IRSTI 28.23.13 https://doi.org/10.26577/JMMCS-2019-3-m5

Exploration of student behavior patterns through digital footprints

Nugumanova A., S. Amanzholov East Kazakhstan State University,
Ust-Kamenogorsk, Kazakhstan, e-mail: anugumanova@vkgu.kz
Mansurova M., Al-Farabi Kazakh National University, Almaty, Kazakhstan,
e-mail: mansurova.madina@gmail.com
Baiburin Ye., S. Amanzholov East Kazakhstan State University Ust-Kamenogorsk,
Kazakhstan, e-mail: ebaiburin@vkgu.kz

In this experimental work, a set of Data Mining methods were used to reveal student behavior patterns by analyzing their digital footprints in social Web. Data were gathered from open social profiles of students learning at one of the universities in Kazakhstan. For this case study, 25 publications appeared in the students' social feeds were selected, and students' digital footprints (namely, information about their likes on these publications) were fixed. Patterns extracted via analysis of these footprints were compared with the results of psychological tests that were carried out before; and finally, the degree to which both these findings corroborated and complemented each other was assessed. Therefore, conducted experiments provided by R ecosystem demonstrated the potential of proposed methods to analyze digital footprints for the sake of educational analytics. Despite the fact that a very small set of data was used, the case study results are quite illustrative.
Key words: digital footprints, Data Mining, clustering, principal components analysis, R language.

Сандық іздер арқылы студенттердің мінез-құлық заңдылықтарын зерттеу

Нугуманова А.Б., С. Аманжолов атындағы Шығыс Қазақстан мемлекетік
университеті, Өскемен қ., Қазақстан, e-mail: anugumanova@vkgu.kz
Мансурова М.Е., әл-Фараби атындағы Қазақ ұлттық университеті, Алматы қ.,
Қазақстан, e-mail: mansurova.madina@gmail.com
Байбурин Е.М., С. Аманжолов атындағы Шығыс Қазақстан мемлекетік университеті,
Өскемен қ., Қазақстан, e-mail: ebaiburin@vkgu.kz

Бұл тәжірибелік жұмыста әлеуметтік желілердегі сандық іздерін талдау арқылы оқушылардың мінез-құлық заңдылықтарын анықтау үшін Data Mining әдістерінің жиынтығы қолданылды. Деректер Қазақстанның жоғары оқу орындарының бірінде оқитын студенттердің ашық әлеуметтік профилдерінен алынды. Осы зерттеу үшін студенттердің әлеуметтік арналарына кіретін 25 жарияланым таңдалды, және әр басылымға студенттердің сандық іздері жазылды (студенттерге Ұнаған басылымдарды атап өткен «ұнату» белгілері туралы ақпарат). Осы сандық іздерді интеллектуалды талдау арқылы алынған заңдылықтар бұрын жүргізілген психологиялық тесттердің нәтижелерімен салыстырылды және соңында осы нәтижелердің екеуі бір-бірін растайтын және толықтыратын дәрежесі бағаланды. Осылайша, R тілдік инфрақұрылыммен жүргізілген тәжірибелер білім беру аналитикасы үшін ұсынылған әдістердің жоғары әлеуетін көрсетті. Өте аз мәліметтер жиынтығы қолданылғанына қарамастан, зерттеу нәтижелері айтарлықтай нәтижелі болды.
Түйін сөздер: сандық іздер, Data Mining, кластеризация, негізгі компоненттері талдау, R тілі.

Исследование паттернов поведения студентов через их цифровые следы

Нугуманова А.Б., Восточно-Казахстанский государственный университет им. С. Аманжолова, г. Усть-Каменогорск, Казахстан, e-mail: anugumanova@vkgu.kz
Мансурова М.Е., Казахский национальный университет имени аль-Фараби, г. Алматы, Казахстан, e-mail: mansurova.madina@gmail.com
Байбурин Е.М., Восточно-Казахстанский государственный университет им. С. Аманжолова, г. Усть-Каменогорск, Казахстан, e-mail: ebaiburin@vkgu.kz

В этой экспериментальной работе набор методов Data Mining использовался для выявления паттернов поведения студентов путем анализа их цифровых следов в социальных сетях. Данные были собраны из открытых социальных профилей студентов, обучающихся в одном из вузов Казахстана. Для данного исследования были отобраны 25 публикаций, попавших в социальные ленты студентов, и по каждой публикации были зафиксированы цифровые следы студентов (а именно, информация об отметках «Нравится», которыми студенты выделяли понравившиеся публикации). Паттерны, извлеченные с помощью интеллектуального анализа этих цифровых следов, сравнивались с результатами психологических тестов, которые были проведены ранее, и в конечном итоге была оценена степень, в которой оба этих результаты подтверждали и дополняли друг друга. Таким образом, проведенные эксперименты, обеспеченные инфраструктурой языка R, продемонстрировали высокий потенциал предлагаемых методов в целях образовательной аналитики. Несмотря на то, что использовался очень небольшой набор данных, результаты исследования оказались достаточно показательными.
Ключевые слова: цифровые следы, интеллектуальный анализ данных, кластеризация, метод главных компонент, язык R.

1 Introduction

Digital footprint is a fairly general concept by which all possible actions performed by an user in a digital environment are expressed [1, 2]. In this work, digital footprints means likes, put by users to posts of other users in social networks. We rely on the methodology for assessing and analyzing digital footprints, proposed by the authors of work [3]. These authors provide an accessible step-by-step "tutorial for social scientists seeking to benefit from the availability of big data sets". They develop two complementary analytical approaches. We interested in the first approach which is aimed to employ cluster analysis and dimensionality reduction to extract patterns from large data sets. Unlike the original work, our data set is rather small, but this did not prevent us from getting the same interesting results as the quoted authors. The proposed methodology is described in the most general form in [4]. In this description, there are 4 main steps:

1. open personal data of users of social networks (the so-called "digital footprints") are collected using special applications (i.e. parsers, Social web crawlers, and so on);
2. as a rule, users also fill out psychological questionnaires (tests);
3. the results of psychological tests can be compared with available open information about user behavior on social networks.
4. all gathered data is not only analyzed using statistical methods, but also is used to build predictive models with the help of machine learning algorithms.

Thus, the psychological and personal characteristics of users (for example, the level of motivation, the degree of openness to the new, the level of subjective well-being, etc.) can be

predicted only on the basis of their open data in social networks – publications, subscriptions, likes, and this is a completely new way to get information about respondents [4]. For example, in work [4], it is proved that only on the basis of Facebook likes a person's gender and ethnic origin can be determined with accuracy of more than 90%, a person's age can be determined with accuracy 75% and his or her openness to experience can be determined with accuracy 43%.

In educational predictive analytics, this approach can be useful for adapting teaching and learning process, based on the personal characteristics of students. One of the most important applications of predictive analytics today is to predict the behavior of students in order to identify those of them who are inclined to drop out of university and therefore who need a special attention.

2 Related work

As it says in [6], Internet provides vast opportunities for individuals to present themselves in different ways, and personality-perception researchers have turned to studying social media in order to ask whether a person's digital traces can reveal aspects of his or her identity. In this context, one of the most popular tools used in the assessment of personality, is the "Big Five"model, also known as the OCEAN model [7] (see Fig. 1). According to this model there are five basic dimensions of personality, namely openness, conscientiousness, extraversion, agreeableness and neuroticism. The characteristics of these dimensions, which we have derived from [6], are shown in Figure 1.

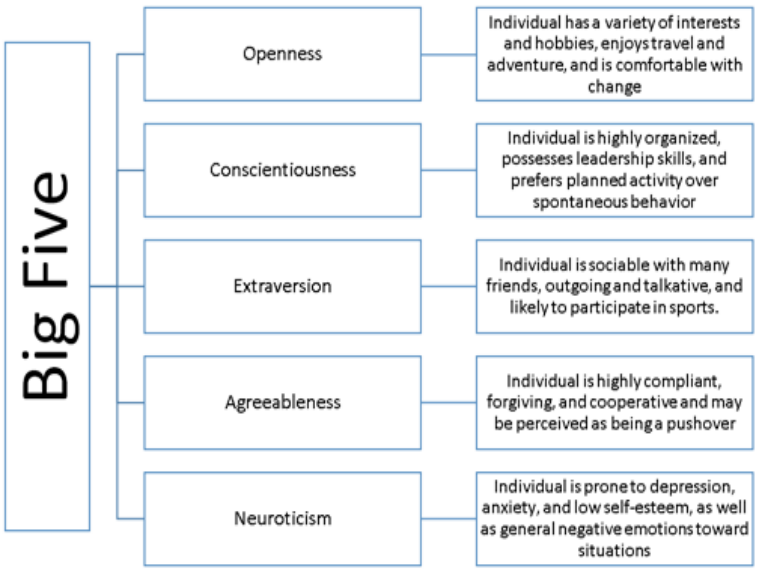

Figure 1: Characteristics of Big Five traits (obtained from work [7])

For example, in [8] the Big Five questionnaire was used for self-reporting of 100 Facebook users. Their profiles were then content analyzed by 35 observers for the presence of 53 cues. These were such signals as a profile picture, the number of friends, the number of

quotes, the use of emoticons (smiles), status updates, comments on others' posts, others' likes, music listed, movies listed, etc. Although all 35 observers were unfamiliar with the profile owners, nevertheless, they were able to accurately estimate such traits of Big Five personality as extraversion, openness, conscientiousness and agreeableness. Only for such a trait as neuroticism, the accuracy of diagnosis was low, and none of the considered signals could be indicated as diagnostic warrant. According to the authors of [7], this finding is consistent with previous studies, which claimed that neuroticism is a trait that is difficult to judge on Facebook.

The authors of [8] analyzed 700 million words, phrases, and thematic concepts collected from messages published on Facebook by 75,000 volunteer respondents who, in addition to this, took standard Big Five personality tests. According to the authors, they found striking differences in language depending on personality, gender and age. For example, the extrovert vocabulary more often included words such as a party, guys, let's, club, ready, tonight, love, and the introvert vocabulary more often included words such as anime, internet, computer, reading. Thanks to visualized word clouds, the authors were able to put forward new hypotheses about the correlation of language with a personality type. The results showed that emotionally stable people wrote about pleasant social activities, such as "sport", "vacation", "beach", "church", "team" and the topic of family time. The results also showed that introverts are interested in Japanese media and that those low in openness like to use social network's shorthands. In addition, authors use extracted open-vocabulary features to predict gender, age and personality factors. They randomly sampled 25% of users as test data, and used the remaining 75% as training data to build predictive models. Support vector machine was used to classify the binary variable of gender, and ridge regression was used to predict age and each factor of personality. Features were first run through principal component analysis to reduce the feature dimension. The best accuracy of prediction reached of 91.9%.

In [9], three fundamental conclusions were made based on the results of a large experimental study. First conclusion is that computer-based personality analysis based on digital footprints (Facebook Likes) is more accurate than human-based analysis using manual questionnaires. Second one is that computer models exhibit a higher degree of consistency (less variation in scores). Third conclusion is that computer predictive models have higher external reliability in predicting life outcomes such as substance use, political attitudes, and physical health; according to some results, they even outperform self-esteem of personality. The study compared the accuracy of human and computer personality judgments using a sample of 86,220 volunteers who completed a 100-point questionnaire.

Authors of [10] also analyzed the ability of digital footprints collected from social networks to predict Big Five personality traits. In addition, they explored the impact of various types of digital footprints on prediction accuracy. The results of the analysis showed that the predictive power of digital footprints corresponded to the standard correlation upper limit for behavior that allows to predict presonality, with correlations ranging from 0.29 (agreeableness) to 0.40 (extraversion). In general, the results showed that the accuracy of prediction increased when the analysis included demographic data of users and several types of digital footprints.

Authors of [11] examined prediction accuracy for Big 5 profile, Holland types, buying behavior types, and Gardner Multiple Intelligence scores from the data on preferences of the images from a pre-defined gallery. 1400 participants filled online questionnaires, and then selected from 20 to 100 images from 300 predefined ones. Each image was associated with a

set of tags belonging to 4 categories: objects in the image; general description of the landscape or situation; verbs describing behavior of objects in image; and emotions associated with the picture. The data was processed by the artificial intelligence module based on the gradient boosting algorithm. The percentage of "liked" images in which certain tags were present was used for prediction. 75% of the data was used to train the model, and the remaining 25% for testing. The mean prediction accuracy reached of 0,83. Thus, the authors believe that their approach is promising, i.e. the tasks associated with the selection of favorite images from the gallery can be used to predict some psychometric characteristics, such as Big 5, Holland and others.

In [12], a systematic literary review was presented which summarized studies on the prediction of user demographic data based on digital footprints. Studies were included in the review if they met the following criteria: (i) they reported results where at least one demographic feature was predicted from at least one form of digital footprint; (ii) they used automated predictive methods; (iii) they studied either explicit or implicit footprints. The authors of the review analyzed 327 papers published before October 2018. In these reviewed papers, 14 demographic attributes were deduced from digital footprints; the most studied attributes were gender, age, location, and political orientation. For each of the identified demographic attributes, the authors provided next information: a platform where digital footprints were found, sample sizes, prediction accuracy, and classification methods used.

Work [13] proposed an approach for predicting consumer decision-making styles by analyzing digital footprints on Facebook. Authors of the work obtained questionnaires and various digital footprint contents from 3304 participants. Footprint contents included "Likes", "Status" and media-content such as photo and video. The authors randomly divided this data in proportion 80/20, i.e., 80% for training and 20% for testing. Their experiments demonstrated the efficiency of proposed approach.

In works [14, 15] methodological aspects of deep and big data studies of Facebook and Instagram through methods involving the use of API data were discussed. Authors of work [14] argued for three major issues: data quality, access and ethical considerations, whereas the author of work [15] was focused on data access and validity.

3 Data and methods

3.1 Data acquisition

Data is gathered from open social profiles of 28 students learning at one of the universities in Kazakhstan. It contains information about 25 posts which these students have liked or not. The information is represented in the form of cross-tabulation matrix whose rows correspond to students, columns correspond to posts and cells correspond to likes. In work [3] this matrix is named as User-Likes matrix. If a student has liked a post, 1 is written at the intersection, otherwise 0 (see Fig. 2). Also, Big Five personality trait profiles of these students are obtained through the standard B5's questionnaire before data gathering starts. Profiles represented in the form of dataset whose rows correspond to students and columns correspond to five traits such as Extraversion, Agreeableness, Conscientiousness, Neuroticism and Openness (see Fig. 3). The R system is used for data processing and mining.

Name	P01	P02	P03	P04	P05	P06	P07	P08	P09	P10
Таншолпан Нурлыбек	0	1	0	1	1	1	1	0	1	0
Таннур Дукай	1	1	0	1	1	1	1	0	1	1
Ринат Кумашев	0	1	0	1	0	1	1	0	1	1
Рауан Мухамжаров	0	1	1	1	1	0	1	1	1	1
Назым Абдеканова	0	1	1	0	1	1	1	1	1	1
Молдир Асылбекова	0	1	0	1	1	1	0	0	1	0
Меруерт Маулет	1	1	0	1	1	1	1	0	1	1
Меруерт Кайдарова	1	1	0	0	0	1	1	1	1	1
Мархаба Карменова	1	0	0	0	0	0	1	1	0	0

Figure 2: A cross-tabulation matrix for students likes

Name	Extr	Agre	Cons	Neur	Open
[illegible]	59	59	60	61	57
[illegible]	45	52	55	47	51
[illegible]	51	39	46	54	57
[illegible]	48	65	53	52	62
[illegible]	33	59	34	40	48
[illegible]	57	68	66	37	66
[illegible]	32	57	52	43	53
[illegible]	52	56	53	58	61
[illegible]	52	59	58	49	49

Figure 3: A dataset of the Big Five personality traits

3.2 Big Five Personality Data exploration

At first, Big Five personality trait profiles have been analyzed through principal component analysis. There it is defined, that first 3 principal components of data explain 88,44% of all information (see Fig. 4).

```
big5 <- read.csv2("profiles.csv", dec=",")
pca <- prcomp(big5, scale = F)
eig <- get_eigenvalue(pca)
fviz_eig(pca, addlabels = TRUE)
var <- get_pca_var(pca)
corrplot(var$contrib, method="square", is.corr = F)
```

The first principal component is a combination of neuroticism and extraversion. The second principal component is a combination of neuroticism and conscientiousness. The third principal component is concentrated primarily on extraversion. The fourth and fifth components express high level of openness and agreeableness respectively (see Fig. 5).

Therefore, these principal components express variables which correlate with each other. For example, agreeableness correlates with conscientiousness, and openness correlates with neuroticism (see Fig. 6).

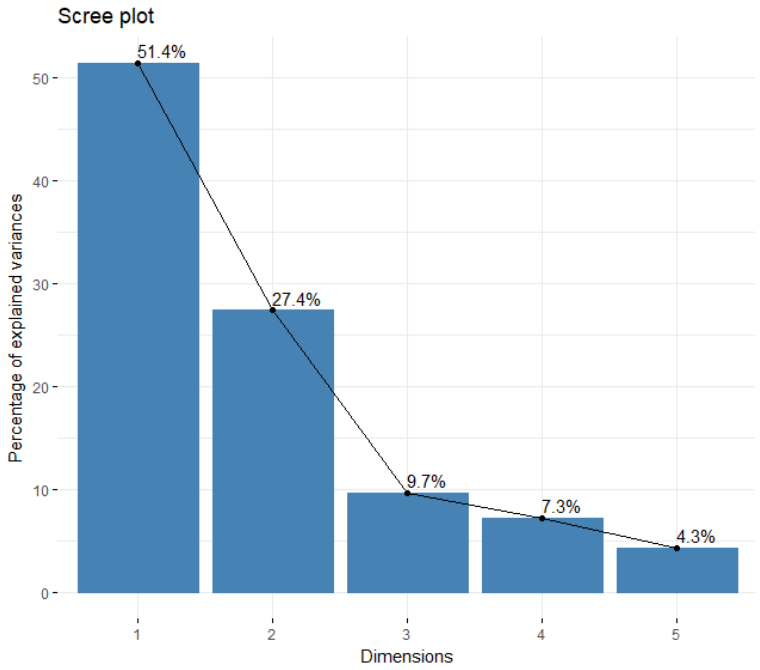

Figure 4: Principal components of Big Five dataset

Then, Big Five dataset is clustered with k-means algorithm. Fig. 7 shows visualization of cluster partitions over principal component 1 (Dim1) and principal component 2 (Dim2), and Fig. 8 shows the same visualization but over variable 2 (Agreeableness) and variable 5 (Openness). Clustering is necessary to gain deeper understanding of data structure.

```
km <- kmeans(big5, 3)
fviz_cluster(km, data = big5)
fviz_cluster(km, choose.vars = c(2,5), data = big5)
```

3.3 Student-Like Matrix decomposition

The next step is singular value decomposition of the Student-Like matrix as it is described in work [3]. Singular decomposition of the matrix is necessary to extract topics (dimensions) in a given collection of posts and compare these topics to profile traits. It has been empirically shown that in this case most appropriate number of dimensions is 6.

```
likes <- read.csv2("likes.csv")
M <- as.matrix(likes)
Msvd <- irlba(M, nv = 6)
u <- Msvd$u
v <- Msvd$v
```

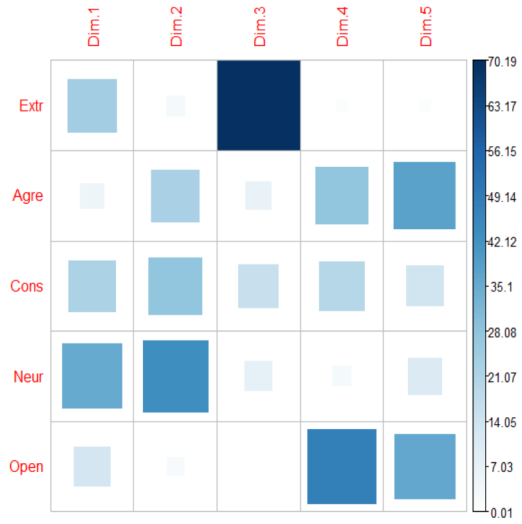

Figure 5: Composition of Principal components

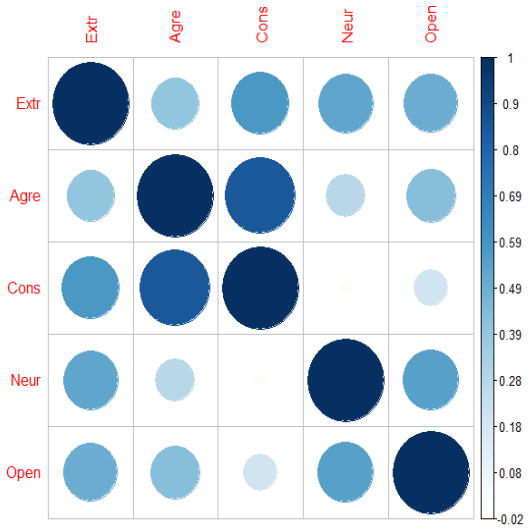

Figure 6: Correlations between big five traits

```
v_rot <- unclass(varimax(Msvd$v)$loadings)
u_rot <- as.matrix(M %*% v_rot)
```

The relationship between SVD dimensions and Big Five personality traits is defined using correlation (see Fig. 9). For example, the first topic is correlated to such traits as Extraversion and Conscientiuousness, and the fourth topic is not correlated to any trait.

```
z <- cor(u_rot, big5, use = "pairwise")
corrplot(z, method="square", is.corr = F)
```

Similarly, the relationship between topics and principal components (clusters) of Big Five data can be defined (see Fig. 10).

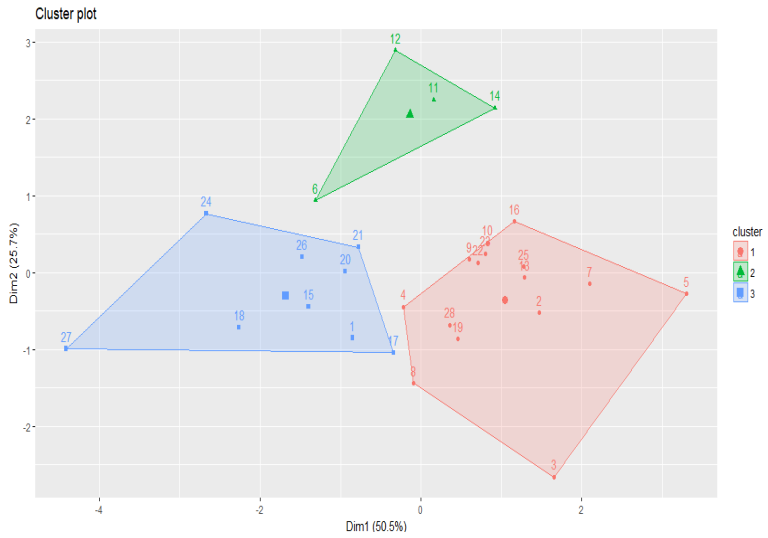

Figure 7: Visualization of cluster partitions over two first principal components

Figure 8: Visualization of cluster partitions over two variables

```
z <- cor(u_rot, pca$x, use="pairwise")
corrplot(z, method="square", is.corr = F)
```

At last, posts with the highest varimax-rotated SVD scores can be defined as below.

```
top <- list()
for (i in 1:6) {
 f <- head(order(v_rot[ ,i], decreasing = T))
 top[[i]]<-colnames(M)[f]
}
```

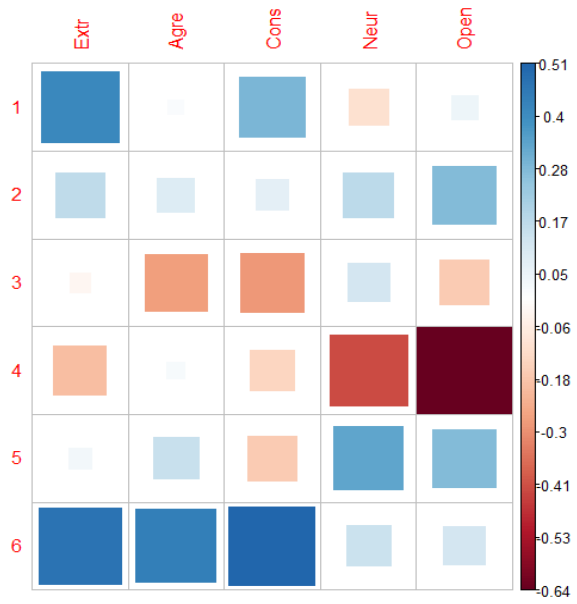

Figure 9: Correlation between topics and traits

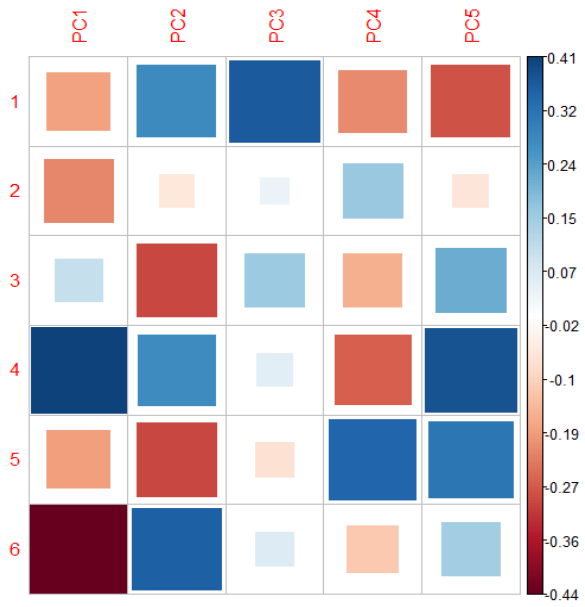

Figure 10: Correlation between topics and principal components of Big Five dataset

For example, a post with the highest score in Topic 6 (Conscientiuosness, Extraversion and Agreebleness) is P12. This post is about motivation to go forward without regard to discouragement, opposition or previous failure. A post with the highest score in Topic 5 (Neuroticism and Openness) is P17. The content of this post is clear without a word (see Fig. 11).

Figure 11: The content of Post No 17

4 Conclusion

Further study of posts with highest scores reveals the strong relationship between clusters defined through Big Five questionnaire and clusters defined through cross-liking. Thus, this experimental work confirms the results obtained by the authors of work [3] and moreover shows that these results can be obtained on smaller data sets. In other related work [4], authors conduct a series of experiments to determine the predictive power of digital footprints gathered from social media over Big 5 personality traits. They investigate how different types of digital footprints impact on prediction of user traits and behavior. Results of analyses show that the predictive power of digital footprints over personality traits is in line with the standard and can improve when demographics and multiple types of digital footprints are employed.

Therefore, the prospects for applying Data Mining methods to digital footprints analysis, are difficult to overestimate. Such an analysis could be useful for predicting student performance, for early identification of troubled students, etc. We plan that our future work will be devoted to this.

5 Acknowledgement

This work was partially supported by the Ministry of Education and Science of the Republic of Kazakhstan under grant No. AP05132933, 2018-2020 “System development for knowledge extraction from heterogeneous data sources to improve the quality of decision-making” and under grant No. BR05236340, 2018-2020 “Creation of high-performance intelligent analysis and decision making technologies for the «logistics-agglomeration» system within formation of the Republic of Kazakhstan digital economics”.

References

[1] Knight A., et al., "Systems to Harness Digital Footprint to Elucidate and Facilitate Ageing in Place." , *Studies in health technology and informatics* vol. 246 (2018): 91-101.

[2] Weaver S.D. and Gahegan M., "Constructing, visualizing, and analyzing a digital footprint." , *Geographical Review* vol. 97, no 3 (2007): 324-350.

[3] Kosinski M., et al., "Mining big data to extract patterns and predict real-life outcomes." , *Psychological methods* vol. 21, no 4 (2016): 493.

[4] Ledovaya Ya., Tikhonov R., Bogolyubova O., "Sotsialnyye seti kak novaya sreda dlya mezhdistsiplinarnykh issledovaniy povedeniya cheloveka [Social networks as a new medium for interdisciplinary research of human behavior]" , *Vestnik Sankt-Peterburgskogo universiteta. vol 16. Psihologiya. Pedagogika.* vol. 7, no 3 (2017).

[5] Kosinski M., Stillwell D. and Graepel T., "Private traits and attributes are predictable from digital records of human behavior" , *Proceedings of the National Academy of Sciences (PNAS).* vol. 110, no 15 (2013): 5802-5805.

[6] Hinds J. and Joinson A., "Human and computer personality prediction from digital footprints" , *Current Directions in Psychological Science.* vol. 2, no 8(2) (2018): 204-211.

[7] McCrae R. R. and Costa P. T., "A five-factor theory of personality. In L. A. Pervin & O. P. John (Eds.)" , *Handbook of personality theory and research* vol. 2 (1999): 139-153.

[8] Hall J. A., Pennington N. and Lueders A, "Impression management and formation on Facebook: A lens model approach" , *New Media & Society.* vol. 16, no 6 (2014): 958-982.

[9] Youyou W., Kosinski M. and Stillwell D., "Computer-based personality judgments are more accurate than those made by humans" , *Proceedings of the National Academy of Sciences.* vol. 112, no 4 (2015): 1036-1040.

[10] Azucar D., Marengo D. and Settanni M., "Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis" , *Personality and individual differences.* vol. 124 (2018): 150-159.

[11] Krainikovsky S., Melnikov M. Y. and Samarev R., "Estimation of psychometric data based on image preferences" , *WEI.* (2019): 75.

[12] Hinds J. and Joinson A.N., "What demographic attributes do our digital footprints reveal? A systematic review" , *PloS one.* vol. 13, no 11 (2018): e0207112.

[13] Chen Y. J. et al., "Predicting Consumers' Decision-Making Styles by Analyzing Digital Footprints on Facebook" , *International Journal of Information Technology & Decision Making (IJITDM).* vol. 18, no 02 (2019): 601-627.

[14] Bechmann A. and Vahlstrup P.B., "Studying Facebook and Instagram data: The Digital Footprints software" , *First Monday.* vol. 20, no 12 (2015).

[15] Sciandra A., "Social media big data: state of the art of some methodological challenges" , *Data Science & Social Research Book of Abstracts.* (2019): 117.