Advances in Social Science, Education and Humanities Research, volume 437
International Scientific Conference "Digitalization of Education: History, Trends and Prospects" (DETP 2020)

Indexing of Social Network Texts for Psychometric Model of Academic Success Prediction

Prokopyev N.A.[1,*] Vakhitov G.Z.[1] Ustin P.N.[2]

[1]*Institute of computational mathematics and information technology, Kazan Federal University, Kazan, Russia*
[2]*Institute of psychology and education, Kazan Federal University, Kazan, Russia*
[*]*Corresponding author. Email: nikolai.prokopyev@gmail.com*

ABSTRACT
The paper presents the technical description of a software package for extracting and processing of text content from social networks for the purpose of their use in developing of an academic success psychometric model. This model is based on a selected set of cognitive behavioral predictors of personal activity in the context of education. Source textual data consists of posts from the Vkontakte profile pages of students of Kazan Federal University. As the main method of data processing, text indexing and word frequency characteristics analysis for academically "successfull" and "not-successful" students. The intermediate results of the developed text indices analysis are also presented.
Keywords: *natural language processing, text indexing, data analysis, social networks, psychometry*

1. INTRODUCTION

Research of the connection of personal psychosocial characteristics with various indicators of their personal activity is an emerging field of study nowadays. Rapid development of information technology, mathematical methods and the possibilities of big data processing allows us to build and verify formal psychometric models for their further use in creating of software systems that can predict some forms of personal activity success. In particular, within the objectives of our project (the main research results are presented in [1-3]), we consider the development of a system for predicting the academic success of students based on data from the information-analytical system of Kazan Federal University and from their profiles on the social network Vkontakte.

One of the important sources of personalized data that can be converted into psychometric characteristics of a person are posts and reposts text found on personal pages. The classical methods of information technology, namely, methods of information retrieval, allow us to structure this data for further automatic processing and mathematical analysis. The article provides a technical description of text indexing process with extracting the most frequent words that are typical for groups of relatively successful, average and unsuccessful students.

1.1. Related Work

The framework of our research is based on the hypothesis of the relationship between psychometric characteristics of a person in his life activity and their digital footprint in social networks. Therefore, we will firstly describe existing studies in this field.

In paper [4], correlations were found between activity in social networks or messengers and three elements of the Big-Five model, such as extraversion, neuroticism, and openness to experience. According to the results of the study, authors conclude that extraversion and openness to experience have a direct correlation with the level of activity in social networks, while the characteristic of neuroticism shows an inverse correlation.

At the same time, authors of [5] take into account the relationship between level of human activity on Facebook social network and all elements of the Big-Five model (extraversion, agreeableness, consciousness, neuroticism, openness to experience). In addition to them, authors also take into account the characteristics of shyness, narcissism, and loneliness. The main research method is psychological tests, which showed that active users of a social network are more extroverted and narcissistic. On the other hand, less active users, according to the authors, are more conscious and lonely.

Authors of papers [6] and [7] discuss somewhat similar topics and come to similar conclusions about positive correlation between less active use of social networks and academic success.

Further analysis of existing studies relates to research about qualitative analysis of various textual data features for machine learning. These papers show that text indexing is an essential step for formalizing text features in a qualitative analysis in humanitarian research projects.

In paper [8], a lexico-statistical analysis of posts and reposts texts from users of VKontakte social network for the subject of the user's professional interests is considered. Such a system, according to the authors, will facilitate for the professional orientation of students. For the task of analysis, the texts were normalized and three

thesauruses for automatic indexing were created: humanitarian, mathematical and scientific. Further, based on these thesauruses, text classifiers of the following types were developed: linear discriminant analysis (LDA), support vector machine (SVM), logistic Regression (LR), decision trees (Trees), random forest (RF). The most stable result was shown by the LR classifier.

The author of [9] touches on the task of automatical determining of text style based on statistical analysis. For this, experimental corpora of scientific, artistic, official and fiction styles were created and indexed for the purpose of numeric feature extraction. A decision trees classifier was built on the basis of these features. The resulting classifier showed the best results for texts in official style, and the worst results when classifying of journalistic texts, which he mistakenly classified as fiction. Also, the classifier sometimes mistakenly classified pop-science texts as fiction.

The main object of research in the paper [10] is the analysis of natural language texts modality. By modality is understood both the relation of the text author to the message, and the relation of the message to reality. The author analyzes existing approaches to extracting of modality from the text, and finds their dependence on the subject area. All methods include, as a first step, markup of the text, which can be done using machine learning, which involves preliminary indexing.

1.2. Our Contribution

In this paper we present an approach to processing of large text dataset from social network profiles using one of the classical technologies of information retrieval – text indexing and words frequency characteristics analysis from obtained indices. This approach allows us to move from unordered and informal natural language texts to the data that is ready for algorithmic processing. Further, this data can be used to develop various kinds of psychometric models.

1.3. Paper Structure

The paper is divided into four chapters. Chapter 2 describes methodology, algorithms, implementation features, and programming libraries used in development. Chapter 3 is devoted to the analysis of intermediate results in form of determining the most frequent words for datasets from "successful" and "unsuccessful" students. Chapter 4 presents the conclusion and plans for further work.

2. METHODOLOGY

2.1. Data acquisition

To begin with, a sample of 21208 student's data was extracted from information-analytical system of KFU. All these students has their personal pages on VKontakte social network publicly available. Further, detailed post data was extracted from these pages using the "requests" Python library and VK API. The following information was extracted for each post: post author identifier, post identifier, post text. In addition, if the post contained a repost, information about this repost was also extracted: repost author identifier, repost identifier, repost text. A total of 4168879 posts were extracted, which were then stored in a separate database.

When forming indices from a general sample of students based on their average academic performance, the following samples were formed: the sample of relatively successful students and the sample of relatively unsuccessful students. The sample of successful students included the upper 15% slice of students with the highest grade point average and amounted to 1300 entries. The lower 15% slice of students with the lowest grade point average made up 835 people and fell into the sample of unsuccessful students.

2.2 Text indexing

The main data source for the formation of necessary text indexes are post texts and repost texts. The indexing process begins with the tokenization of current text which means dividing it into separate token words. During tokenization, the text is cleared of punctuation and extra characters and then some semantically standing apart elements namely hashtags and URL links are extracted from this text. Separate indices are constructed for these elements due to the peculiarity of their semantics in comparison with ordinary words. Whole tokenization process is performed with usage of regular expressions from "regex" Python library.

Before adding ordinary words in the index, it is necessary to convert them into some initial form in order to take into account different forms of the same word in the text. For this, "pymorphy2" Python library is used, which in turn uses the OpenCorpora Russian language corpus to morph words into their linguistic normal form. Only after this processing is the word added to the index, which expresses the number of uses for this word in posts and reposts from the personal page of some student. Thus, the indexing of "Word -> Number of uses" type occurs for each student’s page from the general sample separately for ordinary words, for hashtags, for URL links from posts and from reposts, resulting in 6 indexes per profile. Afterwards, these indices are combined on the basis of the student's affiliation to sample of successful or unsuccessful

students. In addition, aggregation into general indices is performed for the whole initial sample.

In addition to indexes of "Word -> Number of uses" type, which is simply called an index, so-called reverse indexes of "Word -> Number of people from the sample on whose pages this word occurs" type are also constructed for every considered sample. As a result, 36 of the following indices were obtained:

1.1) General index and reverse index of post texts
1.2) General index and reverse index of hashtags from posts
1.3) General index and reverse index of links from posts
1.4) General index and reverse index of repost texts
1.5) General index and reverse index of hashtags from reposts
1.6) General index and reverse index of links from reposts
2.1) Index and reverse index of post texts for unsuccessful
2.2) Index and reverse index of hashtags from posts for unsuccessful
2.3) Index and reverse index of links from posts for unsuccessful
2.4) Index and reverse index of repost texts for unsuccessful
2.5) Index and reverse index of hashtags from reposts for unsuccessful
2.6) Index and reverse index of links from reposts for unsuccessful
3.1) Index and reverse index of post texts for successful
3.2) Index and reverse index of hashtags from posts for successful
3.3) Index and reverse index of links from posts for successful
3.4) Index and reverse index of repost texts for successful
3.5) Index and reverse index of hashtags from reposts for successful
3.6) Index and reverse index of links from reposts for successful

3. RESULTS ANALYSIS

Before analysis of the results, stop words, such as: conjunctions, prepositions, pronouns, auxilary verbs, names, were deleted from indices. Further analytical processing of indices is associated with finding the most frequent unique words for indices of successful and unsuccessful students. Using this approach, it is possible to identify the most characteristic words, hashtags, links for these groups of people. However, the construction of such index subsets can be done in two ways of intersecting of the corresponding indices. Both methods involve sorting of the indices being intersected by the frequency charastericstic in descending order as the first step, but further steps differ in their order of execution.

When using the first method, called “Top Unique”, for indices being intersected S (for successful) and U (for unsuccessful), their complements $S\backslash U$ and $U\backslash S$ are found, after which an upper slice of N words is extracted from them. For this study, N is taken as 1000. Thus, the result of this method consists of the index subsets containing the top 1000 words which are unique to successful students and unsuccessful students.

When using the second method, called “Unique Tops”, for indices being intersected, their upper slices of N words are extrated from them. After this step their complements $S_c\backslash U_c$ and $U_c\backslash S_c$ are found for the obtained index slices S_c and U_c. Thus, the result of this method consists of index subsets containing unique words for successful students and unsuccessful students for the top 1000 slices from source indices.

These methods were applied to previously obtained 14 indices and reverse-indices for successful and unsuccessful groups of students. Both methods give different results when applied to the initial data, but it cannot be said unequivocally that the results given by one method can better reflect the psychometric characteristics than the results of the other method.

Some examples of the results analysis are presented in Tables 1-6 in the form of the most frequently occurring characteristic words in index subsets of post and repost texts, hashtags, and URL links for samples of successful and unsuccessful students from both “Top Unique” and “Unique Tops” methods.

Table 1 Post texts “Unique Tops” index for successful

Word	Word translation	Frequency	Word aspect
солнышко	sun	1387	positive tone
студент	student	274	education
конкурс	contest	266	education, arts
совет	advice	232	positive tone, education
поддержка	support	219	positive tone, difficulties

Table 2 Post texts "Unique Tops" index for unsuccessful

Word	Word translation	Frequency	Word aspect
картина	picture	265	arts
долг	duty	174	army
матч	match	134	sports, games
выставка	exhibition	122	arts
армия	army	86	army

Table 3 Post hashtags "Top Uniques" reverse-index for successful

Word	Tag explanation	Frequency	Word aspect
#kazanvolunteers	Kazan volunteers	9	volunteering
#selet	Selet (organization)	8	organization
#выпускной	graduation	7	event
#работа	job	6	job
#выборы	elections	6	event

Table 4 Post hashtags "Top Uniques" reverse-index for unsuccessful

Word	Tag explanation	Frequency	Word aspect
#uefa	UEFA	2	sports
#отпуск	vacation	2	relaxation
#tatneftarena	Tatneft Arena (stadium)	2	sports
#шашлык	barbecuing	2	relaxation
#LeagueofLegends	League of Legends (game)	2	gaming

Table 5 Repost links "Top Uniques" reverse-index for successful

Word	Site information	Frequency	Word aspect
www.openculture.com	educational media	6	education, news
www.bbc.co.uk	news site	6	news
tass.ru	news agency	5	news
www.khanacademy.org	e-learning platform	5	education, e-learning
webcast.berkeley.edu	scientific organization	5	science, e-learning

Table 6 Repost links "Top Uniques" reverse-index for unsuccessful

Word	Site information	Frequency	Word aspect
t.co	Twitter content site	3	social network
steamcommunity.com	Game shop	3	gaming
www.wday.ru	Women journal	2	entertainment
joyreactor.cc	Entertainment site	2	entertainment
www.hellride.ru	Sports shop	2	sports

4. CONCLUSION

The obtained indices can be used for a qualitative analysis of psychometric characteristics in the research personal activity indicators in both expert and automatic mode. From the intermediate results, we can conclude:
1) Successful students often use words that have a positive tone, as well as words related to education and arts. In addition, they are characterized by words expressing professional difficulties. The hashtags used by them are associated with events, volunteering, organizations, job. Such students often provide links to news sites, e-learning systems and research or educational organizations.
2) Unsuccessful students often use words related to arts, sports, computer games, army. The hashtags they use are associated with computer games, humor, relaxation, and motivational self-development. Such students often provide links to gaming and entertainment sites, as well as to sites of other social networks.
These findings are a non-expert preliminary assessment of the acquired data and do not reflect all aspects of indexes usage in psychometric research. Further work within the framework of our project is related to statistical processing of indices which is supposed to find correlations with the studied personal activity indicators using both classical mathematical methods and machine learning. This will allow us to build a system that allows prediction with some accuracy of the academic success of students based on textual data from their social network profiles.

ACKNOWLEDGMENT

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] F. Gafarov, Z. Enikeeva, G. Vakhitov, K. Nikolaev, Psychometric Predictors of Educational Success of Students-Humanitarians in Social Networks, Journal of International Pharmaceutical Research, 2019.

[2] K.S. Nikolaev, N.M. Davletshin, A.A. Berdnikov, Nejrosetevoj metod klassifikacii tekstov grupp social'noj seti Vkontakte, Sovremennye informacionnye tekhnologii i IT obrazovanie, 2019.

[3] P.N. Ustin, L.M. Popov, R.N. Hakimzyanov, F.M. Gafarov, Vzaimosvyaz' lichnostnyh harakteristik studentov s pokazatelyami ih personal'nogo profilya v social'nyh setyah, kognitivnoe modelirovanie v professional'nom obrazovanii, 2019.

[6] T. Correa, A.W. Hinsley, H.G. de Zúñiga, Who interacts on the Web?: The intersection of users' personality and social media use, Computers in Human Behavior, Volume 26, Issue 2, 2010, pp. 247-253, DOI: 10.1016/j.chb.2009.09.003.

[7] T. Ryan, S. Xenos, Who uses Facebook? An investigation into the relationship between the Big Five, shyness, narcissism, loneliness, and Facebook usage, Computers in Human Behavior, Volume 27, Issue 5, 2011, pp. 1658-1664, DOI: 10.1016/j.chb.2011.02.004.

[8] D.Y. Wohn, R. LaRose, Effects of loneliness and differential usage of Facebook on college adjustment of first-year students, Computers & Education, Volume 76, 2014, pp. 158-167, DOI: 10.1016/j.compedu.2014.03.018.

[9] M. Michikyan, K. Subrahmanyam, J. Dennis, Can you tell who I am? Neuroticism, extraversion, and online self-presentation among young adults, Computers in Human Behavior, Volume 33, 2014, pp. 179-183, DOI: 10.1016/j.chb.2014.01.010.

[10] A.A. Stepanenko, K.S. Shilyaev, Z.I. Rezanova, Atribucija professional'nyh interesov pol'zovatelej

socialnoj seti «Vkontakte» na osnove tekstov tematicheskih grupp i personalnyh stranic, Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya, No. 52, 2018, pp. 130-144. DOI: 10.17223/19986645/52/8.

[11] A.R. Dubovik, Avtomaticheskoe opredelenie stilisticheskoi prinadlezhnosti tekstov po ikh statisticheskim parametram, Komp'yuternaya lingvistika i vychislitel'nye ontologii, Volume 1, 2017, pp. 29-45.

[12] S.R. Egikyan, Sovremennye metody analiza modal'nosti v tekstakh na estestvennom yazyke, Programmnye sistemy: teoriya i prilozheniya, No. 3(34), 2017, pp. 133-167. DOI: 10.25209/2079-3316-2017-8-3-133-167.