Machine learning of software agents for classification sociotypes

Serhii Holub[1,†], Oleksandr Morushko [2,†], Nataliia Khymytsia[2,*,†], Maria Holub[1,†], and Marek Aleksander[3]

[1] *Cherkasy State Technological University, 460 Shevchenko Boulevard, Cherkasy, 18006, Ukraine*
[2] *Lviv Polytechnic National University, 12 Stepana Bandera Street, 79013, Lviv, Ukraine*
[3] *State Higher Vocational School in Nowy Sacz, z, Staszica st. 1, Nowy Sacz, 33300, Poland*

Abstract

For the first time, it was proposed to use the information technology of intellectual monitoring to classify the authors of printed texts by sociotypes. The relevance of the work is determined by the need to automate the processes of describing the properties of the authors of messages, in particular in social networks, in order to develop controlling influences on a group of individuals and individuals. There is a relatively detailed description of the properties of a person belonging to one or another sociotype. This allows for the development of effective management influences. The use of this tool of information warfare aims to protect the information space of Ukraine from hostile actions. Research tasks are formalized as the task of classifying authors of printed texts. Sociotypes were used as classes. Each of the classes contained texts by 3 authors who had already passed the examination and their belonging to one of the sociotypes was established. To build text classifier models, one of the methods of machine learning is used - the method of group consideration of arguments (Group Method of Data Handling). To conduct the experiment, a monitoring software agent was built to perform Text Mining tasks. For each class of messages, a separate dictionary of features was formed and an array of numerical characteristics of texts based on these features was built, which is called the Array of Input Data. The agent was trained on massive input data describing the texts of 2 authors belonging to each class. Testing of agent models was based on the numerical descriptions of the 3rd author of each of the classes. In this way, the testing of models took place on the basis of data that did not participate in the creation of models. As a result of testing, all texts and, accordingly, their authors have been correctly classified. The hypothesis about the possibility of classifying people according to sociotypes through the intellectual analysis of their text messages has been experimentally confirmed.

Keywords

Intelligent monitoring, classification, sociotypes, software agents, machine learning.

SCIA-2024: 3rd International Workshop on Social Communication and Information Activity in Digital Humanities, October 31, 2024, Lviv, Ukraine
* Corresponding author.
† These authors contributed equally.
s.holub@chdtu.edu.ua (S. Holub); oleksandr.o.morushko@lpnu.ua (O. Morusko); nhymytsa@gmail.com (N. Khymytsia); sashasokolovsksa92@gmail.com (M. Holub); aleksandermarek4@gmail.com (M. Aleksander)
0000-0002-5523-6120 (S. Holub); 0000-0001-8872-2830 (O. Morusko); 0000-0003-4076-3830 (N. Khymytsia); 0000-0002-0553-8163 (M. Holub); 0000-0003-2619-1063 (M. Aleksander)

1. Introduction

The main purpose of monitoring software agents is to provide decision-making processes with information in the relevant subject area. An agent approach to the construction of intellectual monitoring tools allows to increase the variety of processes of extracting knowledge from the results of observations of research objects.

A person is a complex object of research. Obtaining information about the properties of this object and their change is an urgent task when planning many strategies in modern information wars. Today, the Ukrainian information space, in particular social networks, is a testing ground for both attack and manipulation technologies, as well as methods and means of protecting the self-awareness of Ukrainians. Printed texts are one of the means of conveying governing influences.

Each person in one way or another has a selective attitude to what surrounds him. She chooses to use certain texts that meet her requirements. In this sense, a selective assignment of verbal material is carried out, which a person structures, supplements, and enriches in the future. Many great writers and poets created their texts from the social mass of language. But these texts are again appropriated both by their contemporaries and descendants, spiritually enriching and developing [1].

One of the possible methods of studying verbal and non-verbal behavioural reactions is the use of socionic analysis. This method is based on the so-called Jungian framework. It contains four pairs of dichotomous traits, namely: extraversion-introversion, logic-ethics, sensory-intuition, rationality-irrationality. By consistently identifying the dominant dichotomous trait in each of the four pairs, we can determine the psychological type of personality. Since there is a considerable amount of scientific research with a detailed description of the features inherent in a particular psychological type, the information obtained as a result of such typing can be of practical use, in particular when choosing the optimal use of one's abilities.

It should be noted that the psychological type of personality is an innate property of each individual. It does not change throughout a person's life.

There are 16 different psychological types in total. They will demonstrate completely different behavioural reactions under stressful conditions. At the same time, individuals with the same (identical) psychological type will behave in almost the same way under the same conditions.

Socionics provides tools for identifying the properties of an individual person, classifying people according to these properties, and constructing mass self-awareness management technologies. Knowing the type of person allows you to actualize your abilities, functionally apply your natural inclinations, and achieve maximum self-realization. On the other hand, an incorrect definition of a person's sociotype can provoke the emergence of additional complexes.

In the context of sociology, these types have a lot in common. In particular, each of them has certain characteristic facial features and similar behavioural reactions, especially in extreme situations. In addition, since identical socionic types have exactly the same mental model, they also profess similar values. Therefore, there is a high probability of a reliable determination of a psychological type by marker words and the general tone of communication.

It is clear that such a determination will be correct only if we can compare verbal and non-verbal signals. This is important in order to exclude the possibility of social mimicry, when someone tries to disguise themselves as another type. In this case, verbal signals will always lag

behind non-verbal signals. If we don't have this option, then when communicating exclusively verbally (in particular, through social media or audio communication), we should focus on the speed of communication. The less time a person has to think about their statements or responses, the more likely they are to work 'in their own way'.

Based on this assumption, the following conclusions can be drawn:

- the speech characteristics of identical psychological types in communication are similar;
- certain characteristic words and methods of communication inherent in each sociotype can be identified;
- computer technologies can be used to reliably recognise a sociotype.

The latter conclusion is extremely important, since the use of computer technologies to recognise the psychological profile of the interlocutor can lead to significant competitive advantages. At the same time, the range of applications for such technologies is extremely wide - from hiring and managing personnel to modelling various types of negotiations. The main advantage is that the computer program that will perform the recognition functions has the ability to constantly improve and expand its capabilities based on the array of data that is constantly being analysed. The more this programme operates, the more accurate the analysis results it will provide.

The initial stage of such a programme may be the selection of texts by several representatives of the same sociotype for initial synthesis by the programme and the creation of an algorithm for recognising similar texts. When we have created 16 different text recognition programs (according to the number of available basic sociotypes), we can start the process of identifying an unknown text by recognising the sociotype of its author.

The purpose of this work is to create and test a method of machine learning of software agent models designed to classify authors of printed texts by sociotypes. The results of the work will be used to build virtual robots that will automate the processes of monitoring large volumes of text messages on social networks and simplify the processes of identifying the properties of their authors.

2. Processes of classification of authors by sociotypes

Today, it is common to use machine learning models for intellectual analysis of text in order to solve the problem of attribution - determining their authorship [2], [3], [4], [5]. The task of text classification is also referred to in the literature as typing. Solving the problem of correct typing is very important. Knowing the type of person allows you to actualize your abilities, functionally apply your natural inclinations and achieve maximum self-realization. On the other hand, an incorrect definition of a person's sociotype can provoke the emergence of additional complexes.

There are three main socionic methods of determining a person's type: visual, verbal and testing. They are based on the use of Jung's basis - the definition of the dominant trait in four dichotomous pairs: extraversion-introversion, logic-ethics, sensory-intuition, rationality-irrationality. The best results are obtained by the simultaneous use of all these methods [6]. But in practice, such a comprehensive approach is rarely used, so one of these methods has to be preferred.

The visual definition is based on the features of the physique, depending on the combination of four dichotomous features. It is not difficult to calculate that the number of different possible combinations is 16. This number is the number of basic socionic types. Therefore, experts who carry out typing can also belong to any of these types. Therefore, it is difficult to determine which visual signs are dominant and which are additional. To a certain extent, this leads to the subjectivity of the evaluation, since the probability of the determination largely depends on the level of the expert's qualification.

Verbal definition consists in the analysis of speech features of the subjects under study. You should pay attention to the tempo rhythm of speech, its phonetic coloring, information saturation and emotional coloring. All this, under the condition of live communication, should be compared with non-verbal reactions, which are also quite informative.

Testing as a research method can be used only with the prior consent of the subject. In addition, the reliability of the obtained results when using this method largely depends on the objectivity and truthfulness of the answers to the test questions. That is why to verify the results obtained as a result of testing, a further interview is used, during which methods of visual and verbal analysis are used.

The main problem that should be solved is the maximum reliability of the classification results. Therefore, the question arises whether it is possible to use the capabilities of information technologies for such an analysis, which would significantly reduce the factor of subjectivity when determining a person's sociotype.

In our opinion, the most adaptive for computer processing and analysis can be the use of machine learning models to analyze the features of the text written by the author of the sociotype, which was previously established by a group of experts. This text should have a sufficient number of characters (at least 500-1000) so that the software agent can find out the speech features of each of the 16 basic sociotypes.

Psycholinguistic classification of texts is studied by many authors. In scientific research [7] the concept of the psychiatrist K. Leonhard, who first investigated the clinical features of accented personalities and pointed out the nature of the accentuation of characters in artistic texts, is represented. The novelty of the article is that it for the first time synthesizes the opinions of scientists regarding the implementation of the accentuation theory, substantiates the importance of the accentuation approach for the analysis of linguistic factage. On the basis of the generalized results, the scientists came to a conclusion about the dynamics of the mentioned problem, as well as determined the possibility of its deepening and implementation in new interpretive linguistic directions. The authors present a classification of texts based on the theory of psychoaccentuation proposed by V. Belyanin. This scientist singles out a system of texts, offers a psycholinguistic classification of them: they are "light", "dark", "funny", "sad", "beautiful", "complex" texts. In the scientist's metalanguage, metaphorized terms for delineating text types appeared.

Galustyan O.A., Zakharenko L.M. and Kazmirenko V.O. in their work consider psycholinguistic text analysis as a method of identification of an unidentified person. Determine the profile of the author of texts in social Internet networks and propose a general scheme for compiling the profile of an unknown author of a written text [8].

According to the classification proposed in the study [9] the corpus is divided into classes determined by application-specific criteria. The authors demonstrate tutorial examples that associate data points with labels that indicate they belong to a particular class. Specifically,

training examples extracted from the political news dataset can have one of three labels that assign them to one of the classes: "senate," "congress," and "legislature." The "With the teacher" classifier model is trained on these educational examples. At the next stage, the test examples that remained unmarked are placed into one of these class-categories using a classifier model.

The history of machine learning methods for classifier models of short texts from social networks is presented in the paper [10].

In the study [11] the definition of the psychological type of a person through social networks with the help of sociological analysis is described. Methods of visual and verbal determination of a person's sociotype are used. The practical examples considered by the authors clearly demonstrate that knowing certain features of the body or the style of human speech with a high probability helps to predict how they can behave in certain situations.

In the paper [12] proposed to determine the human psyche using four pairs of Jung's dichotomous signs. For this, the authors suggest using the following methods of analysis: visual, verbal, and, if possible, testing. An algorithm for using these methods to form teams, assess their cohesion, and determine the optimal leader is proposed.

A work describing the process of using machine learning technologies to classify authors of printed texts by sociotypes could not be found.

3. Research tasks and hypotheses

The purpose of the work is formalized as the task of classifying the authors of the texts. Sociotypes are used as classes.

16 classes are given, the name of which coincides with the name of sociotypes. Each of the classes contains the texts of messages from 3 authors, which are expertly assigned to one of the sociotypes.

It is necessary to assign the author of the new text to one of the available classes.
To solve the problem, it is necessary to build a dictionary of classification features, according to which the authors of other texts would be assigned to one of the sociotype classes with an acceptable level of adequacy.

Since it is not possible to build a dictionary of classification signs in an explicit form, it was decided to build a decisive rule.

The following hypotheses were formulated to fulfil the tasks:

1. Classification of persons according to belonging to one of the sociotypes is possible through intellectual analysis of their text messages.
2. The decisive rule should combine the classification features into a single system, the uniqueness of which is the ability to classify the author of the text.
3. To build a decision rule, it is advisable to use one of the Text Mining methods described in [13]. This method has already been used to solve the task of classifying authors of texts [14], but other agent tasks were performed. Thus, we are talking about adapting a known method to perform a new agent task.

4. Research and discussion of their results

An experiment was conducted to test the proposed hypotheses. One of the monitoring software agents, which is an element of the structure of the virtual robot, described in [15]. To build a

decision rule in the form of a model-classifier, the agent synthesizer of models used a multi-line algorithm of the method of group consideration of arguments Group Method of Data Handling (GMDH) [16].

16 classes have been identified for training. Each of the classes contained texts by 3 authors, regarding which their belonging to a given sociotype was established Table 1.

Characteristics of classes provides a list of classes for model training.

Table 1.
Characteristics of classes

№ in order	Class	Author
1	Logic-Intuitive Extrovert	Jack London
2		Steve Jobs
3		Stephen King
4	Sensory-Ethical Extrovert	Mykhailo Gorbachev
5		Alexander Pushkin
6		Leo Tolstoy
7	Intuitively-Logical Introvert	Mykhailo Zhvanetskyi
8		Niccolo Machiavelli
9		Viktor Saltykov
10	Ethical-Sensory Introvert	Volodymyr Vysotskyi
11		Marilyn Monroe
12		Ivan Turgenev
13	Sensory-Logical Extrovert	Volodymyr Ulyanov
14		Mykhailo Lomonosov
15		Napoleon Bonaparte
16	logic-Sensory Introvert	Mykhailo Bulgakov
17		Yosyp Dzhugashvili
18		Vasyl Shukshin
19	Ethical-	Mykola Gogol
20	Intuitive Extrovert	Mykhailo Lermontov
21		Faina Ranevska
22	Intuitively-Ethical Introvert	Ivan Goncharov
23		Sergey Yesenin
24		Mykhailo Kalinin
25	Intuitively-Logical Extrovert	Karl Marx
26		Yury Nikulin
27		Oleksandr Suvorov
28	Sensory-Ethical Introvert	Evgeny Leonov

29		Mykola II Romanov
30		Antoine de Saint-Exupéry
31	Ethical-Sensory Extrovert	Victor Hugo
32		Jules Verne
33		Ernesto Che Guevara
…	…	…

The dictionary of features was built according to the adaptive method described in 0. When building an array of input data (AID), this method provides an individual list of features in the dictionary for each of the classes. All other texts are described by features. Figure 1: A fragment of the input data array shows a fragment of the AID created for the Sensory-Logical Extrovert class from the individual dictionary created for the class of texts, which is designated as "Own".

Class	a	b	c	d	e	f	g	…
…	…	…	…	…	…	…	…	…
100	51	22	33	23	26	65	5	…
100	56	8	39	7	34	85	12	…
100	54	13	44	13	19	77	12	…
100	52	22	39	12	21	66	6	…
100	62	20	36	12	30	76	11	…
-100	49	22	40	11	27	102	10	
-100	54	9	22	15	29	65	6	
…	…	…	…	…	…	…	…	…

Figure 1: A fragment of the input data array

The texts of two authors from the respective classes were used to train the models. Learned models were tested on an array of input data formed from the features of the texts of the third author of each of the classes (Figure 2: Results of testing the classifier model for the "Sensory-Logical Introvert" sociotype).

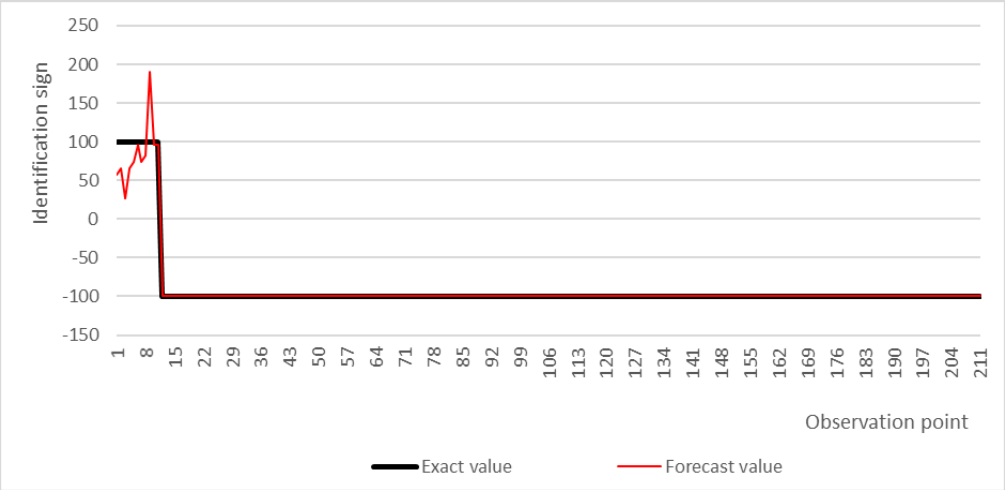

Figure 2: Results of testing the classifier model for the "Sensory-Logical Introvert" sociotype

The results of the test of the classifier model show that the software agent managed to separate the observation points of the Array of Input Data, which describe the texts of authors belonging to the "Sensory-Logical Extravert" sociotype, from the tests whose authors belong to other sociotypes.

A separate classifier model was built for each class. These models were tested on the Array of Input Data, which was built from the features of the texts of third authors. The test of the Array of Input Data contained 211 observation points. The results of testing agent-classifier models are presented in Table 2.

Table 2.
Results of testing classifier models

№ in order	Class	Number of correctly classified points,%
1	Logical-Intuitive Extrovert	100,00
2	Sensory-Ethical Extrovert	100,00
3	Intuitive-Logical Introvert	100,00
4	Ethical-Sensory Introvert	100,00
5	Sensory-Logical Extrovert	100,00
6	Logico-Sensory Introvert	100,00
7	Ethical-Intuitive Extrovert	100,00
8	Intuitive-Ethical Introvert	100,00
9	Intuitive-Logical Extrovert	100,00
10	Sensory-Ethical-Introvert	100,00
11	Ethical-Sensory Extrovert	97,96
12	Logico-Intuitive Introvert	100,00
13	Logico-Sensory Extrovert	98,98
14	Sensory-Logical Introvert	100,00
15	Ethical-Intuitive Introvert	100,00
16	Intuitive-Ethical Extrovert	100,00

All classifier models have completed their task. The vast majority of models correctly recognized all points of the test Array of Input Data. This means that each of these models is able to recognize the author's text belonging to a given sociotype.

5. Conclusions

The sociotype to which a person belongs can be determined by intellectual analysis of his text message. For this, a monitoring software agent and machine learning technology of its models are used.

An experimentally confirmed hypothesis about the construction of a decisive rule by combining the numerical characteristics of the features of the printed text into a single system of the model-classifier features in the form of a classifier model.

It is concluded that the use of the method of adaptive formation of a feature dictionary when converting printed text into an array of its numerical characteristics allows ensuring the necessary informativeness of the input data set and synthesise useful classifier models.

The results of the experiments presented in Table 2 prove the effectiveness of the application of machine learning technology of software agent models to determine the sociotype of a person based on his printed message.

References

[1] T. S. Yatsenko, The problem of verbalisation and cognition of systemic characteristics of the unconscious content of the subject's psyche. Professional education: pedagogy and psychology (1999) 391 - 402.

[2] E. Stamatatos, A survey of modern authorship attribution methods. Journal of the American Society for information Science and Technology, Vol. 60, no 3, (2009), 538-556. doi:10.1002/asi.21001.

[3] R. Zheng, Li. J. Chen, Z. Huang, A framework for authorship identification of online messages: Writing-style features and classification techniques. Journal of the American society for information science and technology, Vol. 57. №. 3, (2006), pp. 378-393. doi:10.1002/asi.20316.

[4] M. AlSallal, R. Iqbal, V. Palade, S. Amin, V. Chang, An integrated approach for intrinsic plagiarism detection. Future Generation Computer Systems, Vol. 96., (2019) pp. 700-712. doi:10.1016/j.future.2017.11.023.

[5] B. Alhijawi, S. Hriez, A. Awajan, Text-based authorship identification - A survey. Paper presented at the 5th International Symposium on Innovation in Information and Communication Technology, ISIICT 2018. (2018), pp. 1-7. doi:10.1109/ISIICT.2018.8613287.

[6] O. O. Morushko, N.O. Khymytsia, R.O. Holoshchuk, Modeling of social information exchange: study guide. Lviv, Publishing House of Lviv Polytechnic, (2023).

[7] V. Papish, et al. Origins and development of the theory of psychoaccentuation in East Slavic linguistics: a fragmentary review (end of the 20th–beginning of the 21st centuries). Acta Polono-Ruthenica, (2021), 3.XXVI: 101-118.

[8] O. A. Galustyan, L.M. Zakharenko, V.O. Kazmirenko (under the editorship of O.I. Motlyakh), Compilation of a psychological profile of an unidentified person based on the characteristics of his written text. Kyiv, (2020).
[9] C. C. Aggarwal, Text classification: basic models. In Machine Learning for Text, Springer, Cham. (2022). https://doi.org/10.1007/978-3-030-96623-2_5
[10] N. Khymytsia, S. Holub, M. Holub, O. Morushko, Personality prediction from social networks: a review of works, CEUR Workshop Proceedings, (2022), 3296, 72–82.
[11] O. Morushko, N. Khymytsia, N. Shakhovska, Determining the psychological portrait of members of web communities through socionic analysis, CEUR Workshop Proceedings, (2020), 2616, 112–124.
[12] O. Morushko, N. Khymytsia, V. Teslyuk, Remote selection of staff based on socionic analysis of social network, Content CEUR Workshop Proceedings, (2022), 3171, 138–149.
[13] S. V. Holub, H. I Martynova, M. S. Holub, Modelyuvannya dialektnoho tekstu v tekhnolohiyi bahatorivnevoho informatsiynoho monitoringu, Matematychni mashyny i systemy, (2016), № 4, 76-83.
[14] S. Holub, N. Khymytsia, M. Holub, O. Morushko: Machine learning of the classifier of authors of social network messages, SCIA-2022, 1st International Workshop on Social Communication and Information Activity in Digital Humanities, October 20, (2022), Lviv, Ukraine. Ceur-ws.org.Vol-3296.paper11.
[15] D. Tolbatov, S. Holub, Construction of models of monitoring agents on several reference forms. CEUR Workshop Proceedings Volume 3126, (2021), 108-112, Short Paper Proceedings of the 2nd International Conference on Intellectual Systems and Information Technologies (ISIT 2021) co-located with 1st International Forum "Digital Reality" (DRForum 2021) Odesa, Ukraine, September 13-19, 2021.
[16] A.G. Ivakhnenko, Inductive method of self-organization of models of complex systems. Kiev, Naukova Dumka, (1981).
M. Holub, O. Piven, Classification of texts in the technology of multilevel information monitoring. Inzynier XXI wieku, Monografia, Wydawnictwo naukowe Akademii Techniczno-Humanistycznej w Bielsku-Bialej, (2016), 119-122.