Metrics for Personal Profiles of Social Network Users

Konstantin Sergeevich Nikolaev[1*], *Fail Mubarakovich* Gafarov[2], and *Pavel Nikolaevich* Ustin[3]

[1]Kazan (Volga region) Federal University, Institute of Psychology and Education, Department of Clinical Psychology and Personality Psychology, Kazan, Russia
[2]Kazan (Volga region) Federal University, Institute of Computational Mathematics and Information Technologies, Department of Information Systems, Kazan, Russia
[3]Kazan (Volga region) Federal University, Institute of Psychology and Education, Department of General Psychology, Kazan, Russia

Abstract. This paper discusses the technical details of obtaining and processing data to determine a set of characteristics of texts from social networks, genre preferences in movies and music genres for students of Kazan Federal University who have different academic performance (successful, average, not-successful). The selection of such characteristics is carried out using machine learning methods (Word2Vec, tSNE). The data obtained is used in the development of a functional psychometric model of cognitive behavioral predictors of an individual's activity within the framework of their educational activities. We also developed a web application for visualizing the obtained data using the Flask engine.

1 Introduction

The analysis of psychological characteristics of a person is a problem that has received a lot of attention in the past few years. The development of information technologies, methods of mathematical statistics and processing of large data sets gave a noticeable boost to the development of this topic. As part of our project, we plan to develop a functional psychometric model of cognitive behavioral predictors of an individual's activity within the framework of their educational activities. One of the applications of this system is to predict the success of students of Kazan Federal University. Current results of our projects are described in [1-4].

One of the sources of quantitative and qualitative data in our project is the social network Vkontakte, as it is a huge repository of personalized user data, among which we can distinguish the majority of current students and potential applicants of our University. However, in most cases, the mentioned data is either unstructured (such as the texts of posts on users' walls), so it is necessary to use methods for automatic processing of information.

* Corresponding author: konnikolaeff@yandex.ru

This paper will focus on the content analysis of the texts of posts and reposts social networks, and will highlight the correlation between music, film preferences and academic success.

1.1. Related Work

Our research is based on the theory that psychometric indicators of a person's personality should be reflected in individual characteristics of their content uploaded to social networks [5, 6].

First, we will briefly describe the works that analyze the relationship between human behavior in the information area and the characteristics of his personality. For example, Correa et al. [7] find out the correlation between human activity in social networks, text messengers and three dimensions in the Big-Five model, namely extraversion, emotional stability and openness to experience. In addition, the influence of gender and age category on this correlation is considered. The authors found that extraversion and openness to experience have a positive correlation with social media activity, and emotional stability shows an inverse correlation. Age and gender influence the relationship between the objects under study, but they do not play a significant role in our work.

Wilson et al. [8] approach the problem from the other side – they develop a model that allows to predict the probability that a person will often use social networks. The data that this study relies on is prepared by students reporting their social media activity and passing two tests: the NEO Five-Factor Personality Inventory and the Coopersmith Self-Esteem Inventory. Then a correlation is set between these data and used to determine the first parameter based on the tests passed.

Ryan et al. [9] consider the relationship between a broader range of psychometric indicators (the Big Five, shyness, narcissism, loneliness) and activity in the social network Facebook. Data was collected by means of psychological tests performed by the subjects. The study found that Facebook users differ from non-users by being more extroverted and narcissistic. On the other hand, non-users are more conscientious and socially isolated.

Nadkarni et al. [10] consider the use of social networks from the perspective of human needs. The authors found that the use of social networks is due to the need to belong and self-presentation. Wohn et al. [11] found that infrequent use of social media correlates with academic success. A similar problem is being considered by Michikyan et al. [12].

Taking into account these studies, we will consider projects that use machine processing of publicly available data in an attempt to identify certain characteristics of the owner of this data.

For example, Souri et al.in [13] suggest a machine learning method that can predict a student's personality type without having to perform a full five-factor psychometric test, that is, based only on data from user pages. The accuracy of the model in this work was 82.2%.

Liyao Ma et al. [14] use decision trees and linear belief functions to successfully process inaccurate data and reduce epistemological uncertainty by prioritizing the most valuable indeterminate entities in the training procedure.

1.2. Our Contribution

This paper presents the collection and processing of data necessary to identify the properties of data that are characteristic of successful and unsuccessful students. The multi-sided nature of the data will allow us to build a more correct mathematical model for further automatic analysis of the relevant data from the personal pages of future students. We will also offer technical details of getting data and suggest a method for automatically selecting their characteristics.

2 Methods

2.1. Data preparation

Data from personal pages of the social network Vkontakte belonging to KFU students were used as the source data for our algorithms. We have extracted the first 4,000 posts from each page from the list of students. We only used post text and repost text data in current research. The rest of the data was used for additional data filtering.

In a separate table of the database, information about audio recordings of a working sample of Vkontakte users was uploaded, namely, the user ID, the name of the song, the authors of the song, and the duration of the song in seconds. We implemented a search for artists in the Yandex.Music database via an API that returns detailed information about the performer when executing the request. In this way, information was generated about the musical genres present in the students' audio recordings.

To determine the genre preferences of students, a table was created in the database containing the following fields: student ID, video file ID, video file owner ID, title, video file description, and duration. Since we determine genre preferences in the field of cinema, we need to remove short videos using the "duration" field.

We created a method of automatic video genre identification using Kinopoisk API and offline cinema database. Results of this method are placed in the database.

2.2 Processing data using Word2Vec

The data obtained in the previous section was processed using the Word2Vec library, which converts text data into a set of vectors using machine learning. Based on these vectors it is possible to obtain the semantic closeness of sets of words based on their proximity in the training texts. The training set was divided into 3 groups (successful, average and unsuccessful students). The division into groups was based on the average grade point in each semester of study. The percentage of successful, average and slow students is 10-80-10, respectively. A separate neural network was trained on each of these groups, whose resulting vectors were reduced to two-dimensional points using the t-SNE machine learning algorithm. The close location of points on the plane means their semantic proximity in the source texts. The resulting array of points was truncated to 1000 words for each group and presented as a word cloud using the Bokeh data visualization library (docs.bokeh.org).

3 Results

This section shows the results of automatic processing for various types of data obtained from students' personal pages.

3.1 Posts text analysis results

As a result of collecting and analyzing texts of posts and reposts from the social network Vkontakte, we get a set of words that are most often found in the texts of posts and reposts of successful, average and unsuccessful students. Thanks to the convenient visualization of results in the form of a cloud of words, we can select a set of topics that are specific to the corresponding group of students. The selected topics are shown in Table 1.

Table 1. The ratio of subjects of texts for successful and unsuccessful students

Successful	Not successful

Congratulations with the holidays	Congratulations with the holidays
Treatment relatives	Treatment relatives
References University	Reading
The study English language	The exam
Reading	Art
The exam	The school and the teachers
Art	Sport
The school and the teachers	Health

In addition, additional Word2Vec models were trained with data grouped by academic areas: technical sciences, natural sciences, and humanities. The received topics are shown in Table 2.

Table 2. Neural network experimental results in various fields

Technical	**Natural science**	**Humanitarian**
Congratulations with the holidays	Congratulations with the holidays	Family greetings
Family greetings	Treatment relatives	Congratulations with the holidays
History	Games industry	Music
The school and the teachers	Food and cooking	Food and cooking
Reading books	Art	Art
The exam	The school and the teachers	Sport
Art	Sport	The school and the teachers
Studying English language	Healthy lifestyle	Healthy lifestyle

3.2 Movie and audio genres analysis results

For the data describing the film genres found on the students' pages, the same grouping was made for the successful, ordinary and unsuccessful students, and within the group corresponding to the successful students, a division was made into three groups according to the direction of study. The model results for movie genres are shown in Tables 3 and 4, for audio genres – in Tables 5 and 6.

Table 3. Top 3movie genres for successful students in various fields

Technical		**Natural science**		**Humanitarian**	
Genre	**Percent**	**Genre**	**Percent**	**Genre**	**Percent**
Dramamovie	19,22%	Drama movie	21,90%	Drama movie	21,87%
Thriller	5,88%	Romance movie	6,67%	Comedy movie	6,84%
Action movie	5,88%	Comedy movie	5,71%	Romance movie	5,83%

Table 4. Top 3 movie genres by category among the successful, regular and underperforming

Successful		**Regular**		**Underperforming**	
Genre	**Percent**	**Genre**	**Percent**	**Genre**	**Percent**
Drama movie	19,22%	Drama movie	21,90%	Drama movie	21,87%
Thriller	5,88%	Romance movie	6,67%	Comedy movie	6,84%

Action movie	5,88%	Comedy movie	5,71%	Romance movie	5,83%

Table 5. Top 3audio genres for successful students in various fields

Technical		Natural science		Humanitarian	
Genre	**Percent**	**Genre**	**Percent**	**Genre**	**Percent**
pop music	12,4%	pop music	10,7%	pop music	14,9%
dance music	11,6%	alternative music	10,0%	dance music	11,2%
rusrap	10,4%	rock music	9,7%	electronics	8,5%

Table 6. Top 3audio genres by category among the successful, regular and underperforming

Successful		Regular		Underperforming	
Genre	**Percent**	**Genre**	**Percent**	**Genre**	**Percent**
pop music	14,4%	pop music	14,4%	pop music	13,8%
dance music	11,1%	rusrap	10,1%	dance music	11,0%
electronics	8,6%	dance music	9,9%	rusrap	10,6%

4 Discussion

In this paper, we have identified the properties of texts and genre preferences in music and film among students with different levels of academic performance. In [15], the authors use similar methods for extracting text characteristics, namely NLP and machine learning, although they pursue slightly different goals – to identify the author of a certain set of entries in social networks. In [16] and [17], researchers determine the relationship between social media use and academic performance as reflected in the GPA (as in our work).

In general, the current work is aimed at clarifying the results of research on the relationship between the use of social networks and academic success. Specifically, it answers the question: what textual characteristics are characteristic of certain groups of students.

5 Conclusion

Qualitative analysis of various indicators of students' pages in the social network Vkontakte showed the following results:

1) In the texts of posts and reposts of excellent students more often than students with lower academic performance, there are words of University topics, mentions of foreign languages, culture, which indicates a greater interest of such students in academic and social life at the University.

2) Analysis of video titles suggests that excellent students are more likely to add videos in foreign languages (including subtitles). In addition, unlike less successful students, they prefer more "serious" genres, such as drama movies, thrillers and action movies.

3) The most revealing was the analysis of performers present on the personal pages of students. Excellent students are characterized by the presence of compositions from the genres of pop rock, classic rock, alternative, and classical music. Perhaps audio recordings show an interesting result, because they reflect a person's aesthetic and cultural preferences better than other indicators.

In the future research we plan to extend our psychometric model to apply it on groups of people other than current students. After finishing this project, we should be able to analyze psychometric parameters of future students of our university.

The study (all theoretical and empirical tasks of the research presented in this paper) was supported by a grant from the Russian Science Foundation (Project No. 19-18-00253, «Neural network psychometric model of cognitive-behavioral predictors of life activity of a person on the basis of social networks»).

References

1. G. Vakhitov, Z. Enikeeva, N. Yangirova, A. Shavalieva, P. Ustin, *Identification of the clusters of social network communities for users with a specific characteristic*, in Proceedings of the 12th International Conference on the Developments in eSystems Engineering (2019)
2. N. Yangirova, Z. Enikeeva, G. Vakhitov, Revista Turismo Estudos & Práticas (RTEP) **2**, 1-10 (2019)
3. A.A. Vyshinskaya, Z.A. Enikeeva, G.Z. Vahitov, *Statisticheskij analiz metrik pol'zovatelej social'noj seti*, Lobachevskie chteniya-2019 (2019)
4. L.M. Popov, P.N. Ustin, *Problema razvitiya lichnosti usloviyah globalizacii: psihologo-pedagogicheskie aspekty*, in Collection of scientific papers of the International Scientific and Practical Conference, Yerevan, Russia (2019)
5. Y. Gvili, O. Kol, S. Levy, European Review of Applied Psychology **70**(2), (2020)
6. C. Huang, Computers in Human Behavior **97,** 280-290 (2019)
7. T. Correa, A. W. Hinsley, H. G. de Zúñiga, Computers in Human Behavior **26**(2), 247-253 (2010)
8. K. Wilson, S. Fornasier, K. White, Behavior, and Social Networking **13**(2), 173-177 (2010)
9. T. Ryan, S. Xenos, Computers in Human Behavior **27**(5), 1658-1664 (2011)
10. A. Nadkarni, S.G. Hofmann, Personality and Individual Differences **52**(3), 243-249 (2012)
11. D.Y. Wohn, R. LaRose, Computers & Education **76**, 158-167 (2014)
12. M. Michikyan, K. Subrahmanyam, J. Dennis, Computers in Human Behavior **33,** 179-183 (2014)
13. A.Souri, S. Hosseinpour, A.M. Rahmani, Hum. Cent. Comput. Inf. Sci. **8**(24), (2018)
14. L.Ma, S. Destercke, Y. Wang, Pattern Recognition **52**, 33-45 (2016)
15. D.K. Srivastava, B. Roychoudhury, Knowledge-Based Systems 1**95**, (2020)
16. E.Alwagait, B. Shahzad, S. Alim, Computers in Human Behavior **51**, 1092-1097 (2015)
17. F.Giunchiglia, M. Zeni, E. Gobbi, E. Bignotti, I. Bison, Computers in Human **82**, 177-185 (2018)