Individual Differences in the Most Frequent Content Word Usage as a New Type of Features in the Authorship Profiling Task

**Tatiana Litvinova**

Voronezh State Pedagogical University, Russian Federation, centr\_rus\_yaz@mail.ru, Açıklama: ORCID <https://orcid.org/0000-0002-6019-3700>

**Viktoriya Zavarzina**

Voronezh State Pedagogical University, Russian Federation, zva0604@yandex.ru, Açıklama: ORCID [https://orcid.org/0000-0001-6484-4606](https://orcid.org/XXXX-XXXX-XXXX-XXXX)

**Polina Panicheva**

Voronezh State Pedagogical University, Russian Federation, ppolin86@gmail.com, Açıklama: ORCID <https://orcid.org/0000-0002-9229-6456>

**Abstract**: Authorship profiling is a task of revealing an author’s characteristics (i.e., gender, age, personality traits, etc.) of a text based on the analysis of linguistic features. This task is not only a purely theoretical but also a practical one. Identifying the characteristics of text authors is a task of great importance in marketing, sociology, forensics, etc. The task of authorship profiling is often approached as that of text classification or clustering. Different types of features have been introduced – lexical, morphological, syntactical ones, etc. In recent years, deep learning (DL) architectures have frequently been applied, along with traditional machine learning methods. However, for AP DL approaches regularly underperform, left behind by classical machine learning approaches, which indicates the complexity of the task. Also, interpretability is the key demand for authorship profiling methods limiting the use of DL methods for AP in real life. There is a need in more sophisticated but interpretable features for AP. In this paper, we propose a completely new type of features for this task – semantic characteristics of the contexts of the most frequent content words extracted using word embedding model, semantic relation extraction method and hand-crafted set of variables reflecting different aspects of word meaning. We present the results of the experiments where we applied this type of features to the texts of two genres extracted from the RusIdiolect dataset for the detection of gender and Big-5 personality traits. We discuss the advantages of this type of features as well as their limitations and further directions of improving the proposed methodology.

**Keywords:** Authorship profiling, Semantics, Word embeddings, Personality detection.

**Citation:** Litvinova, T., Zavarzina, Z., & Panicheva, P. (2024). Individual Differences in the Most Frequent Content Word Usage as a New Type of Features in the Authorship Profiling Task. In M. Shelley & O. T. Ozturk (Eds.), *Proceedings of ICRES 2024-- International Conference on Research in Education and Science* (pp. XXX-XXX), Antalya, Turkiye. ISTES Organization.

**Introduction**

Authorship profiling is a task of revealing the characteristics (i.e., gender, age, emotions, traits, etc.) of an author of a text based on the analysis of linguistic features. This task is not only a purely theoretical but also a practical one. Identifying the characteristics of text authors is a task of great importance in marketing, sociology, forensics, etc. Although this problem is not a new one, with the emergence of a great number of texts (mostly on social media) it has become one of the most important and widely studied issues in the area of text analysis. A series of hackathons has been devoted to this task in its different aspects (PAN events, [https://pan.webis.de](https://pan.webis.de/)) (Potthast et al., 2019). Nowadays, this specific task has gained incredible importance due to the appearance of new risks introduced by generative AI and necessity of research into its ability to mimic an author’s style and characteristics as well as an urgent need in the development of the models aimed at detecting texts generated by AI.

The task of authorship profiling is often approached as that of text classification or clustering. Different types of features have been introduced – lexical, morphological, syntactical ones, etc. Very naïve features such as word unigrams and bigrams and character 3-4-5-grams are the most popular in traditional machine learning approaches (HaCohen-Kerner, 2022). In recent years, deep learning (DL) methods have frequently been applied, along with traditional machine learning methods (HaCohen-Kerner, 2022). Controversial results have been reported regarding the superiority of this or that methodology for different tasks of text classification including authorship profiling, and lots of researchers highlight the superiority of the traditional machine learning approaches (López-Santillán et al., 2023). Also, one of the drawbacks of deep learning methods are the difficulties with the interpretability of the results they provide. For some type of tasks this is not crucial, but it is not the case when we reconstruct the authors behind the texts where we are seeking to identify not only an author's characteristics but also which features of a text are associated with them (i.e., why did the model choose to classify this text as one written by a female?). This aspect of the authorship profiling task has often been overlooked in recent papers which have used deep learning methods making the results of the constructed models less applicable in the domains where the explanation is key.

In this paper, we discuss the results of the experiments aimed at profiling the authors of texts using a new type of features, namely the semantic characteristics of the contexts of the most frequent content words. It is well known that there are some differences in word meanings in an individual’s mental lexicon. Based on this fact, we extracted the contexts of the most frequent content words from a corpus of texts in Russian RusIdiolect supplied with the author metadata (Litvinova, 2020), calculated the semantic characteristics of the contexts of the selected words over a set of features using Concept Mover’s Distance (Stoltz & Taylor, 2019) and pretrained a language model word2vec. Namely, we were able to measure the most frequent content words’s engagement with a set of the concepts using distributional representations of the meaning of words. While constructing a set of the concepts, we relied on the features that had been shown to represent the core modalities of information processing in the neuroimaging literature (Binder et al., 2016). For example, we calculated the engagement of the contexts of the most frequent content words with sensory and motor experiences (e.g., shape and motion) and affective experiences (e.g., happy and sad) using the pretrained language model. Then, using these features, we built the classification models to predict the gender and psychological traits of the authors of texts.

In our paper, we will discuss which content words and semantic features of their contexts of usage contributed most to the results of the predictive models for gender and personality detection. We will also look at the advantages and some limitations of this new type of interpretable features for the authorship profiling task.

**Methods**

**Material description**

*Text data*

Collection of the datasets for personality detection is difficult since it requires recruiting participants, processing of the results of psychological questionnaires, etc. Moreover, not only personality test data but also user-generated texts should be analyzed, which is often not the case (especially when social media data is analyzed). However, recent studies (Zhou et al., 2020) have confirmed that small high-quality datasets perform better than large low-quality ones in the model training of personality testing tasks. With this in mind, we used the carefully constructed dataset RusIdiolect (Litvinova, 2020) which contains both personality test scores and texts written by the respondents in the course of the data collection process.

We analyzed texts of two different genres – picture description and letter to a friend. Examples of both text type are given below.

Picture Description: На картине изображены две женщины. На переднем плане изображена молодая женщина с короткой стрижкой. Грустными глазами она смотрит в сторону. Как будто провожая взглядом уходящего человека. За ней стоит старушка с белыми седыми волосами, покрытыми темным платком. Правую руку она держит подперев подбородок. Глаза как будто смотрят в другую сторону. С первого взгляда лицо женщины очень страшное, все покрыто складками и морщинами. То ли она осуждает эту женщину, то ли сочувствует ей. Понять это не возможно. А если присмотреться к старушке повнимательней, то начинаешь испытывать симпатию. Милая старушка слегка улыбается. Напряженный взгляд молодой женщины провожает человека, смотря ему вслед. Но лицо остается спокойным и без эмоций. Ей все равно уйдет этот человек или нет.

The painting shows two women. In the foreground there is a young woman with short hair. She looks away, her eyes are sad. As if watching someone leaving. Behind her stands an old woman with white gray hair covered with a dark scarf. She holds her right hand under her chin. The eyes seem to be looking in the other direction. At first glance, the woman’s face is very scary, everything is covered with folds and wrinkles. Either she condemns this woman, or she sympathizes with her. It is impossible to understand. And if you take a closer look at the old woman, you begin to feel sympathetic. The sweet old lady is faintly smiling. The young woman’s intense gaze follows the man she is watching leaving. But her face remains calm and without emotion. She doesn’t care whether this person leaves or not.

Letter to a friend: Привет, Данил! Как дела? Я надеюсь, что все хорошо. Последний месяц был очень напряженным и интересным. Я получил права на вождение автомобиля. Я целых полгода ходил на занятия. Это было утомительно, но это того стоило. Теперь я могу управлять авто. Это помогает делать много дел в один день. Побывать в сотнях новых мест, узнать много новых людей. Я понял смысл поговорки: «автомобиль не роскошь, а средство передвижения». И это правда! Учеба дается мне легко. Наша группа очень веселая и сильная. Мы сдаем завтра зачет по информатике. Через месяц у меня сессия. Немного волнуюсь, но да ладно! Расскажи о себе, мне все интересно. Жду ответа! Я знаю, что ты не любишь писать, но надеюсь на ответ. Может, позвонишь, мне будет приятно услышать твой голос. Мы давно не разговаривали по телефону. Твой номер не изменился? Или ты пользуешься только мобильным? Не пропадай. Пока!

Hi Danil! How are you? I hope everything is OK. The last month has been very busy and interesting. I have received my driving license. I have been training for six months. It has been tiring, but it has been worth it. Now I can drive a car. This helps to do a lot of things in a day. Visit hundreds of new places, meet a lot of new people. I have understood the meaning of the saying, “A car is not a luxury, but a means of transportation.” And it is true! Studying comes easy to me. Our group is very cheerful and strong. We are taking a computer science test tomorrow. I am having an exam session in a month from now. I'm a little worried, but oh well! Tell me about yourself, I'm interested in everything. Can’t wait to hear back from you! I know you don't like to write, but I hope to get your answer. Maybe you'll call me, I'll be glad to hear your voice. We haven't talked on the phone for a long time. Has your number changed? Or do you only use mobile? Stay in touch! Bye!

The mean text length is 152 (sd=55.6).

*Participant metadata*

RusIdiolect dataset contains the personality scores obtained with a set of questionaires. For this particular study, we used scores on the *Big-5* test as widely used in personality computing research (Gregory et al., 2015). The theoretical foundation of the questionnaire is the five-factor personality model. Five fundamental factors that help to characterize the structure of a person’s personality have been identified as a result of scientific research by psychologists over a few decades. They include extraversion, agreeableness, conscientiousness, neuroticism, openness to experience. This framework does not imply that personality can be described by merely five factors. Rather, these five dimensions represent personality at a high level of abstraction. Each of the five factors includes a large number of smaller specific characteristics.

Initially, the Big Five was created in English (Goldberg, 1981), but the authors emphasize the universality of this model for a lot of languages and cultures, which is confirmed by empirical research.

While creating RusIdiolect materials, the 5PFQ version of the “Big Five” was used which was translated and adapted in 1995-1999 to domestic conditions of the social environment by psychologists of Kurgan State University (Hromov, 2000) as the only one available at the first stage of the experiment (2006). This version was subsequently used to compare the results.

A high level of extraversion in the interpretation of this test is associated with activity, dominance, sociability, and the search for new experiences; a low level is associated with passivity, subordination, isolation, and avoidance of impressions and attention. Extraversion in this version of the test carries a more significant connotation of dominance than in the other adaptations.

According to the second factor, agreebleness contrasts warmth, cooperation, trustfulness, understanding, respect for others, on the one hand, and indifference, suspicion, misunderstanding, self-respect. There is a more pronounced collectivist attitude in the 5PFQ construct where a greater emphasis is placed not on the test taker, as such but rather on their actions in relation to group members.

The third factor includes components such as accuracy, perseverance, responsibility, self-control, forethought, which are contrasted with sloppiness, irresponsibility, impulsiveness, carelessness, and lack of persistence.

High scores on the fourth factor are associated with such characteristics as anxiety, depression, self-criticism, and emotional lability; low scores are associated with carefree, relaxed, emotional comfort, self-sufficiency, and emotional stability.

High scores on the fifth factor are associated with curiosity, daydreaming, artistry, sensitivity, and plasticity; low scores are associated with conservatism, realism, lack of artistry, insensitivity, and rigidity.

The minimum number of points scored for any main factor is 15, the maximum number is 75. Conventionally, the points can be divided into high (51-75 points), average (41-50 points) and low (15-40 points). In accordance with this, we initially divided the respondents into 3 classes. However, in total for each factor, the number of respondents with low and average ratings was lower than or approximately equal to that of the respondents with high ratings. Therefore in order to make our dataset more balanced, which allows for higher accuracy, we combined individuals with low and average scores into one LowMedium group. The composition of the dataset is shown in Table 1.

Table 1. Participant metadata

|  |  |  |  |
| --- | --- | --- | --- |
| Variables/Register | | Picture description | Letter to a friend |
| Gender | Male | 95 | 75 |
| Female | 166 | 128 |
| Extraversion | LowMedium | 108 | 86 |
| High | 153 | 117 |
| Agreebleness | LowMedium | 101 | 69 |
| High | 160 | 134 |
| Сonscientiousness | LowMedium | 89 | 66 |
| High | 172 | 137 |
| Neuroticism | LowMedium | 132 | 106 |
| High | 129 | 97 |
| Openness | LowMedium | 48 | 35 |
| High | 213 | 168 |
| Total |  | **261** | **203** |

*Feature construction*

At the first stage of analysis, texts were lowecased and lemmatized using udpipe package R. We also used special token for emoji and left it for further analysis. After that, we selected the most frequent content words whose contexts to be analyzed further. We set the minimal document frequency (for each register) where such word occurs to 50 (i.e. in no less than in 20 % of the text).

The resulting list of selected words forpicture description: всё\_PRON “all”, мочь\_VERB “can”, дочь\_NOUN “daughter”, лицо\_NOUN “face”, девушка\_NOUN “girl”, бабушка\_NOUN “granny”, парень\_NOUN “guy”, человек\_NOUN “human”, жизнь\_NOUN “life”, мужчина\_NOUN “man”, я\_PRON “me”, мать\_NOUN “mother”, свой\_DET “own”, возможно\_ADV “possible”, смотреть\_VERB “see”, сторона\_NOUN “side”, взгляд\_NOUN “sight”, думать\_VERB “think”, время\_NOUN “time”, очень\_ADV “very”, хотеть\_VERB “want”, мы\_PRON “we”, весь\_DET “whole”, женщина\_NOUN “woman”, молодой\_ADJ “young”.

Resulting list of selected words forletter to a friend: человек\_NOUN “man”, я\_PRON “me”, свой\_DET “own”, очень\_ADV “very”, думать\_VERB “think”, жизнь\_NOUN “life”, весь\_DET “whole”, время\_NOUN “time”, хотеть\_VERB “want”, всё\_PRON “all”, новый\_ADJ “new”, уже\_ADV “already”, мой\_DET “my”, любить\_VERB “love”, надеяться\_VERB “hope”, давно\_ADV “longago”, много\_ADV “many”, писать\_VERB “write”, ☺ “emoji”, ждать\_VERB “wait”, your “твой\_DET”, друг\_NOUN “friend”, последний\_ADJ “last”, видеться\_VERB “see”, учеба\_NOUN “study”, хорошо\_ADV “nice”, скоро\_ADV “soon”, день\_NOUN “day”, хотеться\_VERB “want2”, учиться\_VERB “learn”, соскучиться\_VERB “miss1”, скучать\_VERB “miss2”, работа\_NOUN “job”)

As a unit of our analysis, we used the contexts of the selected word. We experimented with different window sizes but for the final experiments n=3 (i.e., 3 words before and after key word) were chosen taking into account the length of the documents.

To construct our feature set, we used methods for semantic relation extraction from the texts using word embeddings. Semantic relations are modeled as vector translations in a word embedding space (Stoltz et al., 2024).

Namely, we used text2map function CMDist() (Stoltz & Taylor, 2019) which document-term matrix (DTM) as input, a matrix of word embedding vectors, and concept words or concept vectors. The function uses word counts from the DTM and word similarities from the cosine similarity of their respective word vectors in a word embedding model. The "cost" of transporting all the words in a document to a single vector or a few vectors (denoting a concept of interest) is the measure of engagement, with higher costs indicating less engagement. For intuitiveness the output of CMDist() is inverted so that the higher numbers will indicate more engagement with a concept of interest.

It is possible not only to specify single words from the word embeddings, but also a list of them. In this case the function (get\_centroid) takes a vector extracted from the embedding space in the form of a centroid (which averages the vectors of several words). It is possible to use the offset of several juxtaposing words using a function get\_direction() to extract the engagement of a text with this or that pole of the scale. A widely used example is the gender spectrum, from more feminine to more masculine. Therefore, this method has been used extensively to measure the extent target terms are biased toward man or woman. The words need not be in the DTM, but they must be in the word.

As was shown in a number of papers, different types of psycholinguistic relations could be extracted from word embeddings, e.g., abstractness, sentiment, etc. (see Stoltz et al., 2024). Therefore, we might construct, using existing psycholinguistic databases, a concept consisting of words with a high/low level of abstractness, sentiment, etc. and extract the level of abstractness of the texts.

A different set of concept words was tested, and their particular content was dependent on the task the researcher was solving (more on that: Litvinova & Panicheva, 2024). We propose a novel approach for the construction of these feature set. As our task is to perform complex analysis of the contexts, we set the task to describe them over the comprehensive set of features however theoretically motivated. To do this, we relied on the data from neurobiology about the processing of different type of words depending of their semantics. Specifically, we used the data from Binder et al. (2016) where a brain-based semantics consisting of conceptual primitives defined in terms of the modalities of neural information processing proposed.

This study aimed at developing a representation that captured the aspects of the experience that are central in the acquisition of concepts. The authors organized human experience in 13 different domains - each one corresponding to a variable number of features for which some specialized neural processor has been identified and described in the neuroscientific literature (Binder et al. 2016). In this work, an example of the words rated high on these features was proposed. We used both translations of these words to construct our dimensions and also the data from psycholinguistic databases. E.g., for constructing the dimensions which are related to the visual modality we used the database with a human rating of different words for this modality from database created by Miklashevsky (Miklashevsky, 2018), Russian version of LIWC thesaurus (Pennebaker et al., 2015) was used (Panicheva & Litvinova, 2020), etc.

We also preliminary constructed the list of basic semantic oppositions. They were constructed based on the word meaning components presented in Binder et al. 2016, lists of semantic differentials presented in literature (see Litvinova & Panicheva 2024 for more on this) as well as using a dictionary of Russian antonyms alighned with the set of the most frequent Russian words.

The resulting feature set is presented on github (<https://github.com/Litvinova1984/ICRES2024> ).

We used a pretrained model ruwikiruscorpora\_upos\_cbow\_300\_10\_2021 which was trained on Wikipedia and NCR in late 2021.

DTM on the contexts of the selected words was constructed using quanteda package. Thus, for each contexts of the selected words, 46 features were extracted. Resulting matrix for PIC dataset was 261x1159, for LETTER dataset – 203х1527.

*Exploratory analysis and classification approach techniques*

As our major goal is not only to obtain high quality classification/regression models, but also to search for the most significant features, taking into account the situation when the number of features is greater than that of the samples, as well as possible multicollinearity of the features (which is typical for linguistic features), we used a set of multivariate methods analysis implemented in the mixomics library and specifically designed for such a data set.

We performed PCA for exporatory analysis using both mixomics and Factominer packages and functions of this package allowing us to establish connections between the selected combined features (components) and the qualitative and quantitative characteristics of respondents (which, according to the terminology adopted by those who designed the package, are called additional qualitative (quantitative) variables, i.e., they serve to interpret the results, but are not directly used in the analysis itself). Supplementary variables have no influence on the principal components of the analysis. They are going to help to interpret the dimensions of variability. We can add two different kinds of variables: continuous ones and categorical ones. To facilitate interpretation, we used functions dimdesc and catdes of **Factominer.**

The dimdesc() function calculates the correlation coefficient between a variable an a dimension and performs a significance test. These tables give the correlation coefficient and the p-value of the variables which are significantly correlated to the principal dimensions. Both active and supplementary variables whose p-value is smaller than 0.05 appear.

For both text types, as supplementary continuous variables, we would like to add the variables linked to the scores on 5 personality traits from Big 5 and age. As supplementary categorical variable, we used gender.

For the classification (in the case of a binary variable - the gender of the author of the text), we used the algorithm Partial Least Squares – Discriminant Analysis (PLS-DA) which operates efficiently over large dataframes and is not negatively influenced by collinearity.

While evaluating the classification performance of (s)PLS-DA models, repeated cross-validation is used. We used 5-fold cross-validation with 50 repeats for each model as recommended by the package developers.

We used a balanced error rate (BER) as a measure for the model efficiency. All the tasks (both gender and personality traits detection) were approached as a binary classification problem for the purpose of the comparison with SOTA approaches in other languages and with the aim of constructing more balanced datasets.

**Results and Discussion**

**PIC dataset**

PCA was performed to summarize the data with a multivariate approach and to provide a visual representation of the distances between the texts. Preliminary analysis with elbow method and showed that 5 components were optimal and was used for further experiments. One-way ANOVA models were constructed including the Principal Components as the response variables and gender and personality traits as the explanatory variable; coefficients of determination (R2) and related p-values were calculated as well as the categories coefficient estimates, tested for significant differences from zero (α = .005). Fig. 1 shows the tendency of division of the texts by male and female authors over Dim. 1, which is proved by statistical test (R2 = 0.0543, p=0.0001; estimates for gender: Female =1.976211, p= 0.0001, Male = -1.976211, p= 0.0001).

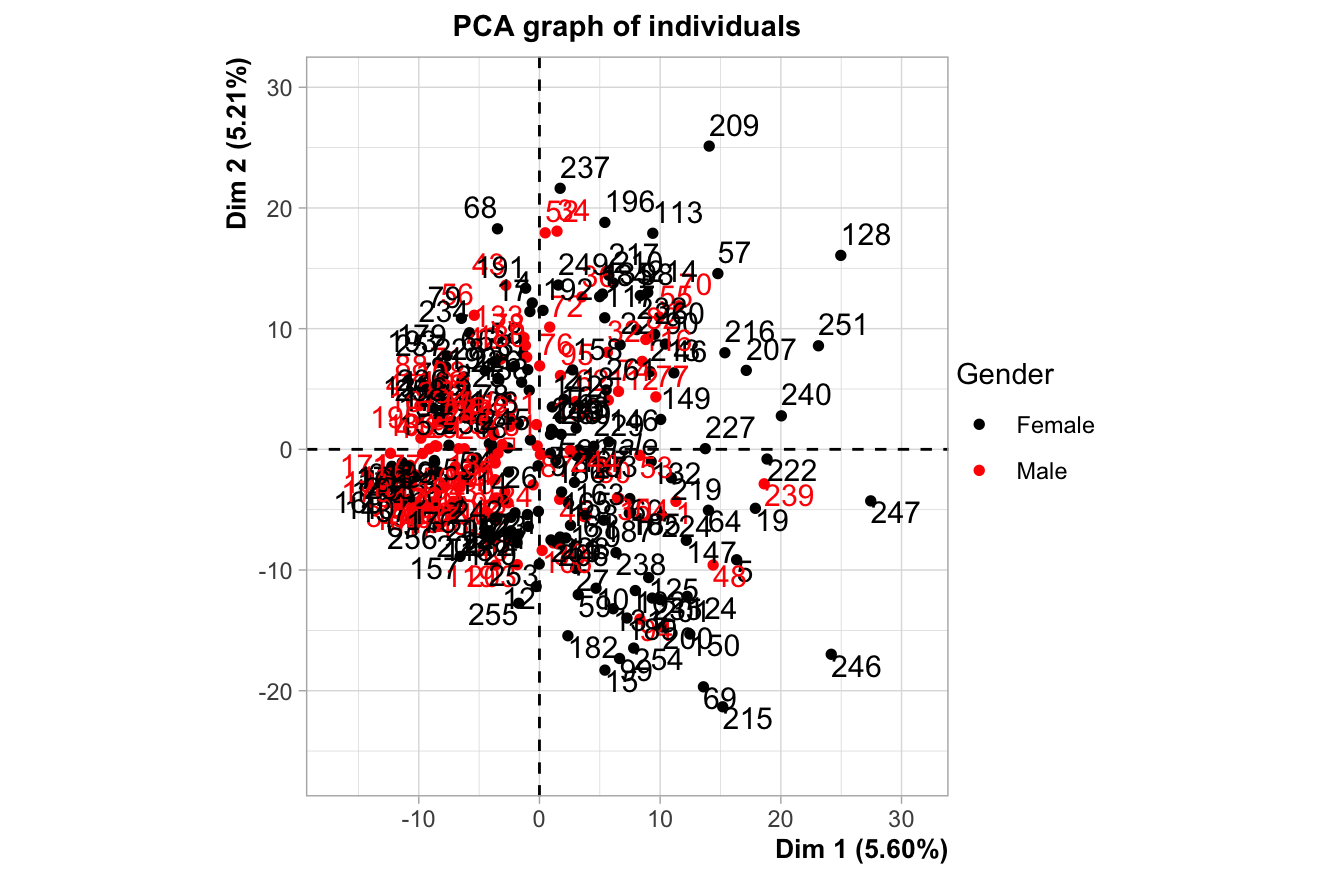


Fig. 1. PCA on PIC dataset

The following features were positively correlated (0.5 and higher, p<0.000000001) with Dim 1: motherSocialSelf, motherSocialLIWC, motherGustNorms, motherSpatialProx, motherVisLIWC, motherAudLIWC, motherSomatNociception, motherCausal, motherVisBody, daughterSocialSelf, daughterSocialLIWC, daughterSpatialProx, daughterSomatNociception (according to estimate, this values are higher in female texts). Negative correlations are registered for motherEmoAngry, motherGustTaste, motherSocialGender, motherEmoSentiment, motherDrive, motherMotorPractice, motherEmoDisgust, daughterSocialGender, daughterEmoAngry, daughterMotorPractice, motherAttentionArousal, motherAudIntens, daughterGustTaste, motherCognitionImage, motherEmoHappy, canSomatTexture (these features have higher levels in texts which are located to the left on Dim 1, i.e. most texts by male and some female texts).

*Gender classification*

For the classification, initially we performed PLS-DA with 10 PCs. For each component, repeated cross-validation (10 ×5−fold CV) is used to evaluate the PLS-DA classification performance (overall and balanced error rate BER), for each type of a prediction distance; max.dist, centroids.dist and mahalanobis.dist). The smallest error rate (both overall and balanced) was obtained for centroids.dist (0.39) and comp=3. They were used for the final model.

Then we constructed a sparse model with the aim to reduce the feature size in mind . We set list.keepX <- seq(50, 200, 10)). We estimate the classification error rate with respect to the number of the selected variables in the model with the function tune.splsda(). The tuning is being performed on one component at a time inside the function and the optimal number of variables to select is automatically retrieved after each component run.

Previously, we determined the number of components to be ncomp = 3 with PLS-DA. Here we set ncomp = 4 to further assess if this was be the case for a sparse model, and use 5-fold cross validation repeated 10 times. We also choose the centroid prediction distance. A model with comp =1 and 200 variables were automatically selected with $BER = 0.401. It is interesting that the female authors were detected at a lower level than those by males (0.512 and 0.273, respectively) which is supported by the visualization (the male texts are more homogeneous) (see Fig. 2).

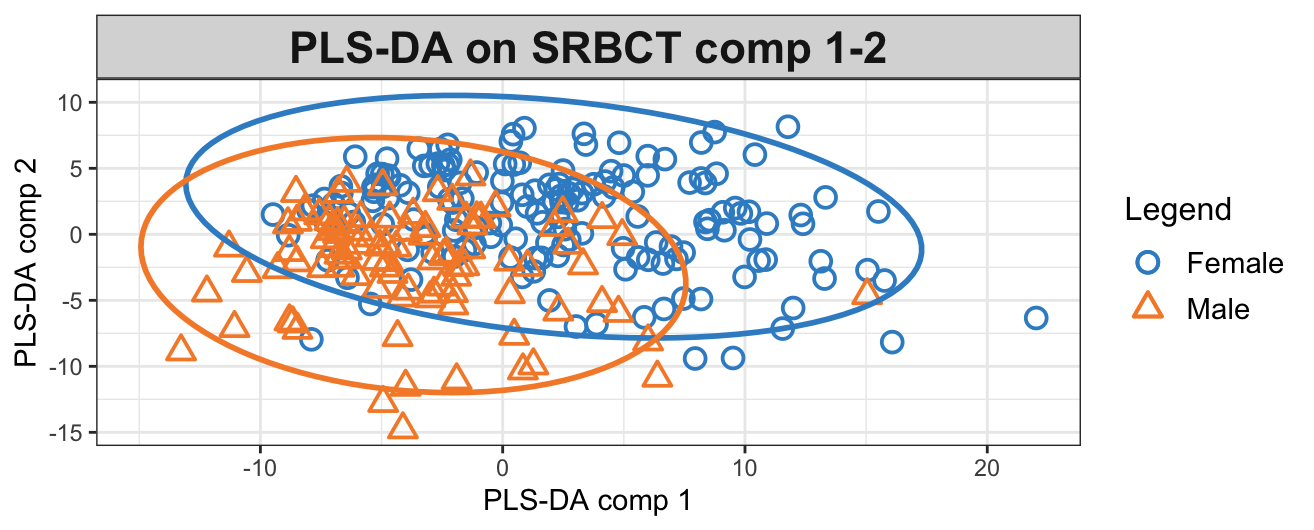


Fig. 2. Visualization of PLS-DA on PIC dataset

In the plot in Fig. 3, the loading weights of each selected variable on each component are represented. The colours indicate the group in which the expression of the selected features is maximal based on the mean value.



Fig. 3. Feature importance for Gender, PIC dataset   
(blue lines indicate the female texts, orange lines indicate the male texts)

We can see that the most important features related to the differences in the meaning of words mother, own, life, time, can, daughther.

*Big 5 classification*

At the first stage of the analysis, we calculated the correlations (p<0.05) between the Big 5 scores and features. The extraversion scores are correlated with Openness at a rather high level (0.58), Agreeableness (0.36) and Conscientiousness (0.25). Among linguistic features different components of the meaning of word me are positively correlated with the scores on extraversion (meDriveNeeds, meCausal, meSocialLIWC, meTemporalLIWC, meCognitionLIWC, meSomatNociception, meVisBody, meSocialSelf, meAudLIWC, meEmoFear), side (sideGustNorms, sideSpatialNumber, sideSocialSelf, sideSomatProprioception, sideTemporalLIWC, sideEmoPleasant, sideCausal, sideSpatialProx, sideSomatLIWC, sideVisColor, sideCognitionLIWC), very (veryOlfacNorms, veryVisColor), as well as life (lifeEmoSentiment), while negative correlations were observed for life features lifeVisFace, lifeEmoPleasant, lifeEmoBenefit, lifeEmoHappy, side features sideDrive, sideAttentionArousal, sideEmoAngry, sideGustTaste, *me* meCognitionImage, meEmoAngry, *very* veryAudNorms.

For all the models related to psychological traits, we set ncomp=10, then tuned the models using the procedure recommended in mixOmics tutorial.

The lowest BER for extraversion = 0,35 was obtained with max.dist and comp=1 (Fig. 4).



Fig. 4. Feature importance for Extraversion, PIC dataset

Agreeableness correlated positively with Openness (0.45), Conscientiousness (0.41), Extraversion (0.36). Linguistic features positively correlated with Agreebleness are the characteristics of mother motherSomatNociception, motherTemporalLIWC, motherEmoFear, motherAudLIWC, motherCausal, motherSocialLIWC (negatively with motherMotorPractice, motherDrive, motherEmoAngry, motherEmoSurprised, motherEmoBenefit), can (canEmoFear, canSocialLIWC, canSocialSelf, canGustNorms) (negatively – canEmoAngry, canDrive, canEmoSentiment, canAttentionArousal), as well as different characteristics of want (VisSize, wantVisIntens, wantDriveNeeds, negatively – wantMotorPractice), *human* (EmoHappy), *own* (ownSpatialUpDown, ownVisIntens, negatively - ownMotorPractice), *whole* (wholeAttentionArousal; negatively – wholeEmoBenefit, wholeCausal), granny (grannyDriveNeeds), face (faceAudNorms), guy (guyVisMotion), life (lifeTempDuration) (negatively – lifeSomatNorms).

As for Agreeableness, the minimum BER using centroid.dist and mahalanobis.dist was 0,36. The best results are obtained with ncomp=2 (Fig. 5).

|  |  |
| --- | --- |
|  |  |

Fig. 5. Feature importance for Agreeableness, PIC dataset

Conscientiousness correlated positively with Agreeableness (0.41), Extraversion (0.25), Openness (0.23). Positive correlations are revealed for *time* timeEmoBenefit, timeEmoPleasant, timeSocialSelf, timeVisLIWC (negatively – timeEmoAngry, timeEmoDisgust, timeDrive), *own* ownVisIntens, ownEmoPleasant, ownSomatSurface (negatively - ownSomatNorms, ownEmoSentiment), *young* youngVisFace, youngVisIntens, youngVisNorms (negatively – youngEmoDisgust, youngCognitionAbstract), *human* humanVisFace, humanSomatSurface, *man* manVisSize, manSomatProprioception, manCognitionLIWC, manSocialSelf, manDriveNeeds, manEmoPleasant (negatively – manMotorPractice, manEmoAngry, manEmoSentiment). The error rate was high for Conscientiousness was 0.4 with ncomp=5 (Fig. 6).

|  |  |
| --- | --- |
|  |  |

Fig. 6. Feature importance for Conscientiousness, PIC dataset

Neuroticism positively correlated withOpenness (0.24) and humanTempDuration (negatively – humanVisColor, humanVisNorms, humanEmoSurprised), weEmoSentiment, weEmoDisgust, weTempDuration (negatively – weVisColor, weEmoPleasant, weEmoHappy, weEmoBenefit), motherVisSize, womanEmoSentiment, womanEmoAngry, womanEmoDisgust (negatively – womanSocialSelf, womanDriveNeeds, womanCognitionLIWC, womanVisNorms), wholeEmoHappy, AllVisIntens, AllSpatialUpDown, AllVisFace (negatively – AllEmoSentiment), daughterDriveNeeds, daughterCognitionLIWC, lifeSomatSurface, lifeGustNorms, lifeEmoPleasant.

The lowest BER for Neouroticism 0.35 was obtained with mahalanobis.dist and 4 components. Loadings on PCs are presented in Fig. 7.

|  |  |
| --- | --- |
|  |  |

Fig. 7. Feature importance for Neuroticism, PIC dataset

Openness correlated with Extraversion (0.58), Agreeableness (0.45), Neuroticism (0.24), Conscientiousness (0.23). Positive correlations were revealed with humanCognitionAbstract, humanTempDuration, humanEmoHappy, youngCognitionLIWC, youngEmoFear, youngCausal, youngAudLIWC, youngSocialLIWC (negatively – youngEmoAngry, youngMotorPractice), meSocialLIWC, meSocialSelf, meSomatProprioception, (negatively – meDrive, meSocialGender), verySomatNorms, verySpatialNumber, veryVisNorms, veryVisColor, ownTemporalLIWC, negatively – wantEmoBenefit.

Openness was the hardest one for the discrimination with a minimal overall ER=0.20 but BER 0.48 (high openness is detected much better, with an individual ER per class 0.12) (Fig. 8).



Fig. 8. Feature importance for Openness, PIC dataset

**Letter dataset**

*Exploratory analysis*

An exploratory analysis via PCA on letter dataset, after a preliminary check, was performed with 4 components (Fig. 9).

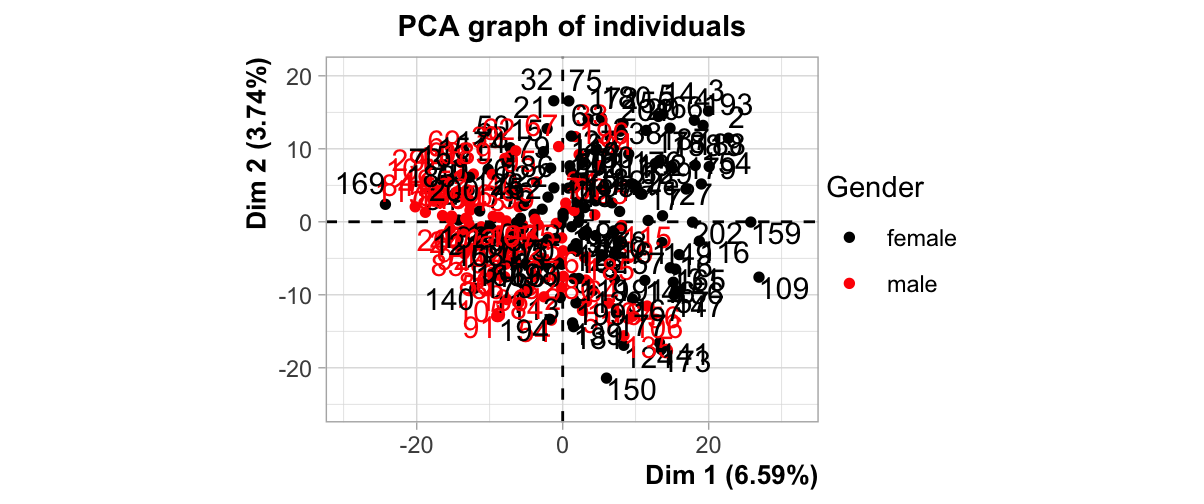


Fig. 9. PCA on Letter dataset

The first dimension is related to the factor “Gender” (R2 = 0.2026373, p<0.00000001), the same is true for Dim 2 (R2 = 0.0244, p= 0.026). Among the features with the highest (among 0.5, p<0.000001) positive correlations with Dim 1 are different characteristics of the word very from RuLIWC (very\_TemporalLIWC, very\_EmoFear, very\_SocialSelf, very\_SocialLIWC, very\_SomatNociception, very\_AudLIWC), whole (VisLIWC, SomatLIWC, SomatNociception, whole\_AudLIWC, whole\_VisBody), love (love\_VisColor, love\_MotorBinder, love\_VisLIWC).

Negative correlations are registered for such characteristics of *very* as very\_MotorPractice, very\_GustTaste, very\_EmoAngry, very\_EmoSurprised, very\_Drive, very\_EmoSentiment, very\_SocialGender), whole (whole\_EmoAngry, whole\_GustTaste, whole\_MotorPractice), *love* (GustTaste, EmoAngry, EmoDisgust, SocialGender, EmoSentiment, Drive). This observation highlights the necessity to take into account not only simple count of words of some semantic groups (LIWC feature, etc.) but also their meaning for individual which could be assessed using the analysis of its contexts.

We can say that, just as for the for PIC dataset, the differences in the meaning of several words are crucial for gender identification.

Lowest BER for Gender was 0.25 (comp=5). The first component highlights the differences in the meaning of words *very, whol*e, as well as *miss*. Component 2 highlights differences in the use of word *all*, *hope* and *emoji* (Fig. 10).

|  |  |
| --- | --- |
|  |  |

Fig. 10. Feature importance for Gender (Letter dataset)

The best sparse model with 200 features yields BER 0.29 which is a worse result than the one obtained on the full feature set. Interestingly, the sparse model selects the characteristics of very, whole, miss1, miss2, all, my, longago, job, miss2.

Let us move to personality trait prediction. The minimal BER for extraversion was 0.35 with ncomp=1.

Extraversion is correlated with emoji features (positively – with SocialLIWC, SomatNociception, AudLIWC, MotorBinder, TemporalLIWC, AudNorms, EmoFear, negatively – with EmoSurprised, EmoAngry, Drive, SocialGender), miss1 (positively – VisBody, SomatLIWC, AudLIWC, EmoHappy, SpatialProx, MotorBinder, negatively – EmoAngry, GustTaste, AudIntens, EmoSurprised, EmoSentiment), very (positively – very\_EmoPleasant, very\_EmoBenefit, very\_SocialSelf, very\_AudLIWC, negatively – with very\_GustTaste, very\_Drive, very\_EmoSentiment, very\_EmoDisgust). The characteristics of these words as well as life, think, learn, study, see contribute most to the classification model (Fig. 11).



Fig. 11. Feature importance for Extraversion (Letter dataset)

The lowest BER for the agreeableness trait was 0.31 (ncomp=2). The most important words for these features (both in terms of correlation and classification) are already (EmoBenefit, DriveNeeds, MotorBinder), job (EmoHappy), life (VisColor), friend, love, think, see (Fig. 12).

|  |  |
| --- | --- |
|  |  |

Fig. 12. Feature importance for Agreeableness (Letter dataset)

Conscientiousness is hard to predict with the overall BER = 0.41 (ncomp=3). Different words contribute to this trait detection with no clear pattern (friend, already, hope, think, man, life), with friend\_VisMotion, life\_VisMotion, already\_AttentionArousal, myEMoHappy, friend\_MotorPractice among characteristics of high Conscientiousness, high values of manTemporalLIWC, manCognitionLIWC, lifeTempduration characterize low Conscientiousness (Fig. 13).



Fig. 13. Feature importance for Conscientiousness (Letter dataset)

The lowest BER for Neuroticism was 0.36 and obtained with ncomp=1. Among the features with the highest correlation with this feature are the semantic features of whole (whole\_VisLIWC, whole\_MotorBinder, whole\_VisBody, etc.), longago, see, miss1, day. This feature contributes to the classification, with a high level of day\_VisMotion, day\_MotorPractice, see\_EmoSentiment indicates low to medium levels, longago\_TempDuration, see\_EmoPleasant, see\_GustNorms – high levels (Fig. 14).



Fig. 14. Feature importance for Conscientiousness (Letter dataset)

The lowest BER for openness = 0.45 with ncomp=2, overall ER=0.21. All models detected a high level with a low ER (up to 0.13). Among features with the highest positive correlation with Openness are the features of write (OlfacNorms, VisNorms, CognitionImage), emoji (GustNorms, SomatSurface, SomatNociception – positively, EmoSentiment, EmoSurprised, VisMotion - negatively), very, nice, think, time, love. These features contribute to the classification (especially *write*) (Fig. 15).

|  |  |
| --- | --- |
|  |  |

Fig. 15.Feature importance for Conscientiousness (Letter dataset)

The metrics of the constructed models (in terms of weighted accuracy, 1 - BER) are summarized in Table 2.

Table 2. Accuracies of the classification models

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Genre | Extraversion | Agreebleness | Conscientiousness | Neouroticism | Openness | AVR | Gender |
| PICTURE | 0.65 | 0.64 | 0.6 | 0.65 | 0.52 | 0,612 | 0.61 |
| LETTER | 0.65 | 0.69 | 0.59 | 0.64 | 0.55 | 0.624 | 0.75 |

**Discussion**

*Gender*

A great deal of research is devoted to inferring an author’s characteristics from a text. Gender is the most popular factor which has been exhaustively studied in authorship profiling. However, as one of the latest reviews shows (HaCohen-Kerner, 2022), the gender accuracy results vary from 52% to 91% (most are less than 85%). The author of the above-mentioned finds these results “quite surprising” and expects “to see higher results when it comes to seemingly relatively simple types of classification especially by gender (only 2 categories) when a large number of teams have competed over the years”. We claim that, in text-based gender detection, there is a strong need to take into account genre characteristics, i.e., such an important factor as the need for construction of the addressee and therefore necessity to play a corresponding gender role which dramatically affects the accuracies of the model and could potentially explain the diversity of the obtained predictive model metrics.

In recent years, DL methods have become widely used for gender classification tasks, however, contradictory results were obtained regarding their superiority over the classical ML methods (e.g.; SVM, LR, and RF) (Lopez-Santillan et., 2023, HaCohen-Kerner, 2022). The current SOTA for Russian informal texts (Sboev et. al., 2020) was obtained with such features as morphological binary vectors, FastText vectors, textual dependency tree structure fed to graph convolution network and LSTM resulting in f1-score 84%. A drawback of such models is the difficulty in assessing the features. For the other models obtained on the same data Rectified Linear Unit (ReLU, 1 Hidden Layer) was the most efficient algorithm, with an accuracy of 0.74+/-0.05 using the imp\_quarter feature selection technique.

The most popular features which are used as input for traditional machine learning algorithms, are word unigrams and bigrams and character 3–4-5 grams and number of words from predefined dictionaries (LIWC is the most popular one). However, in most cases, the authors did not explain why the combinations of these features were the most successful for the classification tasks under discussion. In case of using the LIWC-like features, the efficiency of the models is heavily dependent on the presence of such a word in text and on text length. Our approach, being highly interpretable, allowed us to obtained good results for the texts which imply the construction of the addressee and playing a gender role and does not have such drawbacks as those described above.

*Personality traits*

Text-based personality traits prediction is also a popular direction of research, however, the list of the datasets for this task is limited due to the difficulties in obtaining the appropriate metadata. One of the closest to ours is PAN2015 dataset (Rangel et al., 2015) as having the shortest texts from 294 users Twitter tweets annotated with their Big Five scores. The best results (from 0.646 for consciousness to 0.713 for neuroticism) for this dataset was obtained by Yangfu et al. (2022) who also tackled this problem as a binary classification one and constructed a lexical psycholinguistic knowledge-guided graph of a neural model. They learn personality-aware word embedding that encodes psycholinguistic information in the continuous representations of words.

The authors claim the extraversion and neuroticism can be effectively extracted from a text, because people with neuroticism and extraversion traits often use modal particles and adjectives, while agreeableness personality trait in the text information is not obvious. As they constructed the lexicon to enrich the features with information on a sentiment, their approach is highly dependent on the overall text sentiment level and its content as they are focusing on particular words, on the one hand, and the connection of the predicted personality trait to emotions. However, their results are similar to ours in terms of the established connection of particular words and personality traits, but our approach is more sophisticated as we focus not only on the words themselves but also on their individual meanings. Our approach is not connected to sentiment detection as we analyze word meanings over the whole set of features.

For Russian, research in text-based Big-5 personality prediction is scarce. For example, Stankevich et al. (2018) predicted Big 5 traits (as 3-class problem) from a dataset of 165 VKontakte profiles. However, due to the sparseness of user-generated texts in their dataset, they used only very basic textual features (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 classification of Big 5 traits from the dataset of 1,020 VKontakte profiles. They used lexical features, an aggression lexicon, user profile information, repost data, and reported F1 scores ranging from 61.75% on Openness to Experience to 73.75% on Extroversion. These approaches, however, are not comparable to ours since, first, their dataset consists of mostly non-user generated content, secondly, they also use non-linguistic features.

One of the largest dataset in Russian was gathered by Hull et. al. (2021). It consists of 149K VKontakte posts from 288 consenting participants, with a total of 3.8M word tokens. They used BERT architecture as well as a traditional classifier with a traditional set of features (emoji, the most frequent words and char n-grams, function word to token ratios; NRC emotional lexicon tokens; frequency of morphological features, frequencies of top 100 (RNC) unigrams, etc.) and found out the superiority of the latter.

One trait where the advantage of the set of formal characteristics they used is particularly large is the Agreeableness trait, which might be explained by the usefulness of emoji, emoticons, and/or emotional words for these trait Agreeableness which pretrained model may be ignoring since such “words” may not have appeared in its original training data content. However, the problem of a great amount of non-user generated content which is typical for VKontakte is not discussed in either of the reviewed works.

**Conclusions**

Despite a huge amount of research, the task of inferring an author’s attributes from a written text remains challenging. This is especially true for text-based personality trait detection. Recent works have applied deep learning techniques to this type of tasks, however, with limited success. A lot of works have reported the superiority of traditional algorithm over DL. The difficulties in interpretability also hinder the use of DL for AP tasks. However, traditional approaches usually involve using very basic, linguistically naïve features which adds little to our understanding of individual differences in language usage. Moreover, they are highly dependent on the context (the most frequent n-grams, words and so on).

In this paper, we propose a new approach for AP feature construction which is theoretically motivated and related to the analysis of individual word meanings which is assess through the analysis of the contexts of the most frequent (mostly) content words of the corpus. Using the methods of semantic relation extraction and a neurobiologically motivated set of word meanings, we constructed a set of features which describes the contextual meaning of words. Using this methodology, we were able to achieve the accuracies of the models which are close to SOTA despite the fact that our dataset is quite small both in terms of the text length and the number of texts.

However, our aim was not predicting but gaining the interpretable features. Our approach could be applied to a text of any length and does not depend on the presence in the texts themselves of the words which are used to extract the semantic features.

Regarding the constraints and limitations of our approach, we conducted the experiments using only texts which were created in experimental settings. Therefore, experiments on the other datasets containing texts of different types will help to further improve the methodology. We also tested only one language model trained on the general corpus. Testing of the other types of language models is the next step of our exploratory study.

The Big Five personality traits are not independent, which is proved in our correlation analysis. In this study, we only predict each personality trait individually. In the future, we will design experiments for  joint personality detection tasks.

Using another coding scheme for Big 5 scores as well as working with different questionnaires will also help to increase the validity of the proposed methodology. We also have plans to refine some categories with the appearance of new psycholinguistic databases as well as to test different context window sizes depending on the task at hand and text length. As we used only one classifier, an obvious direction of future research is testing the? other algorithms.

Despite a rather exploratory nature of the present study, we believe that using a novel feature set which describes the characteristics of individual word meanings will contribute to theoretically-driven authorship profiling. This work is only the first step in this direction. Future studies will be related to the collection of the databases of the typical (average) characteristics of the meaning of the most frequent content words (i.e.. key words of the culture) using the proposed methodology for different discourses with a known genre and an author’s characteristics and incorporating this information into the predictive models.

**Acknowledgements**

This study is supported by the grant of Russian Science Foundation no: 21-78-10148.

**References**

Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., & Desai, R. H. (2016). Toward a brain-based componential semantic representation. *Cognitive neuropsychology*, 33(3-4), 130–174.

Goldberg, L.R. (1981). Language and individual differences: The search for universals in personality lexicons. *Review of personality and social psychology*, 2(1), 141–165.

Gregory, P., H., Schwartz, A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D.J., Ungar, L.H. & Martin Seligman, E.P. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, 108(6), 934.

HaCohen-Kerner, Y. (2022). Survey on Profiling Age and Gender of Text Authors. *Expert Systems with Applications*, 199, 117140.

Hromov, A.B. (2000). Pyatifaktornyj oprosnik lichnosti: Uchebno-metodicheskoe posobie [Five factor personality inventory: Study guide]. Kurgan, Kurgan State University Publ., 23 p. (In Russian).

Hull, J.R., Novak, V., Rytting, C.A., Rodrigues, P., Frank, V.M., & Swahn, M. (2021). Personality Trait Identification Using the Russian Feature Extraction Toolkit. Recent Advances in Natural Language Processing.

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

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

Litvinova, T. & Panicheva, P. (2024). Individual differences in the associative meaning of a word through the lens of the language model and semantic differential. *Research Result. Theoretical And Applied Linguistics,* Vol. 10, N 1, 61-93 (in Russian).

Liu, W., Sun, Z., Wei, S., Zhang, S., Zhu, G., & Chen, L. (2024). PS-GCN: psycholinguistic graph and sentiment semantic fused graph convolutional networks for personality detection. *Connection Science*, 36(1).

Lopez-Santillan, J. & Gonzalez, G., Luis C. & Montes, M. & López-Monroy, A. (2023). When attention is not enough to unveil a text’s author profile: Enhancing a transformer with a wide branch. *Neural Computing and Applications*, 35, 1-20.

Miklashevsky, A. (2018). Perceptual Experience Norms for 306 Russian Nouns: Modality Rating. Spatial Localization, Manipulability, Imageability and Other Variables, *Journal of psycholinguistic research*. 47, 3, 641-661.

Panicheva, P. & Litvinova, T. (2020). Matching LIWC with Russian Thesauri: An Exploratory Study. In: Filchenkov, A., Kauttonen, J., Pivovarova, L. (eds.) Artificial Intelligence and Natural Language. AINL 2020. Communications in Computer and Information Science, vol. 1292. Springer.

Pennebaker, J. W, Boyd, R. L., Kayla, J. & Blackburn, K. (2015). The development and psychometric properties of liwc2015. Technical report, University of Texas at Austin, Austin, TX.

Potthast, M., Rosso, P., Stamatatos, E., Stein B. (2019). A decade of shared tasks in digital text forensics at pan. In: Azzopardi L, Stein B, Fuhr N, Mayr P, Hauff C, Hiemstra D (eds). Advances in Information Retrieval. Springer, Cham, pp. 291–300.

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

Sboev, A., Selivanov, A., Rybka, R., Moloshnikov, I. & Bogachev, D. (2020). A Neural Network Model to Include Textual Dependency Tree Structure in Gender Classification of Russian Text Author. In: Misyurin, S., Arakelian, V., Avetisyan, A. (eds) Advanced Technologies in Robotics and Intelligent Systems. Mechanisms and Machine Science, vol 80. Springer, Cham.

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

Stoltz, D.S. & Taylor, M.A. (2019). Concept Mover’s Distance: measuring concept engagement via word embeddings in texts. J Comput Soc Sc 2, 293–313.

Stoltz, D. S. & Taylor, M. A. (2021). Cultural cartography with word embeddings. Poetics, 88.

Stoltz, D. S., Taylor, M. A., & Dudley, J. S. K. (2024). A Tool Kit for Relation Induction in Text Analysis. Sociological Methods & Research, 0(0).

Yangfu, Z., Linmei, H., Nianwen, N., Zhang, W., Wu, B. (2022). A lexical psycholinguistic knowledge-guided graph neural network for interpretable personality detection. *Knowledge-Based Systems*, 249, 108952.

Zhou, J., Huang, J. X., Hu, Q. V., & He, L. (2020). Sk-gcn: Modeling syntax and knowledge via graph convolutional network for aspect-level sentiment classification. *Knowledge-Based Systems*, *205*, 106292.