NEWS
OF THENATIONAL ACADEMY OF SCIENCES OF THE REPUBLIC OF KAZAKHSTAN
PHYSICO-MATHEMATICAL SERIES
ISSN 1991-346X
Volume 1, Number 335 (2021), 45 – 53 **https://doi.org/10.32014/2021.2518-1726.7**

UDC 004.89
IRSTI 28.23.37

S. Ryskulbek[1] , O. Mamyrbayev[2], A. Turganbayeva[1]

[1] Al-Farabi Kazakh National University, Almaty, Kazakhstan;
[2]Institute of Information and Computing Technologies, Almaty, Kazakhstan.
E-mail: sannyrys9@gmail.com

CLASSIFICATION OF PEOPLE BY PSYCHOLOGICAL PERSONALITY TYPES BASED ON THE HISTORY OF CORRESPONDENCE

Abstract. Temperament is a set of innate tendencies of the mind associated with the processes of perception, analysis and decision-making. The purpose of this article is to predict the psychotype of individuals based on chat stories and follow the Keirsi model, according to which the psycho type is classified as a craftsman, guardian, idealist and mind. The proposed methodology uses a version of LIWC, a dictionary of words, to analyze the context of words and uses supervised learning using KNN, SVM, and Random Forest algorithms to train the classifier. The average accuracy obtained was 88.37% for artisan temperament, 86.92% for caregivers, 55.61% for idealists, and 69.09% for rationality. When using the binary classifier, the average accuracy was 90.93% for artisan temperament, 88.98% for caregivers, 51.98% for idealism, and 71.42% for rationality.

Keywords: social networks, ML, NLP, psychotype recognition, Keirsi temperament model, classification.

1. Introduction

Personality defines a set of traits that describe a person's behavior, temperament, and emotions [13]. Personality refers to a set of traits and qualities that form a person's personality. Thus, personality prediction is of interest in the fields of healthcare, psychology, and human resources and has many commercial applications. Several studies have examined the relationship between a person's social media behavior, personality types, and psychological illnesses such as depression and post-traumatic stress Hall [17].

Social networks consist of various types of social sites, including traditional media such as Newspapers, radio and television, and non-traditional media such as Facebook, Twitter, Telegram, etc. social networks [3] Analysis is the process by which patterns can be analyzed and extracted from social media data [13]. In this context, this article develops a system for predicting the psychotypes of Telegram personalities using data in Russian [12]. The temperament model used was introduced by David Keirsey and divides temperament into four categories: guardian; idealist; craftsman; rational. For this purpose, we used the TECLA Framework, which is suitable for working with Russian-English texts. In addition, the LIWC (Linguistic Inquiry and Word Count) dictionary will be used to display contextual analysis of words by temperament [16].

This article is written as a representation on the second section, David Keirsey's substrate model used by TECLA, and section III describes the structure of TECLA. The fourth section presents the methodology and results achieved, and finally, the fifth section summarizes the results and discusses future prospects.

2. Model number Keisey

Temperament is a set of innate tendencies of the mind associated with the processes of cognition, a set of individual psychophysiological features of the individual, analysis and decision-making [1]. People are looking for success, happiness, love, pleasure, etc. In different ways and with different intensity, so they have different types of temperament [5].

Temperament is marked by its history in the sentences of the four humors described by Hippocrates, which underlie the four theories of humor for interpreting human health and disease States [5]. According to this theory, Galen (190 ad) 250 modeled the first typology of temperament [6].

The American psychologist David Keirsey directed the study of temperament to behavior, paying attention to choice, patterns of behavior, unity and consistency. According to Keirsey, psychological type is defined by the drive and interest that motivate us to live, act, move, and play roles in society [12][7] [1 Keirsey].

Artisans are usually impulsive, tend to say what comes to mind and do what suits them. Guardians talk primarily about their duties and responsibilities, as well as how well they comply with the law. Idealists usually act with a clear conscience, a pragmatic mind, behave effectively to achieve their goals, and sometimes ignore rules and customs when necessary [8][11].

Keirsi's temperament can be obtained by comparing the results of the MBTI (Myers-Briggs type indicator) test, a total of 16 psychological types [7][1][17] that classify users according to four parameters. Psychological type is an attitude as an abbreviation consisting of Latin letters starting with E and I (extroverted and introverted). S and N represent sensation and intuition, the process of perception. The letters T and F represent thinking and feeling, usually using logical reasoning, think first and feel later; and J and P represent judgment and perception, that is, relationships that reflect a person's style in the outside world.

Comparison of MBTI with the Keirsi model is performed using the Myers-Briggs classification of abbreviations, as shown in table-1 [7].

Table-1: Classification model psychotic personality of Karsi for MBTI

Keirsey	**Myers-Briggs**			
Artisan	ESTP	ISTP	ESFP	ISFP
Guardian	ESTJ	ISTJ	ESFJ	ISFJ
Idealist	ENFJ	INFJ	ENFP	INFP
Rational	ENTJ	INTJ	ENTP	INTP

3. The TECLA Framework

The TECLA framework (temperature classification framework) was developed by Lima and de Castro [11][12] to provide a modular classification tool.

Psycho is based on the models Carsi and Myers-Briggs [11]. Built in a modular format, it provides greater independence at each stage of the process and allows you to combine and test different technologies in each module [11]. Figure 1 shows the TECLA modules described in detail below.

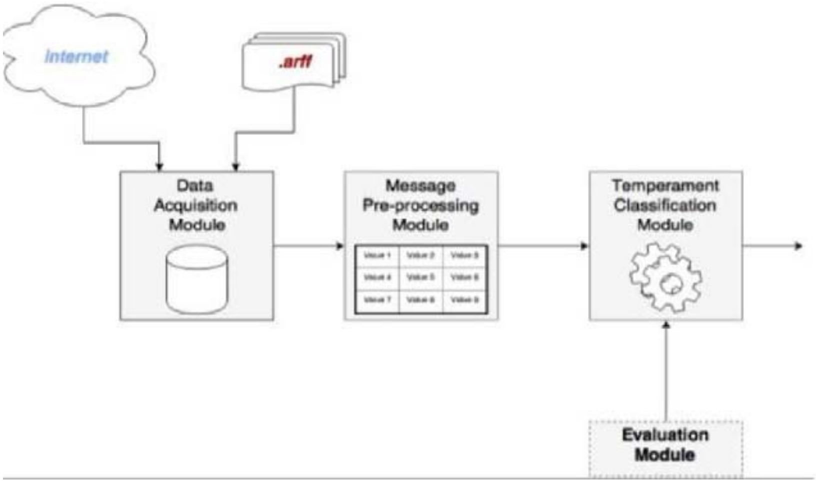

Internet
.arff
Data Acquisition Module
Message Pre-processing Module
Temperament Classification Module
Evaluation Module

Figure 1 - structure of the TECLA structure

• Data collection module: retrieves information from users to be classified, including data, identities and users, user tokens, and a set of user messages.

• Message preprocessing module: processes data by creating a matrix of objects (meta base) represented by meta properties. TECLA information is divided into two categories: grammar and behavior. Behavioral categories use information from Telegram, such as the number of messages, the number of subscribers, followers, favorites, and the number of times users were added to bookmarks. Grammatical categories use information from LIWC, MRC, Taggers, or oNLP [12].

• Module classification of psychological type: psychological type identification of users of social networks. The classification is performed on the models Carsi using a series of classifiers.

• Evaluation module: used to quantify infrastructure performance. In the version proposed in this technical article, TECLA is adapted to work with text written in Russian-English, and uses information provided by LIWC [12].

4. Methodology and results

The descriptions that will be described in this section correspond to the modular structure of the TECLA framework. First, we will explain how each module of the framework was implemented and what the result of its calculations is.

The collection of data. To test this work, we used data from an Excel file in which we collected data (the message and some attributes) from Telegram, available in Russian, and which is provided by the research center [19] for computational linguistics and psycholinguistics (CLiPS). The data set consists of: user ID; data ID; other data ID; validated data ID; MBTI results (Myers-Briggs type indicators); and gender. Data was recorded using the Telegram API [20] in the following form: each given user, the number of users, the number of favorites, and the total number of messages [10] from all users.

The source database consists of 256 user IDs. In this universe, 222 user IDs could not be collected due to an access denial, resulting in there are 213 valid user IDs left. Table 2 provides a descriptive analysis of the database based on the David Keirsey model.

Figure 2 shows the distribution of the user's psychotype. An idealistic temperament is a dominant temperament that makes up a total of 44% of the database.

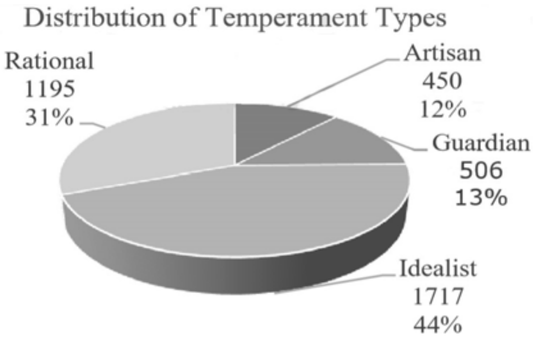


Figure 2 - Distribution of individuals according to temperament

Preprocessing and categorical analysis. At this stage, the application text is prepared. This is part of a classification algorithm consisting of special characters, spaces, numbers, symbols, URLS, tokenization, and stopword removal [4]. After that, we applied the word set technique to indicate the importance of each attribute (token) by weighing each token based on TF-IDF [2].

Another way to organize your documents is to use dictionaries such as the Language Inquiry Word Count (LIWC), which allow you to group words into psychologically relevant categories. LIWC was created by Dr. James Pennebaker to investigate the relationship between language and personality [9][15]. This is a text analysis tool that organizes documents into categories, assigning each word a corresponding category.

Using THE liwc dictionary, we calculated the frequency of words by temperament. The goal is to present the most frequently used word categories for each temperament, as shown in table 3. According to the first analysis, rational temperament usually has a higher average category frequency, which results in the guardian psychotype being taken.

In the second analysis, all substrates are more often found in the following categories: functional words, such as: for, no, very and others; pronouns such as: I, my, me and others; as well as verbs such as:

signs, occurrences and others; socialization, which is a social process, such as user storytelling and others; cognitive category gear mechanism [16]; relativity, rotation, exit and others. There are users who tend to write and hide their identity and tend to submit articles that can Express behavior and show greater awareness and logical thinking.

Another analysis showed that the ppron category for rational and idealistic temperament stands out. And ipron, present, ready, keepers and rational temperament.

This category belongs to the linguistic dimension, which tends to write more articles, pronouns, modal verbs, etc. the Category of people is more often found with an idealistic and rational temperament and tends to write about people. Influence categories are more common among guardians and idealists, and users tend to Express positive, negative, and other emotions. Finally, the admission category has a higher average frequency of temperaments of artisans and caregivers who tend to write about the body, health, and others.

The frequency of words is low, and the categories "Future", "Family", "Anxiety, vision", " Health», "Death, ""Consent," and "filler" have the same value for all temperaments.

Table 3 - Average frequency of LIWC categories by substrate.

Category	Artisan	Guardian	Idealist	Rational
funct	5,22	5,31	5,30	5,39
pronoun	1,85	1,86	1,91	1,93
ppron	1,21	1,22	1,25	1,26
i	0,39	0,36	0,42	0,41
we	0,04	0,04	0,04	0,03
you	0,66	0,70	0,66	0,68
shehe	0,64	0,67	0,65	0,67
they	0,12	0,13	0,13	0,13
ipron	1,22	1,25	1,24	1,27
article	0,77	0,80	0,78	0,81
verb	1,87	1,88	1,88	1,91
auxverb	0,67	0,69	0,68	0,70
past	0,42	0,43	0,43	0,43
present	1,12	1,13	1,12	1,15
future	0,08	0,08	0,08	0,08
adverb	0,48	0,50	0,50	0,50
preps	1,48	1,52	1,46	1,49
conj	0,91	0,90	0,92	0,93
negate	0,24	0,25	0,25	0,25
quant	0,59	0,60	0,61	0,62
number	0,14	0,14	0,15	0,15
swear	0,71	0,71	0,72	0,73
social	2,16	2,17	2,21	2,24
family	0,04	0,04	0,04	0,04
friend	0,11	0,10	0,10	0,09
humans	1,11	1,11	1,15	1,14
Affect	1,08	1,10	1,09	1,08
Posemo	0,70	0,72	0,71	0,69
negemo	0,35	0,34	0,36	0,36
anx	0,05	0,05	0,05	0,05
anger	0,14	0,13	0,14	0,15
sad	0,17	0,16	0,17	0,17
cogmech	4,06	4,14	4,12	4,18
insight	0,72	0,73	0,74	0,75

cause	0,48	0,50	0,49	0,50
discrep	0,68	0,68	0,69	0,70
tentat	0,98	0,99	1,00	1,02
certain	0,38	0,39	0,39	0,39
inhib	0,53	0,55	0,54	0,55
incl	1,40	1,42	1,40	1,41
excl	0,79	0,80	0,80	0,82
percept	0,78	0,79	0,80	0,79
see	0,26	0,26	0,26	0,26
hear	0,18	0,17	0,18	0,18
feel	0,31	0,32	0,31	0,31
bio	0,71	0,69	0,72	0,71
body	0,32	0,31	0,31	0,32
health	0,13	0,13	0,13	0,13
sexual	0,21	0,20	0,21	0,21
ingest	1,08	1,10	1,06	1,06
relativ	2,30	2,36	2,28	2,31
motion	0,76	0,76	0,74	0,76
space	0,98	1,01	0,97	0,99
time	0,98	1,02	0,96	0,96
work	0,23	0,25	0,23	0,24
achieve	0,46	0,49	0,46	0,47
leisure	0,31	0,32	0,31	0,31
home	0,06	0,06	0,06	0,05
money	0,29	0,30	0,29	0,31
relig	0,08	0,09	0,09	0,08
death	0,06	0,06	0,06	0,06
assent	0,13	0,13	0,13	0,13
nonfl	0,26	0,27	0,26	0,27
filler	0,03	0,03	0,03	0,03

Classification. During the experiment, 4% of the total data set was randomly selected and used because of the size of the processed matrix and the availability of computing resources. To perform substrate classification, we used the following classifiers available in Scikit-learn [14]. KNN; SVM; and random forest. Each temperament is divided into binary problems, as suggested by Lima and de Castro [12]. For testing, cross-validation of folders 6 and 10 was used, and accuracy, accuracy, recall, and F-measures were calculated. For the KNN classifier using K-nearest neighbor feature classification, K = 1, K = 2, K = 3, and cosine similarity were used to determine neighbors. The tests were divided into liwc [19] and TF-IDF [20] dictionary.

Table 2 - telegram users ' Temperament and data (A = artisan, G = guardian, I = idealist, R = rational)

	A	G	I	R	Total
Users	450	506	1.717	1.195	3.868
Datas_Statuses_Count	12.343.807	15.648.860	65.593.286	45.198.150	138.784.103
Data _Base	674.211	738.755	2.570.646	1.751.624	5.735.236
Friends	168.893	225.371	825.969	640.529	4.012.741
Favorites	1.768.903	2.371.924	10.006.749	6.683.984	1.860.762
Contact_num	292.413	423.549	1.497.093	1.799.686	4.012.741

5. LIWC

Table 4 shows the results of testing the 10-fold launch of 6 folders in TECLA. The values shown in bold represent the best average accuracy and F measurements obtained by the classifier for each temperature.

For the artisan's temperament, the KNN algorithm with K = 1 gave an average accuracy of 80.44% and an F-measure of 88.91%. The better the performance, that is, the more accurate the function markup, the better is the SVM algorithm with an average accuracy of 88.37%, followed by a random forest with an average accuracy of 87.95%. SVM had better average accuracy and 100% responsiveness, while Random Forest had better accuracy than SVM.

As for the guardian's temperament, the most obvious prediction comes from the SVM algorithm with an average accuracy of 86.92% and a measure of F. This was followed by a 93% random forest and an average accuracy of 86.32%. The lowest average accuracy (78.36%) was found in KNN with K = 1.

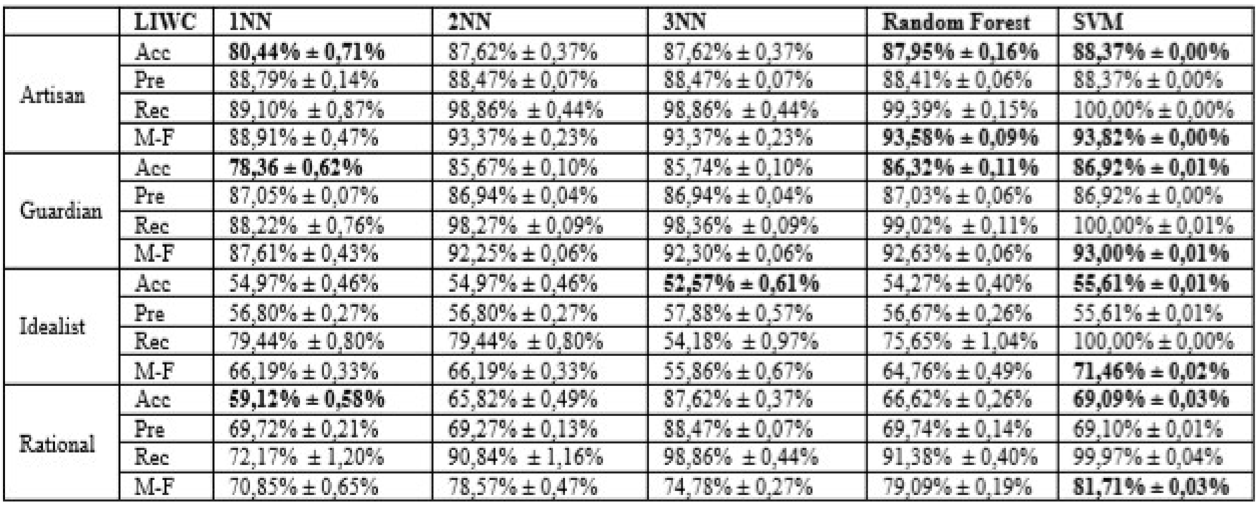
SVM showed the best results even with an ideal and reasonable temperament.

The idealistic substrate had an average accuracy of 55.61%, an F measurement of 71.46%, a rational substrate of 69.09%, and an f measurement. These two substrates exceeded the average accuracy and were suitable for marking objects .50%. Overall, the SVM has the highest accuracy for all temperaments, but for artisan and guardian temperaments, Random Forest showed an average accuracy very close to SVM.

6. TF-IDF

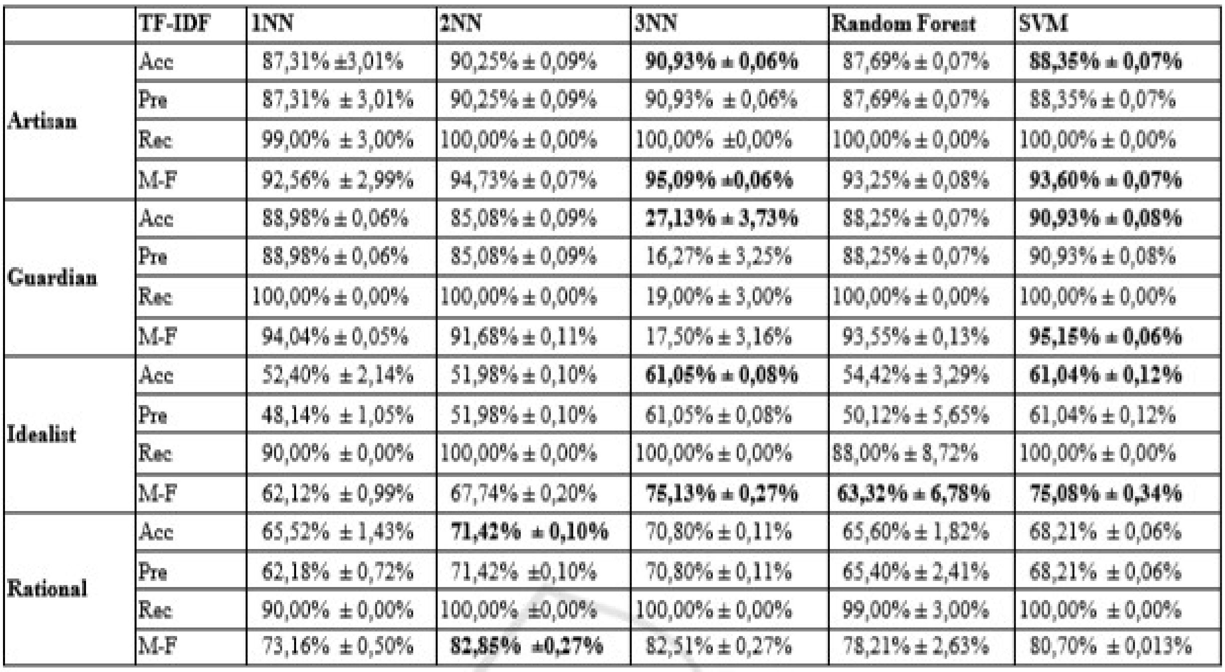
Table-5 shows the results of cross-checking the execution of 10 folders 10 times in a TECLA environment. The values shown in bold represent the best average accuracy and the results of the F measurement obtained using the binary classifier for each temperature. Artisan temperament achieved high average accuracy (90.93%) with KNN, K = 3 and F-measurement of 95.09%, which makes the results obtained by KNN reliable (K = 3). The accuracy is 88.35%. SVM had the highest average accuracy for sentinel temperament, but in this case KNN for K = 3 performed very poorly. One hypothesis about this low value is an imbalance in the database, so when you increase the number of neighbors, the algorithm cannot label objects. For the idealistic temperament, SVM and KNN were practical (K = 3). For rational temperament, the best performance was 82.85% of the average accuracy for the KNN algorithm with K = 2. Once again, SVM turned out to be the algorithm that showed the best average performance among those tested.

Table 4 - Accuracy (Acc), precision (Pre), re-call (Rec), and F-measurement (M-F)
for 4 media using 6 folders and 10 replays.



	LIWC	1NN	2NN	3NN	Random Forest	SVM
Artisan	Acc	**80,44% ± 0,71%**	87,62% ± 0,37%	87,62% ± 0,37%	**87,95% ± 0,16%**	**88,37% ± 0,00%**
	Pre	88,79% ± 0,14%	88,47% ± 0,07%	88,47% ± 0,07%	88,41% ± 0,06%	88,37% ± 0,00%
	Rec	89,10% ± 0,87%	98,86% ± 0,44%	98,86% ± 0,44%	99,39% ± 0,15%	100,00% ± 0,00%
	M-F	88,91% ± 0,47%	93,37% ± 0,23%	93,37% ± 0,23%	**93,58% ± 0,09%**	**93,82% ± 0,00%**
Guardian	Acc	**78,36 ± 0,62%**	85,67% ± 0,10%	85,74% ± 0,10%	**86,32% ± 0,11%**	**86,92% ± 0,01%**
	Pre	87,05% ± 0,07%	86,94% ± 0,04%	86,94% ± 0,04%	87,03% ± 0,06%	86,92% ± 0,00%
	Rec	88,22% ± 0,76%	98,27% ± 0,09%	98,36% ± 0,09%	99,02% ± 0,11%	100,00% ± 0,01%
	M-F	87,61% ± 0,43%	92,25% ± 0,06%	92,30% ± 0,06%	92,63% ± 0,06%	**93,00% ± 0,01%**
Idealist	Acc	54,97% ± 0,46%	54,97% ± 0,46%	**52,57% ± 0,61%**	54,27% ± 0,40%	**55,61% ± 0,01%**
	Pre	56,80% ± 0,27%	56,80% ± 0,27%	57,88% ± 0,57%	56,67% ± 0,26%	55,61% ± 0,01%
	Rec	79,44% ± 0,80%	79,44% ± 0,80%	54,18% ± 0,97%	75,65% ± 1,04%	100,00% ± 0,00%
	M-F	66,19% ± 0,33%	66,19% ± 0,33%	55,86% ± 0,67%	64,76% ± 0,49%	**71,46% ± 0,02%**
Rational	Acc	**59,12% ± 0,58%**	65,82% ± 0,49%	87,62% ± 0,37%	66,62% ± 0,26%	**69,09% ± 0,03%**
	Pre	69,72% ± 0,21%	69,27% ± 0,13%	88,47% ± 0,07%	69,74% ± 0,14%	69,10% ± 0,01%
	Rec	72,17% ± 1,20%	90,84% ± 1,16%	98,86% ± 0,44%	91,38% ± 0,40%	99,97% ± 0,04%
	M-F	70,85% ± 0,65%	78,57% ± 0,47%	74,78% ± 0,27%	79,09% ± 0,19%	**81,71% ± 0,03%**

Table 5 - Accuracy (Acc), Precision (Pre), Recall (Rec), and F-measure (M-F)
for four temperaments using 10 folders and 10 iterations



	TF-IDF	1NN	2NN	3NN	Random Forest	SVM
Artisan	Acc	87,31% ±3,01%	90,25% ± 0,09%	**90,93% ± 0,06%**	87,69% ± 0,07%	**88,35% ± 0,07%**
	Pre	87,31% ± 3,01%	90,25% ± 0,09%	90,93% ± 0,06%	87,69% ± 0,07%	88,35% ± 0,07%
	Rec	99,00% ± 3,00%	100,00% ± 0,00%	100,00% ±0,00%	100,00% ± 0,00%	100,00% ± 0,00%
	M-F	92,56% ± 2,99%	94,73% ± 0,07%	**95,09% ±0,06%**	93,25% ± 0,08%	**93,60% ± 0,07%**
Guardian	Acc	88,98% ± 0,06%	85,08% ± 0,09%	**27,13% ± 3,73%**	88,25% ± 0,07%	**90,93% ± 0,08%**
	Pre	88,98% ± 0,06%	85,08% ± 0,09%	16,27% ± 3,25%	88,25% ± 0,07%	90,93% ± 0,08%
	Rec	100,00% ± 0,00%	100,00% ± 0,00%	19,00% ± 3,00%	100,00% ± 0,00%	100,00% ± 0,00%
	M-F	94,04% ± 0,05%	91,68% ± 0,11%	17,50% ± 3,16%	93,55% ± 0,13%	**95,15% ± 0,06%**
Idealist	Acc	52,40% ± 2,14%	51,98% ± 0,10%	**61,05% ± 0,08%**	54,42% ± 3,29%	**61,04% ± 0,12%**
	Pre	48,14% ± 1,05%	51,98% ± 0,10%	61,05% ± 0,08%	50,12% ± 5,65%	61,04% ± 0,12%
	Rec	90,00% ± 0,00%	100,00% ± 0,00%	100,00% ± 0,00%	88,00% ± 8,72%	100,00% ± 0,00%
	M-F	62,12% ± 0,99%	67,74% ± 0,20%	**75,13% ± 0,27%**	**63,32% ± 6,78%**	**75,08% ± 0,34%**
Rational	Acc	65,52% ± 1,43%	**71,42% ± 0,10%**	70,80% ± 0,11%	65,60% ± 1,82%	68,21% ± 0,06%
	Pre	62,18% ± 0,72%	71,42% ±0,10%	70,80% ± 0,11%	65,40% ± 2,41%	68,21% ± 0,06%
	Rec	90,00% ± 0,00%	100,00% ±0,00%	100,00% ± 0,00%	99,00% ± 3,00%	100,00% ± 0,00%
	M-F	73,16% ± 0,50%	**82,85% ±0,27%**	82,51% ± 0,27%	78,21% ± 2,63%	80,70% ± 0,013%

7. Conclusion

Temperament affects how we perceive the world and react to it. Understanding temperament is very important in our lives and is important for correctly positioning ourselves in the market. In General, you can determine your temperament using a test such as the Myers-Briggs type indicator (MBTI). The hypothesis of this study is that only with the help of data obtained from a person's social networks, it is possible to passively determine their psychotype. To do this, we used a database containing MBTI results from Telegram users. These data were used to create a model for predicting temperament. Documents (data) are a RUSSIAN liwc dictionary, in which words are grouped by category. Word frequency calculations were performed to show the categories most commonly spoken by artisans, keepers, idealistic and rational psychotypes. In this analysis, you can determine the user's email trends, perception, among other things, related to the topics that are most identified. The data was structured using LIWC and TF-IDF. For classification via LIWC, the best accuracy results were obtained for the temperament of the artisan and guardian trained with SVM. For TF-IDF, the highest average accuracy was for artisan, guardian, and idealist temperament, and the SVM algorithm was also highlighted. For expressions using TFIDF, the best average accuracy was observed for the artisan and guardian temperament of the KNN (K = 3) and SVM algorithms.

As a future work, we plan to conduct a case study using the TECLA framework with a database of a set of volunteer users, so that the data can be used to train TECLA by answering the MBTI test form and sharing a social profile. Make a frame and sort it by temperament. Another improvement is to examine the content of the document to find out how much the imbalance base hinders obtaining these results, to understand why the accuracy of the classifier is low and whether processing of an unbalanced class is required.

С. Рыскулбек[1], О.Ж Мамырбаев[2], А. Р. Турганбаева[1]

[1]Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан;
[2]ҚР БҒМ БК Ақпараттық және есептеуіш технологиялар институты, Алматы, Қазақстан

АДАМДАРДЫ ХАТ АЛМАСУ ТАРИХЫ НЕГІЗІНДЕ ПСИХОЛОГИЯЛЫҚ ТҰЛҒА ТҮРЛЕРІ БОЙЫНША ЖІКТЕУ

Аннотация. Темперамент дегеніміз – қабылдау, талдау және шешім қабылдау үдерісіне байланысты ақыл-ойдың туа біткен тенденцияларының жиынтығы. Мақаланың мақсаты – әлеуметтік желілердегі қарым-қатынас негізінде жеке адамның психотипін болжау және Кейрси моделіне сүйену, соған сәйкес психотип шебер, қамқоршы, идеалист және ақыл ретінде жіктеледі. Ұсынылған әдістемеде LIWC нұсқасы, сөз жиынтығы, сөз контексін талдау үшін KNN, SVM және Random Forest алгоритмдерін бақылау арқылы классификатор оқытуды көздейді. Темпераментпен жұмыс істеу үшін орташа есеппен алынған дәлдік қолөнершілерге 88,37%, тәрбиешілерге 86,92%, 55,61% идеалистер мен рационалдылық үшін 69,09% дәлдікті құрайды. Екілік жіктеуішті қолданған кезде орташа дәлдік қолөнершінің темпераменті үшін 90,93%, тәрбиешілер үшін 88,98%, идеализм үшін 51,98% және ұтымдылық үшін 71,42% құрады.

Түйін сөздер: әлеуметтік желілер, ML, NLP, психотипті тану, Кейрси темпераментінің моделі, классификация.

С. Рыскулбек[1], О.Ж Мамырбаев[2], А. Р. Турганбаева[1]

[1]Казахский национальный университет им. аль-Фараби, Алматы;
[2] Институт информационных и вычислительных технологий КН МОН РК, Алматы, Казахстан

КЛАССИФИКАЦИЯ ЛЮДЕЙ ПО ТИПАМ ПСИХОЛОГИЧЕСКОЙ ЛИЧНОСТИ НА ОСНОВЕ ИСТОРИИ ПЕРЕПИСКИ

Аннотация. Темперамент – это совокупность врожденных склонностей ума, связанных с процессами восприятия, анализа и принятия решений. Цель этой статьи – предсказать психотип людей на основе чатов и следовать модели Кейрси, согласно которой психотип классифицируется как ремесленник, опекун, идеалист и ум.

Предлагаемая методология использует версию LIWC, набор слов для анализа контекста слов и использует контролируемое обучение с использованием алгоритмов KNN, SVM и Random Forest для обучения классификатора. Средняя полученная точность сотавила 88,37% для темперамента ремесленника, 86,92% – для воспитателей, 55,61% – для идеалистов и 69,09% – для рациональности. При использовании бинарного классификатора средняя точность составила 90,93% для темперамента ремесленника, 88,98% – для воспитателей, 51,98 – для идеализма и 71,42% – для рациональности.

Information about authors:

Turganbayeva A., Senior Lecturer, Al-Farabi Kazakh National University, Almaty, Kazakhstan; turalma@mail.ru; https://orcid.org/0000-0001-9723-4679;

Mamyrbayev O., PhD, Associate Professor, head of the Laboratory of computer engineering of intelligent systems at the Institute of Information and Computational Technologies, Almaty, Kazakhstan; morkenj@mail.ru; https://orcid.org/0000-0001-8318-3794;

Ryskulbek S., Master's degree student, Al-Farabi Kazakh National University, Almaty, Kazakhstan; sannyrys9@gmail.com, https://orcid.org/0000-0002-8711-4398

REFERENSES

[1] Calegari M. D., & Gemignani O. H. (2006). Temperamento e Carreira. São Paulo: Summus Editorial.

[2] Feldman R., & Sanger J. (2007). The Text Mining Handbook: Advanced approaches in Analyzing Unstructured Data. Cambridge university press.

[3] Gundecha P., & Liu H. (2012). Mining social media: a brief introduction. New Directions in Informatics, Optimization, Logistics, and Production. Informs, pp. 1-17.

[4] Haddi E., Liu X., & Shi Y. (2013). The role of text preprocessing in sentiment analysis. Procedia Computer Science, 17, 26-32.

[5] Hall C. S., Lindzey G., & Campbell J. B. (2000). Teorias da Personalidade. Porto Alegre: Artmed.

[6] Ito P. d., & Guzzo R. S. (2002). Diferenças individuais: temperamento e personalidade; importância da teoria. Estudos de Psicologia, pp. 91-100.

[7] Keirsey D. (1998). Please Undestand Me II: Temperament, Character, Intelligence. Prometheus Nemesis Book Company.

[8] Keirsey D. M. (1996). Keirsey.com. (Corporate Offices) Acesso em 12/10/2017 de 10 de 2017, disponível em ://www.keirsey.com/4temps/overview_temperaments. Asp

[9] Komisin M. C., & Guinn C. I. (2012). Identifying personality types using document classification methods. In: FLAIRS Conference.

[10] Kwak H., Lee C., Park H., & Moon S. (2010). What is Twitter, a social network or a news media? Proceedings of the 19th international conference on World wide web. ACM., 591-600.

[11] Lima A. C. (2016). Mineração de Mídias Sociais como Ferramenta para a Análise da Tríade da Persona Virtual. São Paulo.

[12] Lima A. C., & de Castro L. N. (2016). Predicting Temperament from Twitter Data. Advanced Applied Informatics (IIAI-AAI), 2016 5th IIAI International Congress on. IEEE.

[13] Nor Rahayu N., Zainol Z., & Yoong T. L. (2016). A comparative study of different classifiers for automatic personality prediction. Control System, Computing and Engineering (ICCSCE), 2016 6th IEEE International Conference on. IEE, 435-440.

[14] Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., . . . Perro, M. (2011). Scikitlearn: Machine learning in Python. Journal of Machine Learning Research 12.Oct, 12, 2825-2830.

[15] Pennebaker J. W., & King L. A. (1999). Linguistic styles: language use as an individual difference. Journal of personality and social psychology, 77(6), 1296.

[16] Pennebaker J. W., Boyd R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015.

[17] Plank B., & Dirk H. (2015). Personality Traits on Twitteror-How to Get 1, 500 Personality Tests in a Week. Proceedigs of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2015), pp. 92-98.

[18] Spencer J., & Uchyigit G. (2012). Sentimentor: Sentiment analysis of twitter data. Proceedings of European conference on machine learning and principles and practice of knowledge discovery in databases.

[19] Verhoeven B., Daelemans W., & Plank B. (2016). Twisty: A Multilingual Twitter Stylometry Corpus for Gender and Personality Profiling. Proceedings of the 10th International Conference on Language Resources and Evaluation.

[20] Xavier O. C., & Carvalho C. L. (2011). Desenvolvimento de Aplicações Sociais A Partir de APIs em Redes Sociais Online. UFG. Goiânia.