisio, M. (2018). Twitpersonality: Computing personality traits from tweets using word embeddings and supervised learning. *Information*, *9*(5), 127. https://doi.org/10.3390/info9050127

Cattell, R. B., & Kline, P. E. (1977). *The scientific analysis of personality and motivation.* Academic Press.

Celard, P., Vieira, A. S., Iglesias, E., & Borrajo, L. (2020). LDA filter: A latent dirichlet allocation preprocess method for weka. *Plos One*, *15*(11), e0241701. https://doi.org/10.1371/journal.pone.0241701

Christian, H., Suhartono, D., Chowanda, A., & Zamli, K. Z. (2021). Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging. *Journal of Big Data*, *8*(1), 68. https://doi.org/10.1186/s40537-021-00459-1

Colovic, P., Mitrovic, D., & Smederevac, S. (2005). Evaluation of Big Five model in Serbian culture by FIBI questionnaire. *Psihologija*, *38*(1), 55–76. https://doi.org/10.2298/PSI0501055C

Čolović, P., Smederevac, S., Milin, P., & Mitrović, D. (2012). *The structure of mid- to most frequent lexical personality descriptors in the Serbian language*. 16th European Conference on Personality, Book of Abstracts, 204–205, July 10-14, Trieste, Italy. https://www.researchgate.net/publication/263152111_The_structure_of_the_mid-_to_most_frequent_lexical_personality_descriptors_in_the_Serbian_language

Čolović, P. & Filipović Đurđević, D. (2019). *Lexical personality dimensions in the land of distributional semantics*. 3rd World conference on Personality, April 2-6, Hanoi,

Vietnam, World Association for Personality Psychology, Programme and Abstracts, 51.

Čolović, P., Smederevac, S., & Mitrović, D. (2014). Velikih Pet Plus Dva: Validacija Skraćene Verzije. *Primenjena Psihologija*, *7*(3–1), 227. https://doi.org/10.19090/pp.2014.3-1.227-254

Condon, D. M., Coughlin, J., & Weston, S. J. (2022). Personality Trait Descriptors: 2,818 Trait Descriptive Adjectives Characterized by Familiarity, Frequency of Use, and Prior Use in Psycholexical Research. *Journal of Open Psychology Data*, *10*(1), 1. https://doi.org/10.5334/jopd.57

Condon, D. M., & McDougald, S. (2022). *Frequency of use metrics for American English person descriptors: Extensions of Roivainen's internet search methodology* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/9gtj7

Cutler, A., & Condon, D. M. (2022). Deep lexical hypothesis: Identifying personality structure in natural language. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/pspp0000443

De Raad, B., & Mlačić, B. (2020). The Big Five Personality Trait Factors. In *Oxford Research Encyclopedia of Education*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190264093.013.894

De Raad, B., Mulder, E., Kloosterman, K., & Hofstee, W. K. (1988). Personality-descriptive verbs. *European Journal of Personality*, *2*(2), 81–96. https://doi.org/10.1002/per.2410020204

De Raad, B., & Ostendorf, F. (1996). Quantity and quality of trait-descriptive type nouns. *European Journal of Personality*, *10*(1), 45–56. https://doi.org/10.1002/(SICI)1099-0984(199603)10:1<45::AID-PER245>3.0.CO;2-6

De Raad, B., Smederevac, S., Čolović, P., & Mitrović, D. (2018). Personality traits in the Serbian language: Structure and procedural effects. *Journal of Research in Personality*, *73*, 93–110. https://doi.org/10.1016/j.jrp.2017.11.008

Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document Numérique*, *17*(1), 61–84. https://doi.org/10.3166/dn.17.1.61-84

Fischer, R., Karl, J. A., Luczak–Roesch, M., Fetvadjiev, V. H., & Grener, A. (2020). Tracing personality structure in narratives: A computational bottom–up approach to unpack writers, characters, and personality in historical context. *European Journal of Personality*, *34*(5), 917–943. https://doi.org/10.1002/per.2270

Golbeck, J., Robles, C., Edmondson, M., & Turner, K. (2011). Predicting personality from twitter. *2011 IEEE Third International Conference on Privacy, Security, Risk and*

Trust and 2011 IEEE Third International Conference on Social Computing, 149–156. https://doi.org/10.1109/PASSAT/SocialCom.2011.33

Goldberg, L. (1981). Language and Individual Differences: The Search for Universals in Personality Lexicons. In L. Wheeler (Ed.), *Review of Personality and Social Psychology* (pp. 141–165). Sage Publication. https://projects.ori.org/lrg/PDFs_papers/universals.lexicon.81.pdf

Goldberg, L. R. (1990). An alternative" description of personality": The big-five factor structure. *Journal of Personality and Social Psychology*, *59*(6), 1216. https://doi.org/10.1037/0022-3514.59.6.1216

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(suppl_1), 5228–5235. https://doi.org/10.1073/pnas.0307752101

Gurtman, M. B. (2009). Exploring personality with the interpersonal circumplex. *Social and personality psychology compass*, *3*(4), 601–619. https://doi.org/10.1111/j.1751-9004.2009.00172.x

Henss, R. (1995). From aal to zyniker. Personality descriptive type nouns in the german language. *European Journal of Personality*, *9*(2), 135–145. https://doi.org/10.1016/0191-8869(95)00093-L

Hofstee, W. K. (1990). The use of everyday personality language for scientific purposes. *European Journal of Personality*, *4*(2), 77–88. https://doi.org/10.1002/per.2410040203

Hofstee, W. K., De Raad, B., & Goldberg, L. R. (1992). Integration of the big five and circumplex approaches to trait structure. *Journal of Personality and Social Psychology*, *63*(1), 146. https://doi.org/10.1037/0022-3514.63.1.146

Jaimes Moreno, D. R., Carlos Gomez, J., Almanza-Ojeda, D.-L., & Ibarra-Manzano, M.-A. (2019). Prediction of Personality Traits in Twitter Users with Latent Features. *2019 International Conference on Electronics, Communications and Computers (CONIELECOMP)*, 176–181. https://doi.org/10.1109/CONIELECOMP.2019.8673242

Kern, M. L., McCarthy, P. X., Chakrabarty, D., & Rizoiu, M.-A. (2019). Social media-predicted personality traits and values can help match people to their ideal jobs. *Proceedings of the National Academy of Sciences*, *116*(52), 26459–26464. https://doi.org/10.1073/pnas.1917942116

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, *110*(15), 5802–5805. https://doi.org/10.1073/pnas.1218772110

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO personality inventory. *Multivariate Behavioral Research*, *39*(2), 329–358. https://doi.org/10.1207/s15327906mbr3902_8

Ljubešić, N., & Klubička, F. (2014). {bs,hr,sr}WaC - Web Corpora of Bosnian, Croatian and Serbian. *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, 29–35. https://doi.org/10.3115/v1/W14-0405

Ljubešić, N., & Klubička, F. (2016). Serbian web corpus srWaC 1.1. *http://nlp.ffzg.hr/resources/corpora/srwac/*. https://www.clarin.si/repository/xmlui/handle/11356/1063

Mavis, G., Toroslu, I. H., & Karagoz, P. (2021). Personality Analysis Using Classification on Turkish Tweets: *International Journal of Cognitive Informatics and Natural Intelligence*, *15*(4), 1–18. https://doi.org/10.4018/IJCINI.287596

Nikita, M. (2020). *ldatuning: Tuning of the Latent Dirichlet Allocation Models Parameters*. https://CRAN.R-project.org/package=ldatuning

Norman, W. T. (1967). *2800 personality trait descriptors*: *Normative operating characteristics for a university population*. Ann Arbor, MI, USA: Department of Psychology, University of Michigan. ttps://eric.ed.gov/?id=ed014738

Oljača, M., Filipović Đurđević, D., & Čolović, P. (2018). Struktura pridevskih opisa ličnosti u pisanom jeziku. *66. naučno-stručni skup kongres psihologa Srbije, Knjiga rezimea*, Društvo psihologa Srbije, Zlatibor 30. maj—2. jun, 117.

Passakos, C. G., & De Raad, B. (2009). Ancient personality: Trait attributions to characters in Homer's Iliad. *Ancient Narrative*, *7*, 75–95.

Paulhus, D. L., & John, O. P. (1998). Egoistic and Moralistic Biases in Self-Perception: The Interplay of Self-Deceptive Styles With Basic Traits and Motives. *Journal of Personality*, *66*(6), 1025–1060. https://doi.org/10.1111/1467-6494.00041

Paulsen, G. (2011). *Causation and dominance: A study of finnish causative verbs expressing social dominance*. Åbo Akademi University Press.

Pedović, I. N. (2021). Relacije osobina ličnosti, situacionih varijabli i stilova odgovaranja na upitnike samoprocene. *Univerzitet u Nišu.* https://nardus.mpn.gov.rs/handle/123456789/18429

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, ourselves. *Annual Review of Psychology*, *54*(1), 547–577. https://doi.org/10.1146/annurev.psych.54.101601.145041

Peres, A. J. D. S. (2018). *The personality lexicon in Brazilian Portuguese: Studies with natural language* [Doctorate, Universidade de Brasília]. https://doi.org/10.26512/2018.01.T.32067

Ponweiser, M. (2012). *Latent Dirichlet Allocation in R* (WorkingPaper 2; Theses / Institute for Statistics and Mathematics). WU Vienna University of Economics and Business. https://research.wu.ac.at/en/publications/latent-dirichlet-allocation-in-r-3

Qiu, L., Lin, H., Ramsay, J., & Yang, F. (2012). You are what you tweet: Personality expression and perception on Twitter. *Journal of Research in Personality*, *46*(6), 710–718. https://doi.org/10.1016/j.jrp.2012.08.008

Quercia, D., Kosinski, M., Stillwell, D., & Crowcroft, J. (2011). Our twitter profiles, our selves: Predicting personality with twitter. *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, 180–185. https://doi.org/10.1109/PASSAT/SocialCom.2011.26

R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Roivainen, E. (2015a). Personality Adjectives in Twitter Tweets and in the Google Books Corpus. An Analysis of the Facet Structure of the Openness Factor of Personality. *Current Psychology*, *34*(4), 621–625. https://doi.org/10.1007/s12144-014-9274-x

Roivainen, E. (2015b). The Big Five Factor Marker Adjectives Are Not Especially Popular Words. Are They Superior Descriptors? *Integrative Psychological and Behavioral Science*, *49*(4), 590–599. https://doi.org/10.1007/s12124-015-9311-9

Saucier, G. (2003). Factor structure of english-language personality type-nouns. *Journal of Personality and Social Psychology*, *85*(4), 695. https://doi.org/10.1037/0022-3514.85.4.695

Schwartz, H., Eichstaedt, J., Kern, M., Dziurzynski, L., Lucas, R., Agrawal, M., Park, G., Lakshmikanth, S., Jha, S., & Seligman, M. (2013). Characterizing geographic variation in well-being using tweets. *Proceedings of the International AAAI Conference on Web and Social Media*, *7*, 583–591. https://doi.org/10.1371/journal.pone.0073791

Sidorov, G., Gelbukh, A., Gómez-Adorno, H., & Pinto, D. (2014). Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model. *Computación y Sistemas*, *18*(3), 491–504. https://doi.org/10.13053/cys-18-3-2043

Smederevac, S., Mitrović, D., & Čolović, P. (2007). The structure of the lexical personality descriptors in serbian language. *Psihologija*, *40*(4), 485–508. https://doi.org/10.2298/PSI0704485S

Watanabe, K., Xuan-Hieu, P., & Watanabe, M. K. (2023). *seededlda: Seeded Sequential LDA for Topic Modeling*. https://CRAN.R-project.org/package=seededlda

Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, *44*(3), 363–373. https://doi.org/10.1016/j.jrp.2010.04.001

Zhao, Y., Guo, Y., He, X., Wu, Y., Yang, X., Prosperi, M., Jin, Y., & Bian, J. (2020). Assessing mental health signals among sexual and gender minorities using Twitter data. *Health Informatics Journal*, *26*(2), 765–786. https://doi.org/10.1016/j.patrec.2020.07.035

Supplementary materials

Word frequencies

Table 1

Serbian personality-descriptive adjectives frequencies

feature	frequency	rank	docfreq	group
normalan	1,419.82389	1	612	all
kriv	1,283.12564	2	500	all
lud	1,119.85462	3	439	all
zadovoljan	680.86449	4	258	all
zanimljiv	623.26385	5	232	all
tužan	616.56779	6	211	all
dosadan	596.20580	7	214	all
drag	549.19478	8	197	all
blag	543.70284	9	197	all
nervozan	525.44637	10	176	all
ponosan	517.14859	11	182	all
ljubomoran	515.25601	12	180	all
iskren	466.22339	13	164	all
realan	419.90874	14	146	all
slab	388.54598	15	125	all
hladan	380.44458	16	130	all
emotivan	350.72601	17	117	all
skroman	322.61937	18	106	all
običan	301.87219	19	98	all
vredan	291.39467	20	94	all
važan	267.98983	21	84	all
bezobrazan	246.00805	22	77	all
kulturan	245.57095	23	78	all
prirodan	244.63803	24	73	all
duhovit	243.73417	25	75	all
depresivan	233.71615	26	69	all
genijalan	224.11137	27	69	all
tvrdoglav	191.86669	28	54	all
naivan	190.06670	29	58	all

feature	frequency	rank	docfreq	group
jednostavan	190.06670	29	58	all
pošten	187.22022	31	57	all
usamljen	183.30785	32	51	all
komplikovan	181.94255	33	54	all
čudan	179.51959	34	52	all
zabavan	175.75671	35	53	all
sposoban	172.87071	36	52	all
besan	169.97637	37	51	all
tih	161.24155	38	48	all
romantičan	156.26142	39	44	all
kreativan	149.46745	40	44	all
zavisan	146.49978	41	43	all
vaspitan	140.53391	42	41	all
spor	140.53391	42	41	all
nasmejan	140.53391	42	41	all
odgovoran	138.42626	45	38	all
zatvoren	131.94506	46	37	all
prijatan	131.50495	47	38	all
inteligentan	119.74549	48	33	all
različit	119.74549	49	33	all
pokvaren	119.30468	50	34	all
stidljiv	117.11959	51	31	all
zabrinut	116.66458	52	32	all
originalan	114.02620	53	30	all
aktivan	113.57051	54	31	all
osetljiv	113.12929	55	32	all
uspešan	113.12929	55	32	all
zauzet	113.12929	55	32	all
nesrećan	110.91931	58	29	all
pristojan	110.02143	59	31	all
perverzan	107.34126	60	29	all
tolerantan	106.89957	61	30	all
moralan	103.76322	62	29	all
moćan	103.76322	62	29	all
očajan	101.05414	64	27	all

feature	frequency	rank	docfreq	group
ljubazan	97.44506	65	27	all
brutalan	91.06250	66	25	all
napredan	84.61038	67	23	all
agresivan	81.35638	68	22	all
samostalan	78.99542	69	19	all
interesantan	78.52761	70	20	all
bezosećajan	78.08264	71	21	all
bahat	74.78820	72	20	all
mračan	74.78820	72	20	all
savremen	74.78820	72	20	all
suptilan	71.47204	75	19	all
mudar	68.57984	76	17	all
dubokouman	68.13302	77	18	all
glupav	68.13302	77	18	all
sebičan	68.13302	77	18	all
spontan	68.13302	77	18	all
nestrpljiv	64.76985	81	17	all
površan	64.76985	81	17	all
pažljiv	60.60660	83	10	all
talentovan	57.96523	84	15	all
opterećen	54.97096	85	13	all
nasilan	54.52037	86	14	all
miroljubiv	51.04446	87	13	all
ironičan	47.53511	88	12	all
vulgaran	47.53511	88	12	all
aseksualan	47.53511	88	12	all
napet	43.98952	91	11	all
ambiciozan	43.98952	91	11	all
grub	43.98952	91	11	all
okrutan	40.40440	94	10	all
tradicionalan	40.40440	94	10	all
nesiguran	40.40440	94	10	all
optimističan	40.40440	94	10	all
nežan	36.77578	98	9	all
dominantan	36.77578	98	9	all

feature	frequency	rank	docfreq	group
rezervisan	36.77578	98	9	all
komunikativan	36.77578	98	9	all
samouveren	36.77578	98	9	all
baksuzan	36.77578	98	9	all
ljigav	33.09880	104	8	all
hladnokrvan	33.09880	104	8	all
savestan	33.09880	104	8	all
primitivan	33.09880	104	8	all
posesivan	33.09880	104	8	all
skeptičan	33.09880	104	8	all
vedar	33.09880	104	8	all
neuredan	33.09880	104	8	all
stabilan	33.09880	104	8	all
ravnodušan	33.09880	104	8	all
sujeveran	33.09880	104	8	all
umeren	33.09880	104	8	all
racionalan	33.09880	104	8	all
ubedljiv	33.09880	104	8	all
umiljat	33.09880	104	8	all
sentimentalan	33.09880	104	8	all
srčan	29.36739	120	7	all
svestran	29.36739	120	7	all
intelektualan	29.36739	120	7	all
poštovan	29.36739	120	7	all
ranjiv	29.36739	120	7	all
ogorčen	29.36739	120	7	all
zloban	29.36739	120	7	all
plemenit	25.57373	127	6	all
operativan	25.57373	127	6	all
poslušan	25.57373	127	6	all
povučen	25.57373	127	6	all
sirov	25.57373	127	6	all
istrajan	25.57373	127	6	all
radostan	25.57373	127	6	all
sujetan	25.57373	127	6	all

feature	frequency	rank	docfreq	group
beskoristan	25.57373	127	6	all
borben	25.57373	127	6	all
zaljubljiv	25.57373	127	6	all
pitom	22.19190	138	4	all
preosetljiv	21.70735	139	5	all
temeljan	21.70735	139	5	all
srdačan	21.70735	139	5	all
nedokazan	21.70735	139	5	all
uvredljiv	21.70735	139	5	all
veseo	21.70735	139	5	all
atraktivan	21.70735	139	5	all
pristrasan	21.70735	139	5	all
mio	21.70735	139	5	all
promašen	21.70735	139	5	all
diskretan	21.70735	139	5	all
anksiozan	21.70735	139	5	all
živčan	18.95764	151	2	all
dostojanstven	17.75352	152	4	all
bezgrešan	17.75352	152	4	all
druželjubiv	17.75352	152	4	all
zamišljen	17.75352	152	4	all
elokventan	17.75352	152	4	all
nerazuman	17.75352	152	4	all
prilagodljiv	17.75352	152	4	all
religiozan	17.75352	152	4	all
zaostao	17.75352	152	4	all
vickast	17.75352	152	4	all
konzervativan	17.75352	152	4	all
žestok	17.75352	152	4	all
luckast	14.21823	164	2	all
čedan	13.68996	165	3	all
superioran	13.68996	165	3	all
slatkorečiv	13.68996	165	3	all
pedantan	13.68996	165	3	all
strog	13.68996	165	3	all

feature	frequency	rank	docfreq	group
impulsivan	13.68996	165	3	all
umišljen	13.68996	165	3	all
setan	13.68996	165	3	all
nepoverljiv	13.68996	165	3	all
snažan	13.68996	165	3	all
bespomoćan	13.68996	165	3	all
izopačen	13.68996	165	3	all
apolitičan	13.68996	165	3	all
frustriran	13.68996	165	3	all
načitan	13.68996	165	3	all
slobodouman	13.68996	165	3	all
lažljiv	13.68996	165	3	all
prefinjen	13.68996	165	3	all
bezazlen	13.68996	165	3	all
velikodušan	13.68996	165	3	all
besraman	13.68996	165	3	all
usiljen	10.08088	186	1	all
bezbrižan	9.47882	187	2	all
oprezan	9.47882	187	2	all
povodljiv	9.47882	187	2	all
problematičan	9.47882	187	2	all
samokritičan	9.47882	187	2	all
snalažljiv	9.47882	187	2	all
inertan	9.47882	187	2	all
maštovit	9.47882	187	2	all
zastrašujući	9.47882	187	2	all
izdržljiv	9.47882	187	2	all
isključiv	9.47882	187	2	all
zavodljiv	9.47882	187	2	all
gord	9.47882	187	2	all
melanholičan	9.47882	187	2	all
odlučan	9.47882	187	2	all
rasejan	9.47882	187	2	all
buntovan	9.47882	187	2	all
šarmantan	9.47882	187	2	all

feature	frequency	rank	docfreq	group
vešt	9.47882	187	2	all
odmeren	9.47882	187	2	all
ćutljiv	9.47882	187	2	all
erotičan	9.47882	187	2	all
principijelan	9.47882	187	2	all
ciničan	9.47882	187	2	all
neposredan	9.47882	187	2	all
brižan	9.47882	187	2	all
nemaran	9.47882	187	2	all
nepouzdan	9.47882	187	2	all
prevrtljiv	9.47882	187	2	all
dinamičan	9.47882	187	2	all
poletan	9.47882	187	2	all
kompetentan	9.47882	187	2	all
provokativan	9.47882	187	2	all
licemeran	9.47882	187	2	all
kolegijalan	5.04044	221	1	all
neumoljiv	5.04044	221	1	all
hirovit	5.04044	221	1	all
prevaran	5.04044	221	1	all
oštrouman	5.04044	221	1	all
šaljiv	5.04044	221	1	all
zajedljiv	5.04044	221	1	all
vragolast	5.04044	221	1	all
pristupačan	5.04044	221	1	all
pravdoljubiv	5.04044	221	1	all
divalj	5.04044	221	1	all
radoznao	5.04044	221	1	all
priprost	5.04044	221	1	all
suzdržan	5.04044	221	1	all
entuzijastičan	5.04044	221	1	all
nepromišljen	5.04044	221	1	all
ponizan	5.04044	221	1	all
privlačan	5.04044	221	1	all
sažaljiv	5.04044	221	1	all

feature	frequency	rank	docfreq	group
popustljiv	5.04044	221	1	all
bezvoljan	5.04044	221	1	all
haotičan	5.04044	221	1	all
neživljen	5.04044	221	1	all
koristoljubiv	5.04044	221	1	all
škrt	5.04044	221	1	all
prostodušan	5.04044	221	1	all
čuvaran	5.04044	221	1	all
temperamentan	5.04044	221	1	all
sumnjičav	5.04044	221	1	all
konvencionalan	5.04044	221	1	all
učtiv	5.04044	221	1	all
tanan	5.04044	221	1	all
teatralan	5.04044	221	1	all
misaon	5.04044	221	1	all
dovitljiv	5.04044	221	1	all
gramziv	5.04044	221	1	all
galantan	5.04044	221	1	all
zbunljiv	5.04044	221	1	all
svadljiv	5.04044	221	1	all
disciplinovan	5.04044	221	1	all
indiskretan	5.04044	221	1	all
zadrt	5.04044	221	1	all
nadmen	5.04044	221	1	all
ohol	5.04044	221	1	all
drzak	5.04044	221	1	all
pragmatičan	5.04044	221	1	all
zlonameran	5.04044	221	1	all
cmizdrav	5.04044	221	1	all

Table 2

Fifteen topics based on Serbian 383 adjectives: coefficients' differences

topics	Griffiths_plus_Deveaud_std	CaoJuan_plus_Arun_std	difference
15	-0.69	-0.25	-0.44
14	-0.01	-0.64	0.63
13	-0.12	-0.54	0.42
12	0.57	-1.10	1.67
11	0.49	-0.77	1.26
10	-0.33	0.28	-0.62
9	1.32	-0.65	1.98
8	0.82	0.02	0.80
7	1.53	-1.40	2.93
6	0.10	-0.08	0.18
5	0.46	0.55	-0.09
4	-1.30	2.19	-3.49
3	-0.66	1.27	-1.93
2	-2.18	1.11	-3.29

Table 3

Fifteen topics based on Goldberg's adjectives: coefficients' differences

Topics	Griffiths_plus_Deveaud_std	CaoJuan_plus_Arun_std	difference
15	-0.28	0.45	-0.73
14	-0.14	-0.94	0.81
13	-1.50	1.20	-2.70
12	0.98	-0.59	1.57
11	-1.25	0.85	-2.10
10	0.60	-0.13	0.73
9	-0.11	-0.50	0.38
8	0.37	0.60	-0.23
7	0.53	-0.70	1.23
6	1.03	-0.30	1.33
5	2.06	-2.25	4.32
4	-1.18	1.36	-2.54
3	-0.65	-0.16	-0.49
2	-0.46	1.10	-1.56

Table 4

Serbian topics and word categories: Overall category distribution per topic

Categories	Serbian Topic1	Serbian Topic2	Serbian Topic3	Serbian Topic4	Total
Abilities, talents or their absence (f)	4.00	5.00	1.00	6.00	16.00
Emotions, moods and cognitions (f)	8.00	9.00	6.00	11.00	34.00
Pure evaluation (f)	9.00	7.00	8.00	10.00	34.00
Roles and relationships (f)	7.00	7.00	7.00	8.00	29.00
Social effects: reactions of others (f)	1.00	6.00	2.00	4.00	13.00
Social status (f)	4.00	0.00	1.00	0.00	5.00
States and activities (f)	1.00	0.00	2.00	0.00	3.00
Temperament and character traits (f)	31.00	29.00	40.00	26.00	126.00
Value orientations (f)	2.00	4.00	0.00	2.00	8.00
Abilities, talents or their absence (p)	0.01	0.02	0.00	0.02	0.05
Emotions, moods and cognitions (p)	0.03	0.03	0.02	0.04	0.12
Pure evaluation (p)	0.03	0.03	0.03	0.04	0.13
Roles and relationships (p)	0.03	0.03	0.03	0.03	0.12
Social effects: reactions of others (p)	0.00	0.02	0.01	0.01	0.04
Social status (p)	0.01	0.00	0.00	0.00	0.01
States and activities (p)	0.00	0.00	0.01	0.00	0.01
Temperament and character traits (p)	0.12	0.11	0.15	0.10	0.48
Value orientations (p)	0.01	0.01	0.00	0.01	0.03

Note. (f) - frequency; (p) - proportion

Table 5

Serbian topics and word categories: Category by topic

Categories	Serbian Topic1	Serbian Topic2	Serbian Topic3	Serbian Topic4
Abilities, talents or their absence (p)	0.06	0.07	0.01	0.09
Emotions, moods and cognitions (p)	0.12	0.13	0.09	0.16
Pure evaluation (p)	0.13	0.10	0.12	0.15
Roles and relationships (p)	0.10	0.10	0.10	0.12
Social effects: reactions of others (p)	0.01	0.09	0.03	0.06
Social status (p)	0.06	0.00	0.01	0.00
States and activities (p)	0.01	0.00	0.03	0.00
Temperament and character traits (p)	0.46	0.43	0.60	0.39
Value orientations (p)	0.03	0.06	0.00	0.03

Note. (f) - frequency; (p) - proportion