Досліджується проблема розробки ефективного способу визначення авторства текстів (на матеріалі публікацій відомих українських журналістів). Більшість наявних методів потребують попередньої обробки тексту, що тягне за собою нові витрати при розв'язанні поставленої задачі. У випадку, коли кількість можливих авторів можна мінімізувати, такий підхід є часто надлишковим. Ще одним недоліком наявних підходів є те, що переважна більшість їх застосовувалися до іншомовних текстів і не враховували особливостей української мови. Тому було вирішено розробити підхід, що дозволяє визначити автора тексту українською мовою без попередньої обробки та дає високі результати точності, а також встановити, які типи штучних нейронних мереж забезпечують мінімальну похибку для українських публіцистів.

Розроблений метод використовує багатошаровий персептрон прямого поширення, алгоритм навчання з учителем, векторизацію HashingVectoriser, оптимізатор Adam. Визначено, що при невеликій кількості ітерацій (4–5 ітерацій) навчання штучної нейронної мережі отримується досить висока точність визначення авторства публіцистичних текстів та досить мале значення похибки. Використано більше 1000 фрагментів текстів трьох українських авторів. У результаті проведених експериментів було встановлено, що застосовування розробленого підходу до розв'язання поставленої задачі дає змогу досягти досить високих результатів. У текстах, що містять не менше 500 символів, точність сягає 91 %, а максимальна кількість ітерацій навчання штучної нейронної мережі при цьому не перевищує 15. Такі результати досягнуті насамперед завдяки ефективному підбору методу векторизації на підготовчому етапі та структури штучної нейронної мережі

Ключові слова: визначення авторства, аналіз тексту, штучні нейронні мережі, багатошаровий персептрон, векторизація тексту

UDC 681.518
DOI: 10.15587/1729-4061.2020.195041

IDENTIFICATION OF AUTHORSHIP OF UKRAINIAN-LANGUAGE TEXTS OF JOURNALISTIC STYLE USING NEURAL NETWORKS

M. Lupei
Postgraduate Student*
E-mail: maxim.lupey@gmail.com
A. Mitsa
PhD, Associate Professor, Head of Department*
E-mail: alex.mitsa@gmail.com
V. Repariuk *
E-mail: reparuck@ukr.net
V. Sharkan
PhD, Associate Professor
Department of Journalism**
E-mail: vasylsharkan@gmail.com
*Department of Information Management Systems and Technologies**
**Uzhhorod National University
Narodna sq., 3, Uzhhorod, Ukraine, 88000

Received date 03.12.2019
Accepted date 06.02.2020
Published date 28.02.2020

1. Introduction

With the advancement of technology, artificial neural networks are increasingly being used to solve certain tasks that take a lot longer for a person than a computer to solve. Some of such relevant issues include the identification of the primary source, identification of the authorship of anonymous texts, fight against plagiarism, determining belonging of a text to a certain author during legal expert examination. At present, there are many approaches to solving them, based on different methods, and yielding different results of accuracy. However, the issue of developing a universal method that will produce the best results, that is, will provide the highest accuracy in authorship identification with the consumption of fewer resources, remains unresolved.

Identification of authorship of Ukrainian publicists who have a similar style with the use of computational approaches is the problem that has been little explored. The approaches to its solution can have their own peculiarities and differ from the generally accepted universal approaches. That is due to the fact that the main parameters, which researchers try to separate, are significantly affected by some general language features, rather than by individual features of the author's language style. Given exactly the specificity of a particular language, one can develop an effective approach to solving the problems of authorship identification.

2. Literature review and problem statement

In Ukrainian linguistics, the greatest attention is traditionally paid to studying the belles-lettres style and less – to other styles, including journalistic. This also concerns studying the individual style of authors: more studies deal with the language of writers than with the language of publicists [1]. The study of Ukrainian linguists refers mainly to the speech of individual authors or the language of specific works, whereas little attention in traditional linguistics is paid to the issue of identification of authorship of texts.

A series of papers, which explore the stylometric parameters of scientific texts in Ukrainian have been published in recent years. In particular, in paper [2], the identification of the author's style is based on comparative analysis of the author's speech coefficients: speech coherence, vocabulary diversity, syntax complexity indices, concentration and exclusivity regarding the author's passage and other analyzed passages for further comparison and determining the degree of belonging of the analyzed text to a particular author. Support-vector machines (SVM) are used at one of the stages. A disadvantage is a small sampling of about 200 single-author papers of the technical area written by about 100 different authors. In article [3], the dynamics of a change of different parameters of the author's style (the number of different language units – words in a text, sentences, prepositions, conjunctions, the number of words with frequency 1 and with frequency 10 or more) is traced. The methods developed in these works of the scientific style can be applied to other functional styles of the Ukrainian language.

A comparative study of statistical parameters of the publicist ("newspaper") and other styles is presented in a number of articles. Thus, article [4] aims to address the issue of distinction of the scientific, fiction, journalistic and conversational styles at the phonological level using mathematical-statistical methods based on the data about the use of consonants in the texts by English-speaking authors, and in article [5], the statistical parameters of conversational and journalistic styles are examined separately. However, these studies are performed using the material of the English language. In this regard, there remains a need to develop the optimum structure of the ANN to identify the authorship of texts the journalistic style in the Ukrainian language.

The approaches to authorship identification can combine accumulated knowledge from the theory of image recognition, mathematical statistics and probability theory, neural networks, cluster analysis, Markov chains, and others [6–11]. Paper [6] studies the state of the problem today; it is noted that if there are texts by 3–4 authors in the training and testing samples, trained classifiers confidently demonstrate up to 85 % of the accuracy of identification of authorship of a text in the test sample. Article [6] proposed duplex architecture with the use of the support vector machine (SVM). Paper [7] examined the problem of recognition of short texts taken from the Internet in order to detect criminals. Given the fact that the messages are quite short, it was important to have a lot of information about possible candidates. The approach in paper [7] was based on the use of a support vector machine (SVM), but the preparatory stage required the use of Stylemetric Analysis. In paper [8], the problem of identifying the authorship of e-mails for 12 people, each of which created 10 letters up to 150 words, was explored. The accuracy, in this case, was 75–80 %. The method *K* of closest neighbors was used for recognition. In article [9], two approaches based on multiple-discriminant analysis (MDA) and support vector machine (SVM) were proposed. Their effectiveness for various problems, including the identification of authorship of disputable texts of the collection of papers "Federalist", is compared, and the issue of the authorship of the Message to Hebrews in the New Testament is explored. Some problems are solved quite effectively, but at the same time, the research lacks clear recommendations on a universal approach to text authorship identification, which for any tasks will work equally effectively. In paper [10], the authors used support-vector machines (SVM), the machine learning algorithm that designs the plane through a multidimensional hyperspace, dividing incident training into target classes, and Platt sequential minimum optimization (SMO). This method also combines various functions of text analysis and does not require its prior processing. This method solves the problem of recognition of belonging of a text fragment to Bengali authors Rabindranath Tagore and Sarat Chandra Chattopadhyay. However, it uses a rather small sampling (only 1–2 texts by each author) and solves the problem of recognizing if a text fragment belongs to the certain work rather than the problem of authorship identification. Transferring this approach to the identification of authorship of Ukrainian publicists will not give such good results. In paper [11], it was proposed to use neural networks based on complex neural cells with generalized activation functions. Article [12] showed that the use of neural networks with two-threshold activation functions can significantly improve the recognition ability of the network. In paper [13], the approaches using stylistic features were proposed as more reliable style markers than, for example, the units of the lexical-semantic level, because stylistic markers are less consciously controlled by the author. Three known indicators based on statistical similarity are used to obtain the individual effect: cosine-similarity (COS), ChiSquare measure (CS) and Euclidean distance (ED). The model of machine learning includes three different modules: decision trees (DT), neural networks (NN), and support vector machines (SVM). However, even the most effective of these methods give results (83.3 %), which can be improved. Study [14] deals with the problem of authorship identification among 55 classics of the world literature and proposes to use the probabilistic approach for it. The identification accuracy for various authors differs significantly, although for some of them it is quite high and exceeds 90 %. The known problem of Marlow-Shakespeare, in which the used method refutes the hypothesis that Christopher Marlow is the co-author of the early plays of his peer Shakespeare, is considered separately. In article [14], it is proposed to use Kullback-Leibler Divergence (KLD) to identify authorship. The effective combination of vectorization and the architecture of artificial neural networks makes it possible to reduce computation costs and obtain high accuracy in the problems of identification of authors [15]. This process should be also considered for authorship identification among well-known Ukrainian publicists.

The considered approaches are adjusted to the specifics of each of the problems. This allows arguing that by working out the approach that uses the peculiarities of a particular problem, it is possible to improve the accuracy of authorship identification, in particular, for journalistic texts in Ukrainian.

3. The aim and objectives of the study

The aim of this study is to determine the effective combination of the vectorization method and the artificial neural nets structures to identify the authorship of texts of journalistic style in the Ukrainian language.

To achieve the aim, it is necessary to consider different methods of text vectorization and different architectures of neural networks, to determine their most effective combinations, to find out its accuracy, as well as the speed of the ANN learning process for the problem of authorship identification of journalistic texts in the Ukrainian language.

4. Method for studying the authorship identification of a journalistic text in the Ukrainian language with the help of ANN

The texts by Y. Makarov, I. Losev, O. Pokalchuk in the quantity of 50 texts by each author published in the "Ukrainian Week" and "Weekly Mirror Ukraine" during 2015–2019 (by chronological principle from the latest to the oldest ones) were chosen for analysis. The total number of word usage in the studied texts is more than 150 thousand (Y. Makarov – 32,758, I. Losev – 61,556, O. Pokalchuk – 72,286). The largest number of word usage is in the texts by O. Pokalchuk, the smallest – in the texts by Y. Makarov. Using a specially created program, the texts (including the headings, however, without specifying the name of the author) were divided into fragments (their total number is 1,194), not less than 500 characters, but to the end of the sentence. Multilayer neural networks of direct propagation Multi-Layer Perceptron [16], an algorithm for supervised learning [17, 18] and a corresponding technology stack (Table 1) were used for authorship identification.

Table 1

The stack of used technologies

Programming languages	Python 3.6, C++
Architecture and learning algorithm	ANN of direct propagation, error backpropagation learning
Libraries	Numpy, Matplotlib, Tensor-Flow, Pandas, HashingVectorizer
Implementation environment	Jupiter on Google Cloud

Perceptron implements the function $f():R^m \rightarrow R^1$ through training on a dataset where m is the input data size, 1 – output size [19]. The algorithm of error backpropagation will be used as the learning algorithm [20]. Having obtained and processed the input data $X=(x_1, x_2, ..., x_n)^T$ output y, construct a nonlinear function approximator for classification or regression; however, there may be a certain number of layers (hidden layers) between the input and the output layer. Perceptrons with a different number of layers (Fig. 1) are considered. Each neuron in a hidden layer converts the value from the previous layer with the weight line adding $w_1x_1+w_2x_2+...+w_nx_n$ with a non-linear activation function, $f(.):R^1 \rightarrow R^1$, which is implemented via *relu* or *sigmoid*. At the output, there appears a mean value of output signal y. The algorithm of error backpropagation will be used as the learning algorithm.

The developed approach works according to the following scheme (Fig. 2).

The initial stage is to search for the texts on the Internet to form the dataset. The selected texts undergo further processing using the program written in the C++ programming language. This program removes unnecessary spaces and blank parts of a text and forms a .csv file where all information is split and grouped according to the predetermined parameters. The artificial neural network subsequently works with this. The next step is vectorization, which is an extremely important section, and the result of the work is very sensitive to the conducted stage of input data vectorization. Then, the texts are divided into learning and training ones. In the flowchart, it is the K-fold record, where K indicates how many times the texts were divided into testing and learning ones in various ways. Further, the artificial neural network learning, which implements the testing identification of the text's author and gives results, works with the texts; it all repeats the definite number of times. In the end, the results of experiments are collected and accuracy and research error are determined.

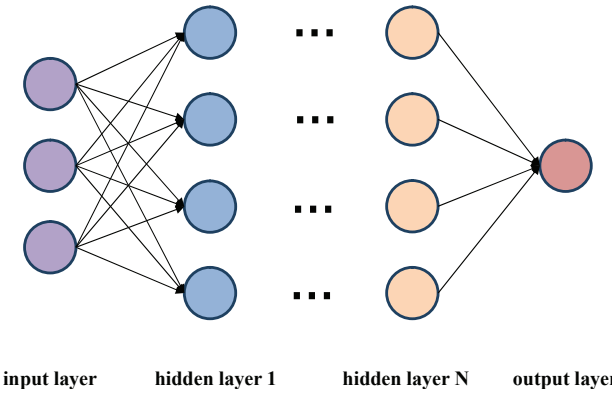

Fig. 1. Structure of artificial neural network

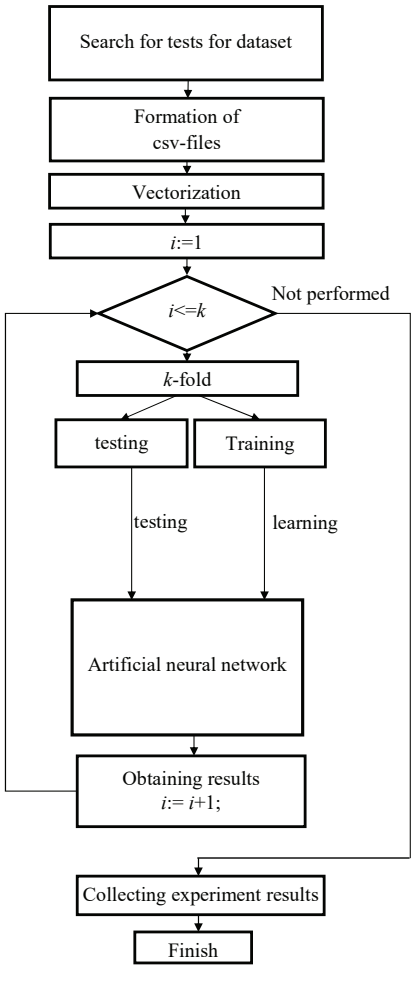

Fig. 2. General scheme of operation of the developed approach

We will separate the problem of finding effective vectorization, which plays a very important role in creating an effective method [15]. Such vectorization methods as ASCII Converter, Simple Vectorizer, and Hashing Vectorizer were considered. The Hashing Vectorizer demonstrated much

better results in conjunction with the considered architectures of artificial neural networks. This vectorization method operates at two stages. At the first stage, we convert the collection of text documents into the frequency matrix of lexemes. In the second stage, we convert the collection of text documents into a 2-dimensional matrix. It contains a count of the number of lexemes (or binary information about occurrence), normalized as the frequency of a lexeme, if norm='11' or projected on a Euclidean unit of a sphere if norm='l2'. This implementation of the text vectorizer uses a hash-trick to find the name of the marker line to display the integer index.

The Adam Optimization Algorithm for Deep Learning is also used. This optimization algorithm based on the first-order gradients of the stochastic target function rests on adaptive estimates of the moments of lower order. The method is simple to implement, efficient to apply and has low memory requirements, invariant to the diagonal scaling of gradients and is well suited for problems that have large sizes of data or parameters. This method is also suitable for non-stationary purposes and problems associated with very noisy or sparse gradients. Hyperparameters for this method have intuitive interpretation and usually require little customization. Empirical results show that the Adam is effective in practical application and shows better results compared to the methods of stochastic optimization. During the research, the variant of this method, which is based on the infinity norm, was used.

The issue of the number of iterations to perform while training the artificial network is very important. This study was carried out and the results were displayed in Fig. 3. We see how accuracy increases and error decreases, depending on the number of training iterations. As one can see, accuracy improves as early as on the 4^{th}–5^{th} iterations and does not increase so rapidly.

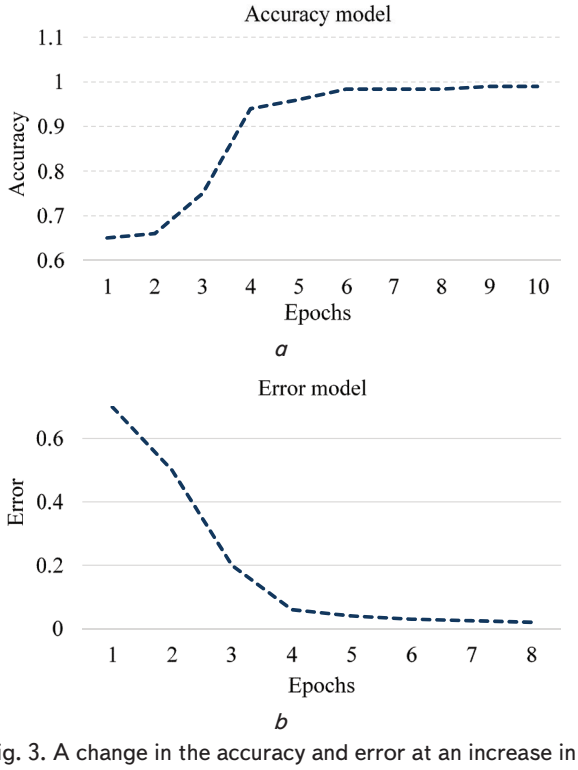

Fig. 3. A change in the accuracy and error at an increase in the number of iterations: *a* — increase in accuracy; *b* — decrease in error

It should be noted that the number of iterations required to achieve the desired accuracy can be reduced by 10–15 % through the use of polynomial neurons [21] in the first hidden layer of the network.

The basic parameters and toolsets, which were used during computational experiments, are assigned in Table 2.

Table 2

Basic parameters and toolsets

Sample	Characters in a text	Vector dimensionality	Converter	Model	Optimizer	loss
0:1193	>500	10000	Hashing Vectorizer	keras. Sequential	adam	binary cross entropy

The explored basic parameters and toolsets are constant in all experiments. All experiments use the sample of 1,194 fragments of texts, which are subsequently divided into two datasets – training and testing. These datasets are divided into completed sentences with the dimensionality of not less than 500 characters. From them, we form vectors of dimensionality of 10,000, which are fed to the input of a neural network. The Hashing Vectorizer algorithm is used in the formation of these vectors. In all experiments, the model of the neural network keras.Sequential, the Adam Optimization Algorithm for Deep Learning and loss function binary cross entropy, are used.

5. The results of the study of the Ukrainian-language journalistic text belonging to a particular author

As a result of the conducted computational experiment, we determined the effectiveness of using different structures of artificial neural networks for the problem of identification of belonging of a text in the journalistic style to one of the three authors. The results of the operation of the developed software are given in Table 3.

Table 3

Results of experiments

No.	*K*-FOLD	Structure of neurons of the ANN	Maximal number of iterations	VALUES	ACCURACY
1	1|10	[10, relu][10, relu] [1, sigmoid]	15	$F1 \approx 0.8609$	0.9161
2	1|10	[10, relu] [10, relu][10, relu] [1, sigmoid]	15	$F1 \approx 0.8275$	0.9027
3	1|10	[10, relu] [1, sigmoid]	15	$F1 \approx 0.8565$	0.9128
4	1|10	[10, relu][10, relu] [10, sigmoid] [1, sigmoid]	15	$F1 \approx 0.8231$	0.9011
5	1|2	[10, relu][10, relu] [1, sigmoid]	15	$F1 \approx 0.6896$	0.8374
6	1|5	[10, relu][10, relu] [1, sigmoid]	15	$F1 \approx 0.8539$	0.9111
7	1|5	[10, relu][10, relu] [1, sigmoid]	5	$F1 \approx 0.5787$	0.8029
8	1|5	[10, relu][10, relu] [1, sigmoid]	10	$F1 \approx 0.8058$	0.8893

Table 3 shows that depending on the changes of certain parameters or architectures of the network, results become bet-

ter or worse. The most effective architecture of neurons in the ANN was determined based on the database from the studied fragments of texts by 3 authors (the texts are evenly distributed in alphabetical order beginning with the first word). The 10-fold cross-validation revealed that the architecture of neurons of ANN [10, relu][10, relu][1, sigmoid] is the most effective. The total number of characters in the texts is more than 500, the maximum number of iterations is 15, the average accuracy is 0.9161, and the average error is *f*1: 0.8609. It was experimentally proved that such parameters in the sample are most effective (computational experiment No. 1) among other parameters.

Consider separately for each author the accuracy of authorship identification using the developed method. In addition, check the influence of dimensionality of the vectors, which are fed to the input into the neural network, on the accuracy of authorship identification.

As Table 4 shows, the authorship of publicist I. Losev was identified most effectively. We also see that at an increase in the dimensionality of vectors from 10000 to 100000, accuracy significantly increases in the identification of authorship of Pokalchuk from 0.9347 to 0.9559. For other authors, an increase in dimensionality affects the accuracy insignificantly.

Consider how the authorship of I. Losev was identified using the example of samples from the dataset.

As Table 5 shows, the fragment of the text from the dataset with index 115 was not recognized with the help of the developed method. Identification by experts-philologists of certain stylistic features of an author in such fragments of texts would make it possible to take it into consideration in the developed method and to increase the recognizability for particularly for him.

Table 4

Results of experiments separately for each author

Author	*K*-FOLD	Dimensionality	Structure of neurons of ANN	Maximal number of iterations	VALUES	ACCURACY
Losev	1|10	100,000	[10, relu][10, relu][1,sigmoid]	15	*F*1≈0.9633	0.9661
Losev	1|10	10,000	[10, relu][10, relu][1,sigmoid]	15	*F*1≈0.9617	0.9644
Makarov	1|10	100,000	[10, relu][10, relu][1,sigmoid]	15	*F*1≈0.7013	0.8688
Makarov	1|10	10,000	[10, relu][10, relu][1,sigmoid]	15	*F*1≈0.6856	0.8603
Pokalchuk	1|10	100,000	[10, relu][10, relu][1,sigmoid]	15	*F*1≈0.9545	0.9559
Pokalchuk	1|10	10,000	[10, relu][10, relu][1,sigmoid]	15	*F*1≈0.9331	0.9347

Table 5

Example of the text to be recognized

Index in dataset	Example of text	Is it Losev?	How did the method identify?
1	2	3	4
55	А сам верховний, вочевидь, не дуже й наполягав. Тоді як такі нестандартні кадрові рішення були успішно апробовані в інших країнах, приміром у США під час війни між Північчю та Півднем, коли доброволець, простий американець із північних штатів Улісс Ґрант зібрав загін патріотів і почав воювати. Він брав міста південців одне за одним. Ґрант мав суттєву ваду: схильність до спиртних напоїв. Доносчики повідомили президентові Авраму Лінкольну про той факт. Але Лінкольн, який дав Ґранту звання бригадного генерала, знав, що це найуспішніший полководець його армії.	1	1
302	Є класична істина, що окремо взята людина, будь-якого рівня, завжди розумніша (це не означає — моральніша, чесніша, праведніша тощо) за групу в частині осмислення прийнятих рішень. Але далі тема розпадається на професійну дискусію про видове і родове виживання. Вона нуднувата, однак головна ідея в тому, що як себе людина позиціює — особистістю чи частиною цілого, така їй і карма. Довгі роки українські політтехнологи жили з цією максимою душа в душу, і, в принципі, все працювало. Імітація змін при тотальних лінощах і саботажі, взагалі, всіх влаштовувала, і навіть Захід, нашого годувальника.	0	0
364	Знадобилося півстоліття, щоб переконатися: загадки не вирішуються такими грубими інструментами, а геній точних наук, навіть нобелівський лауреат, не мусить претендувати автоматично на роль мислителя-гуманітарія (хоча претензії спостерігаються регулярно). Розбіжності між расами справді існують. Статистично. Хоча нині в Америці, зокрема, саме поняття «раса» вважається застарілим, в академічному світі радше оперують категорією «популяція». Ці розбіжності стосуються й таких особливостей, як когнітивні (пізнавальні) здібності.	0	0
820	Зі знанням — як зі словом 'реформи'. Під 'перевернути' всі автоматично розуміють 'змінити на ліпше'. Нас із дитинства навчали: що більше знаєш, то ліпше. Не пояснюючи в подробицях, кому, власне кажучи, від цього ліпше, наскільки й чому. А залежить від того, хто навчає. Знання — такий самий товар, як і все інше. Факт — це сировина, інформація — напівфабрикат. А ми вже прилаштовуємо інформації свою особисту рамочку, ставимо в домі на почесне місце й віддаємо шану. Якщо почесті соціально схвалені (як у культах і релігіях), то знання має шанси набути соціальної цінності, і його можна комусь продати.	0	0
821	Розрізняється сім видів інформаційних війн: за командування і контроль, війни розвідок, електронна, хакерська та кібервійни (так, вони всі різні), психологічна, за економічну інформацію. Кожна з них ведеться на трьох рівнях (особистому, корпоративному, глобальному) й у двох сферах — цивільній і військовій. Держава має потребу в інформаційній безпеці, як кістки організму потребують кальцію. Держава — кістяк країни, суспільство — її плоть. Але що з того, коли добре кальцинований череп скалиться тобі з історичної могили, якщо ти не Гамлет?	0	0

Continuation of Table 5

1	2	3	4
899	Театральні ефекти замість наближення до адекватного розуміння ситуації. В Україні є окремі експерти, проте немає експертного середовища як сталої специфічної спільноти зі своїм етичним кодексом, своїми внутрішніми нормами, принципами, визнаними правилами верифікації (перевірка інформації та публічних тверджень, блокування фейків), зі своєю ієрархією репутацій тощо. Між іншим, постійні проблеми із соціологічними фундаціями в Україні також пов'язані з відсутністю соціологічного середовища, коли численні окремі соціологи та компанії існують у режимі вільнохаотичного буття, без механізмів саморегуляції та самоконтролю, поза професійно-громадською думкою, без атмосфери цеху професіоналів, без взаємної цехової відповідальності.	1	1
920	Тобто група схожим чином формує своє уявлення про соціальний успіх, і члени групи прагнуть йому відповідати. Культурні люди, спілкуючись переважно з собі подібними, інколи дивуються, чому в процесі групових заклинань щось шукане не з'являється. Але не дуже. Бо процес їм замінює результат. Люди результату виганяються з культурного середовища як бездуховні й цинічні, що, взагалі-то, правда. Але середовище від цього не змінюється, продовжуючи живитися праною й амрітою власного виготовлення. І ось культурний наратив зіштовхується з тим-таки снарядом соціальних макропроцесів, і все розмітається в емоційне клоччя.	0	0
845	Сплеск обурення з приводу антиукраїнської пропаганди, тиражованої низкою впливових і не дуже мас-медіа, нагадує волання про харасмент вельми дорослих голлівудських (і просто політичних) актрис. Змістовно проблукавши років сорок по лос-анджелеській пустелі, вони раптово усвідомили, що увесь цей час жили в муках сорому і страху, які терміново потрібно монетизувати. Більш як чверть століття наші претенденти на месіанство діловито топталися на місці зі своїми особистими скрижалями. За цей час кремлівське інформаційне сміття засіяло наш інформаційний ландшафт такими териконами брехні, що вони стали сприйматися так само природно, як Карпати.	0	0
115	Але тільки злагодженість дій здатна забезпечити їх ефективне функціонування. Яким чином українське громадянське суспільство в конкретному вищезгаданому естонсько-українському проекті є не тільки сприймаючою стороною, а й такою що віддає? У рамках спільної протидії російським інформаційним операціям того ж таки 2016-го було створено спільну мережу, що об'єднує українських експертів з їхніми колегами з країн Балтії та деяких інших східноєвропейських країн. Вона дістала назву 'Ліга стійкості' (Resilience League) і негайно викликала пропагандистську істерику в інформаційних ресурсах, які працюють на Кремль.	0	1

6. Discussion of results of studying the identification of authorship of a journalistic text in the Ukrainian language

The developed approach after training makes it possible to verify online quickly and effectively belonging of a text or its fragment to each of the three studied publicists.

The proposed simple scheme (Fig. 2) enabled achieving high accuracy in the conducted research into the identification of authorship of journalistic texts in Ukrainian. This can be conditioned by several factors. First of all, an effective combination of the vectorization method and the ANN structure, which showed the highest results (Tables 3, 4) was chosen automatically from a large number of possible variants. In addition, the authors selected for analysis are some of the most well-known contemporary Ukrainian publicists, who have a rather unique and recognizable writing style and whose works are published in the leading Ukrainian editions. The stylistic expressiveness of the studied texts is also determined by peculiarities of the journalistic style: an author has a better opportunity to express his "self" in comparison, for example, with the official-business or scientific style.

Given this, it is appropriate in the future to test this method of research on a broader empirical material. Firstly, a greater number of authors (perhaps, with not so pronounced author's individual language peculiarities) and their journalistic texts should be selected. Secondly, it might be interesting to take the texts belonging to other styles (scientific, belles-lettres).

The specific features of the style of a particular author are traditionally separated in linguistics, but the reverse process (authorship identification by a text fragment) without using the ANN is a very difficult task. Traditional methods of studying the individual style of authors in linguistics are based on the general features, identified by a person (at the lexical, morphological, and syntactical levels). The ANN itself chooses the criteria of distinguishing the authors, and these criteria can be based on the phenomena not noticeable for a linguist. Therefore, the next stage of the study is to trace the features that help the ANN to identify the authorship.

Unlike the approaches in other studies, the proposed approach at the preparatory stage does not identify any specific elements that will act as the parameters in authorship identification. This process is fully passed on to the vectorization procedure and the ANN. It is the vectorization procedure that converts input texts into the matrix, which records the frequency parameters of a lexeme [15]. The main advantage of this approach is that large volumes of input datasets are well-scaled because there is no need to keep a dictionary in memory. The shortcomings include the fact that there is no possibility to calculate the reverse conversion (from characteristics indices to line names), which can be a problem when trying to perform self-analysis – figuring out what features are the most important for the model. In addition, there may occur some coincidences: separate lexemes can be displayed in one function index. However, in practice, this rarely happens if n_features is large enough (for example 2^{18} for problems of text classification). Note that an increase in the dimensionality of an input vector is directly proportional to the accuracy of the model.

By developing this direction, we plan to increase the number of input data and of authors, to use other types of ANN and input data processing.

7. Conclusions

Correctly chosen text vectorization at the preparatory stage makes it possible to increase substantially the effectiveness of identification of authorship of Ukrainian-language texts of the journalistic style with the help of the ANN. At the same

time, it allows reducing computational costs and avoiding routine procedures, which are often carried out at the preparatory stage. Among the considered vectorization methods, such as ASCII Converter, Simple Vectorizer and Hashing Vectorizer, the most effective was the Hashing Vectorizer method for all the considered architectures of artificial neural networks. Among the considered architectures of neural networks, architecture [10, relu][10, relu][1,sigmoid].proved to be the most effective for identification of the authorship of journalistic Ukrainian-language texts. The developed approach ensures high accuracy (for one of the authors it exceeds 96 %), which corresponds to the level of the most effective methods. The advantages also include a small number of iterations of artificial neural network learning (4–5 iterations). The approach can contribute to more effective identification of authorship of texts written in flexion languages, including Ukrainian. At the same time, the results of the ANN operation (qualification of a text as belonging or not belonging to a certain author) present interesting empirical material for further linguistic studies, since it is important to find out in the context of philology, by which language elements it is possible to determine correctly whether a text belongs to the author.

References

1. Yermolenko, S. Ya. (2007). Linhvostylistyka: osnovni poniattia, napriamy y metody doslidzhennia. Ukrainska linhvostylistyka XX – pochatku XXI st.: systema poniat i bibliohrafichni dzherela. Kyiv: Hramota.
2. Lytvyn, V., Vysotska, V., Pukach, P., Nytrebych, Z., Demkiv, I., Senyk, A. et. al. (2018). Analysis of the developed quantitative method for automatic attribution of scientific and technical text content written in Ukrainian. Eastern-European Journal of Enterprise Technologies, 6 (2 (96)), 19–31. doi: https://doi.org/10.15587/1729-4061.2018.149596
3. Lytvyn, V., Vysotska, V., Pukach, P., Nytrebych, Z., Demkiv, I., Kovalchuk, R., Huzyk, N. (2018). Development of the linguometric method for automatic identification of the author of text content based on statistical analysis of language diversity coefficients. Eastern-European Journal of Enterprise Technologies, 5 (2 (95)), 16–28. doi: https://doi.org/10.15587/1729-4061.2018.142451
4. Khomytska, I., Teslyuk, V. (2016). The Method of Statistical Analysis of the Scientific, Colloquial, Belles-Lettres and Newspaper Styles on the Phonological Level. Advances in Intelligent Systems and Computing, 149–163. doi: https://doi.org/10.1007/978-3-319-45991-2_10
5. Khomytska, I., Teslyuk, V. (2017). Modelling of phonostatistical structures of the colloquial and newspaper styles in english sonorant phoneme group. 2017 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). doi: https://doi.org/10.1109/stc-csit.2017.8098738
6. Marchenko, O. O., Nykonenko, A. O., Rossada, T. V., Melnikov, E. A. (2016). Authorship attribution system. Shtuchnyi intelekt, 2, 77–85. Available at: http://dspace.nbuv.gov.ua/bitstream/handle/123456789/132051/08-Marchenko.pdf?sequence=1
7. Bhargava, M., Mehndiratta, P., Asawa, K. (2013). Stylometric Analysis for Authorship Attribution on Twitter. Lecture Notes in Computer Science, 37–47. doi: https://doi.org/10.1007/978-3-319-03689-2_3
8. Calix, K., Connors, M., Levy, D., Manzar, H., MCabe, G., Westcott, S. (2008). Stylometry for e-mail author identification and authentication. Proceedings of CSIS Research Day. Pace University.
9. Ebrahimpour, M., Putniņš, T. J., Berryman, M. J., Allison, A., Ng, B. W.-H., Abbott, D. (2013). Automated Authorship Attribution Using Advanced Signal Classification Techniques. PLoS ONE, 8 (2), e54998. doi: https://doi.org/10.1371/journal.pone.0054998
10. Chakraborty, T. (2012). Authorship identification in bengali literature: a comparative analysis. Available at: https://arxiv.org/pdf/1208.6268.pdf
11. Kotsovsky, V., Geche, F., Batyuk, A. (2015). Artificial complex neurons with half-plane-like and angle-like activation function. 2015 Xth International Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT). doi: https://doi.org/10.1109/stc-csit.2015.7325430
12. Kotsovsky, V., Geche, F., Batyuk, A. (2019). On the Computational Complexity of Learning Bithreshold Neural Units and Networks. Lecture Notes in Computational Intelligence and Decision Making, 189–202. doi: https://doi.org/10.1007/978-3-030-26474-1_14
13. Gamon, M. (2004). Linguistic correlates of style. Proceedings of the 20th International Conference on Computational Linguistics - COLING '04. doi: https://doi.org/10.3115/1220355.1220443
14. Zhao, Y., Zobel, J. (2007). Searching with style: Authorship attribution in classic literature. In Proceedings of the thirtieth Australasian conference on Computer science, 62, 59–68.
15. Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). Efficient estimation of word representations in vector space. Available at: https://arxiv.org/pdf/1301.3781.pdf
16. Cai, C., Xu, Y., Ke, D., Su, K. (2015). A Fast Learning Method for Multilayer Perceptrons in Automatic Speech Recognition Systems. Journal of Robotics, 2015, 1–7. doi: https://doi.org/10.1155/2015/797083
17. Bodyanskiy, Y., Pliss, I., Kopaliani, D., Boiko, O. (2018). Deep 2D-Neural Network and its Fast Learning. 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP). doi: https://doi.org/10.1109/dsmp.2018.8478578
18. Haykin, S. (1994). Neural networks: a comprehensive foundation. Prentice Hall PTR, 768.
19. Neural network models (supervised). Available at: https://scikit-learn.org/stable/modules/neural_networks_supervised.html
20. Backpropagation Algorithm. Available at: http://ufldl.stanford.edu/wiki/index.php/Backpropagation_Algorithm
21. Kotsovsky, V., Geche, F., Batyuk, A. (2018). Finite Generalization of the Offline Spectral Learning. 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP). doi: https://doi.org/10.1109/dsmp.2018.8478584