Machine learning approach to authorship attribution of literary texts

Urszula Stańczyk, Krzysztof A. Cyran

***Abstract*—**Machine learning approaches are employed in the variety of feature extraction and classification tasks because of their efficiency in dealing with huge amount of data. The paper addresses the application of Artificial Neural Networks as the processing engine for textual analysis of literature oriented at authorship attribution, which falls within the scope of data mining techniques. Author identification is a problem of significant importance not only from academic or historic point of view as in cases of disputed authorship of some literary works but in more current and sinister affairs of forensic nature as well. To solve the problem various methodologies have been employed or invented, belonging either to statistic-dedicated computations or machine learning algorithms, the example of the latter category presented in the paper.

***Keywords*—**Machine learning, stylometry, artificial neural networks, author identification

I. Introduction

STYLOMETRY denotes quantitative analysis of some written text that yields information about the style it is composed with and through that about the author of this text. Thus as the main stylometric tasks, belonging within information retrieval domain [1], there are considered author characterisation, similarity detection, and finally, considered as the most important, author identification.

Author characterisation brings conclusions about the author, such as gender, education, social background etc. Similarity detection involves comparing texts of several authors in order to find, if they exist, some properties in common. Author identification (or attribution) means attributing an unknown text to a writer basing on some feature characteristic or measure. It can be used when several people claim to have written some text or when no one is able or willing to identify the real author of this text.

Stylometry is most often used for detection of plagiarism, finding authors of anonymously published texts, for disputed authorship of literature or in criminal investigations within forensic linguistic domain.

Two critical issues of the stylometric analysis are: selection of descriptors that characterise texts and authors, and analytical techniques applied to the task.

The typical textual analysis procedure (invariant of particular methodology employed) starts with training during which there are used texts of known authors for whom there are computed characteristics of selected features, then follows the stage of verification when for unattributed texts there are obtained the same descriptors to be compared with previously calculated results. Then from the available set of possible authors there is chosen the one that matches most closely.

Features selected [2] in stylometric methods must constitute the writer invariant (called also authorial or author's invariant), a property of a text which is invariant of its author, that is it is similar in all texts of this author and different in texts of different authors. It is generally agreed that writer invariants exist yet establishing what properties of a text should be used is the question that stands open [3].

Usually analytical techniques applied to stylometric tasks employ either statistic or machine learning approaches. Statistical computations are used in Markovian Models (MM), Principal Component and Linear Discriminant Analysis (PCA and LDA), cluster analysis, Cumulative Sum (CUSUM or QSUM), while machine learning involves application of Artificial Neural Networks (ANN), Genetic Algorithms (GA), Support Vector Machines (SVM), Rough Set Theory (RST), decision trees, and other methods [4].

In this paper there is presented application of Artificial Neural Networks to authorship attribution considered as a classification task [5]. Texts studied are literary works of two Polish writers, B. Prus and H. Sienkiewicz, who lived and wrote around the end of the 19th and beginning of the 20th Century. Feature selected to describe texts are lexical and syntactical components that show promising results when used as writer invariants because they are used rather subconsciously and reflect the individual writing style which is difficult to be copied.

The preference of Artificial Neural Networks to other methods is explained by the simplicity of implementation made by commercially available software packages such as BrainMaker, and large set of input data for computations to be performed [6]. Properly trained neural networks possess generalisation properties that allow for high accuracy of classification that is required [7].

Manuscript received September 30, 2007; Revised Received Dec.1, 2007.

K. A. Cyran is with the Institute of Informatics, Silesian University of Technology, Gliwice, 44-100 Poland, phone: +48-32-237-2500; fax: +48-32- -237-2733; e-mail: krzysztof.cyran@ polsl.pl.

U. Stańczyk is with the Institute of Informatics, Silesian University of Technology, Gliwice, 44-100 Poland, phone: +48-32-237-2969; fax: +48-32- -237-2733; e-mail: urszula.stanczyk@ polsl.pl.

II. Objectives of Stylometry

The primary aim of stylometry is to remove uncertainty about the author of some text, which can be used in literary tasks of textual analysis for works edited, translated, with disputed authorship or anonymous, but also with forensic aspect in view to detect plagiarism, forgery of the whole document or its constituent parts, verify ransom notes, etc.

Stylometric analysts claim that each writer possesses some unique characteristic, called the authorial or writer invariant, that keeps constant for all texts written by this author and perceivably different for texts of other authors. To find writer invariants there are used style markers which are based on textual properties belonging to either of four categories: lexical, syntactic, structural, and content-specific.

Lexical descriptors provide statistics of total number of words or characters, average number of words per sentence, characters per sentence or characters per word, frequency of usage for individual letters or distribution of word length.

Syntactic features reflect the structure of sentences, which can be simple or complex, or conditional, built with punctuation marks. Structural attributes express the organisation of text into paragraphs, headings, signatures, embedded drawings or pictures, and also special font types or its formatting that go with layout.

Content-specific properties recognise some keywords: words of special meaning or significant importance for the given context.

Unfortunately, the convenience of using contemporary word editors and processors works against preserving individual author styles due to its available options of "copy and paste". It makes imitation of somebody else's style much easier and that is why modern stylometric techniques aim at exploiting the computational powers of computers to analyse patterns within subconsciously used common parts of speech, as opposed to historical approaches that emphasised some rare standing out elements of a text which could be noticed by virtually anybody and thus likely to be faked.

A. Historical View

Stylometry evolved mainly from historical textual analysis methods dedicated to proving or disproving authenticity of documents or settling questions of authorial identity for anonymous or disputed texts.

As early as in 1439 Lorenzo Valla proved the forgery of the Donation of Constantine by comparing the Latin used in other documents dated to 4th Century that were unquestionably original.

Yet these early attempts could hardly rely on anything else but striking elements of texts such as distinct vocabulary or specific language structures.

The new era for stylometry downed in 1887 when Mendenhall proposed to use not qualitative but quantitative measures such as word length, its average and distribution. This was followed by Yule and Morton, in 1938 and 1965, who selected sentence length as descriptive feature for authorship identification [8].

Numerical measurements of texts were not fully exploited at first but the development of computers with their high and permanently increasing computational powers made possible the application of statistical-oriented analysis to constantly growing corpus of texts in the cyberspace of Internet, enabling also to employ algorithms from machine learning domain to stylometric tasks.

B. Methodologies employed

Contemporary stylometric procedures are typically representatives of either computer-aided statistic based analysis, or artificial intelligence techniques.

In statistical analysis there are used computations of probabilities and distributions of occurrences for single letters or other characters such as punctuation marks, words, patterns of words or sentences [9].

One such method, called QSUM or CUSUM, was developed by Jill M. Farringdon [10]. The name of this method comes from the step of calculating the cumulative sum for two textual features. The first of these is the sentence length whose deviations from the average are plotted as the graph for the whole text sample of some known author. As the second descriptor typically there is chosen either the usage of the 2 and 3 letter words, using words starting with a vowel, or the combination of these two together. The two descriptors reflect the writing habits and are the key to detecting the author. If the two graphs match, the writer is identified.

Markovian Models consider a text as a sequence of characters (letter, punctuation marks, spaces, etc.) that corresponds to a Markov chain [11]. In probabilistic model of natural language letters appear with some probability, depending on which characters precede them. In the simplest model there is considered only the immediate predecessor which gives rise to the 1^{st} order Markov chain. Thus for all pairs of letters in the alphabet there are obtained matrices of transition frequencies of one letter into another. These statistics are calculated for all texts by known authors and for some unattributed text as the true author there is selected the one with the highest probability [12].

Methods such as Linear Discriminant Analysis, Principal Component Analysis or cluster analysis aim to reduce the dimensionality for input data and if procedures applied to texts of both known and unknown authors give the same result, the question of authorship identification is settled.

Genetic Algorithms provide an example of artificial intelligence technique [13] applied in stylometric analysis. The whole procedure starts with definition of a set of rules describing textual properties. Next these rules are checked against the text of known authorship and each rule if evaluated for fitness, basing on which score some rules (with the lowest score) are discarded leaving only these with fitness satisfying some criterion (selection process). The selected rules are slightly modified (mutation) and some new added, after which they are tested again. The process continues till there is obtained some number of rules that best describe features of the known text. At this point the evolved rules can

be tested on a text of unknown author and if their fitness remains the same, the author is found.

Artificial Neural Networks are well suited to classification tasks by their ability to deal efficiently with large amount of data, especially in continuous domain since they do not require discretisation as for example classical rough sets. As the processing engine applied to research this paper presents, ANN with their topologies and training methods are described in the next section with more detail.

III. Artificial Neural Networks

Nervous systems existing in biological organism for years have been the subject of studies for mathematicians who tried to develop some models describing such systems and all their complexities. Artificial Neural Networks emerged as generalisations of these concepts with mathematical model of artificial neuron due to McCuloch and Pitts described in 1943, definition of unsupervised learning rule by Hebb in 1949, and the first ever implementation of Rosenblatt's perceptron in 1958.

The efficiency and applicability of ANN to computational tasks have been questioned many times, especially at the very beginning of their history the book "Perceptrons" by Minsky and Papert, published in 1969, caused dissipation of initial interest and enthusiasm in applications of neural networks. It was not until 1970s and 80s, when the backpropagation algorithm for supervised learning was documented that ANN regained their status and proved beyond doubt to be sufficiently good approach to many problems.

Artificial Neural Network can be looked upon as a parallel computing system comprised of some number of rather simple processing units (neurons) and their interconnections. They follow inherent organizational principles such as the ability to learn and adapt, generalisation, distributed knowledge representation, and fault tolerance.

Neural network specification comprises definitions of the set of neurons (not only their number but also their organisation), activation states for all neurons expressed by their activation functions and offsets specifying when they fire, connections between neurons which by their weights determine the effect the output signal of a neuron has on other neurons it is connected with, and a method for gathering information by the network that is its learning (or training) rule [14].

A. Topology

From topology point of view neural networks can be divided into two categories: feed-forward and recurrent networks. In feed-forward networks the flow of data is strictly from input to output cells that can be grouped into layers but no feedback interconnections can exist. On the other hand, recurrent networks contain feedback loops and their dynamical properties are very important.

The most popularly used type of neural networks employed in pattern classification tasks is the feedforward network which is constructed from layers and possesses unidirectional weighted connections between neurons. The common examples of this category are Multilayer Perceptron or Radial Basis Function networks, out of which the former will be addressed in more detail.

Multilayer Perceptron (MLP) type is more closely defined by establishing the number of neurons from which it is built, and this process can be divided into three parts, the two of which, finding the number of input and output units, are quite simple, whereas the third, specification of the number of hidden neurons can become crucial to accuracy of obtained classification results.

The number of input and output neurons can be actually seen as external specification of the network and these parameters are rather found in a task specification. For classification purposes as many distinct features are defined for objects which are analysed that many input nodes are required. The only way to better adapt the network to the problem is in consideration of chosen data types for each of selected features. For example instead of using the absolute value of some feature for each sample it can be more advantageous to calculate its change as this relative value should be smaller than the whole range of possible values and thus variations could be more easily picked up by Artificial Neural Network. The number of network outputs typically reflects the number of classification classes.

The third factor in specification of the Multilayer Perceptron is the number of hidden neurons and layers and it is essential to classification ability and accuracy. With no hidden layer the network is able to properly solve only linearly separable problems with the output neuron dividing the input space by a hyperplane. Since not many problems to be solved are within this category, usually some hidden layer is necessary.

With a single hidden layer the network can classify objects in the input space that are sometimes and not quite formally referred to as simplexes (single convex objects that can be created by partitioning out from the space by some number of hyperplanes) whereas with two hidden layers the network can classify any objects since they can always be represented as a sum or difference of some such simplexes classified by the second hidden layer.

Apart from the number of layers there is another issue of the number of neurons in these layers. When the number of neurons is unnecessarily high the network easily learns but poorly generalises on new data. This situation reminds auto-associative property: too many neurons keep too much information about training set rather "remembering" than "learning" its characteristics. This is not enough to ensure good generalization that is needed.

On the other hand, when there are too few hidden neurons the network may never learn the relationships amongst the input data. Since there is no precise indicator how many neurons should be used in the construction of a network, it is a common proctice to built a network with some initial number of units and when it trains poorly this number is either increased or decreased as required. Obtained solutions are

usually task-dependant.

B. Activation Functions

Activation or transfer function of a neuron is a rule that defines how it reacts to data accumulated through its inputs that all have certain weights.

Typically there is used linear or semi-linear function, a hard limiting threshold function or a smoothly limiting threshold such as a sigmoid or a hyperbolic tangent. Due to their inherent properties (out of which the most important are whether they are linear, continuous or differentiable) different activation functions perform with different efficiency in task-specific solutions.

For classification tasks sigmoid is the most popularly used activation function (non-linear, continuous, differentiable) defined as

$$y(n) = \frac{1}{1+e^{-\beta n}} \tag{1}$$

where n (net) is a scalar product

$$n = \mathbf{W} \cdot \mathbf{X} = \mathbf{W}^T \mathbf{X} = \sum_{j=0}^{J} w_j x_j \tag{2}$$

of the weight $\mathbf{W}$ and input vectors $\mathbf{X}$, with $j = 0$ reserved for offset t, by setting $x_0 = 1$and $w_0 = -$ t

C. Learning Rules

In order to produce the desired set of output states whenever a set of inputs is presented to a neural network it has to be configured by setting the strengths of the interconnections and this step corresponds to the network learning procedure. Learning rules are roughly divided into three categories of supervised, unsupervised and reinforcement learning methods.

The term *supervised* indicates an external teacher who provides information about the desired answer for each input sample. Thus in case of supervised learning the training data is specified in forms of pairs of input values and expected outputs. By comparing the expected outcomes with the ones actually obtained from the network the error function is calculated and its minimisation leads to modification of connection weights in such a way as to obtain the output values closest to expected for each training sample and to the whole training set.

In unsupervised learning no answer is specified as expected of the neural network and it is left somewhat to itself to discover such self-organisation which yields the same values at an output neuron for new samples as there are for the nearest sample of the training set.

Reinforcement learning [15] relies on constant interaction

between the network and its environment. The network has no indication what is expected of it but it can induce it by discovering which actions bring the highest reward even if this reward is not immediate but delayed. Basing on these rewards it performs such re-organisation that is most advantageous in the long run.

In commonly used Mulilayer Perceptron networks typically there are applied some variants of supervised backpropagation method. The classical backpropagation algorithm modifies the vector of all weights $\mathbf{W}$ accordingly to the descent direction of the gradient

$$\Delta \mathbf{W} = -\eta \nabla e(\mathbf{W}) \tag{3}$$

(with η being the learning rate) of the error occurring on the output of the network

$$e(\mathbf{W}) = \frac{1}{2} \sum_{m=1}^{M} \sum_{i=1}^{I} \left(d_i^m - y_i^m(\mathbf{W}) \right)^2 \tag{4}$$

that is a sum of errors for all M training facts on all output neurons, each defined by the difference between the expected outcome d_i^m and the one generated by the network $y_i^m(\mathbf{W})$.

The modification of weights associated with network interconnections can be performed either after each of the training samples or after finished iteration of the whole training set.

The important factor in this algorithm is the learning rate η whose value when too high can cause oscillations around the local minima of the error function and when too low results in slow convergence. This locality is considered the drawback of the backpropagation method but its universality is the advantage.

IV. Experiments

Stylometric analysis that was performed within research presented in this paper can be seen as the multistage process, as follows

- the first step was selection of the training and testing examples - texts to be studied,
- next stage was taken by the choice of textual descriptors to be analysed - the writerprints of the authors of previously selected texts,
- then followed the third phase of calculating characteristics for all descriptors that were later used for training of the neural network,
- specification of the network with its topology and learning method can be seen as the fourth step of the whole procedure,
- the fifth consisted of the actual training of the network,
- the sixth stage was dedicated to testing,

- and the final one corresponded to analysis of obtained results and coming up with some conclusions and possible indicators for improvement.

Described process was applied several times to different input data, with various topologies of neural networks giving total classification results with accuracy that varied from just few to 100%. As it is not possible to include all detailed results, only some sumarising selection is provided, with calculated averages of outcomes for individual testing samples and works.

A. Texts Used

In research there were used texts of two famous Polish writers, Henryk Sienkiewicz and Bolesław Prus. Their novels and short works provide the corpora which is wide enough to make sure that characteristic features found basing on the training data can be treated as representative of other texts and this generalized knowledge can be used to confirm or discount the possibility of either of considered writers being recognised as the author of a text of unknown origin.

Obviously literary texts can greatly vary in length, what is more, all stylistic features can be influenced not only by different timelines within which the text is written but also by its genre. The first of these issues is easily dealt with by dividing long texts, such as novels, into some number of smaller parts of approximately the same size.

Described approach gives additional advantage in classification tasks as even in case of some incorrect classification results of these parts the whole text can still be properly attributed to some author by basing the final decision on the majority of outcomes instead of all individual decisions for all samples.

Whether the genre of a novel is reflected in lexical and syntactic characteristics of it is the question yet to be answered. If the influence is significant, then lexical and syntactic features cannot be used as the writer invariant as unreliable. On the other hand, this can be rectified by including within the training data set fragments of texts being representatives of not only one but several genres. In fact the more the better. For intended implementation of the classifier with Artificial Neural Networks, which efficiently deal with large amount of data, adding samples to the training set simply means better coverage of the input space that is important in continuous case.

Hence in the training set there were included samples coming from two volumes of "Faraon" and the first volume of "Lalka" by Prus, and two volumes of "Potop" and the first volume of "Krzyżacy" by Sienkiewicz, 70 for the former case and 98 for the latter, giving the total of 168. The testing set comprised all other available works by the two writers, but detailed classification results are provided only for the third volume of "Faraon", the second volume of "Lalka", "Emancypantki" and "Placówka" by Prus, and the third volume of "Potop", the second volume of "Krzyżacy", "Quo vadis" and "Rodzina Połanieckich" by Sinkiewicz, while for all other works that include both other novels and some number of short stories there are given just counted averages.

B. Feature Selection

Establishing features that work as effective discriminators of texts under study is one of critical issues in research on authorship analysis.

In the research there where three groups of textual descriptors used, lexical and syntactic, the first of which was the usage of function words, the second application of punctuation marks, and the third the combination of the other two as follows

- Set 1 of descriptors contained nine function words: "ale", "i", "nie", "to", "w", "z", "że", "za", "na",
- Set 2 of descriptors comprised eight punctuation marks: a comma, a full stop, a semi-colon, a bracket, an exclamation mark, a quotation mark, a colon, and a question mark (Set 2),
- Set 3 of descriptors consisted of both function words and punctuation marks including: "i", "nie", "to", "bo", "po", a comma, a full stop and an exclamation mark.

C. Architecture of ANN

As the base topology of Artificial Neural Network there was applied the feed-forward Multilayer Perceptron with sigmoid activation function trained by backpropagation algorithm. The number of inputs equaled the number of textual descriptors used, thus it was either nine of eight.

There was either one or two hidden layers with as few neurons within them as possible for preserving generalisation properties but achieving convergence during training with tolerance at most 0.4 for all training samples recognised properly.

For all structures of ANN there were used two outputs. Actually, it was possible to use a single output and by interpretation of its active state as one class and inactive output state the second class the task would have been solved as well, but with such approach a text would always be attributed to either one or another author and classification binary with undecided verdict impossible. Two outputs allow to recognise the situation that the network could not easily recognise the writing style of any of previously trained authors and is unable to properly classify some sample of text.

V. Results and Discussion

As stated previously, for validation purposes there were used samples from all other works of both writers that were available, consisting of other parts of the same novels used previously during training and from different novels and short works. Due to their volume texts were divided in such a way

that usually a chapter constituted a sample but maintaining approximately the same lengths for samples. For succinct presentation the results given are as averages for whole volumes or novels and other works instead of details for all individual samples.

Set 1 of lexical descriptors defined by function words was tested on the neural network having the architecture composed of one hidden layer which contained 13 neurons. The results of classification performed by this network are given in the Table 1.

Table 1 Classification results for lexical descriptors, ANN with one hidden layer.

Author	Text	Classification 1
Prus	"Faraon" vol. 3	94.4%
	"Lalka" vol. 2	63.2%
	"Emancypantki"	89.9%
	"Placówka"	9%
	other works	55.5%
	total	75%
Sienkiewicz	"Potop" vol. 3	80%
	"Krzyżacy" vol. 2	80%
	"Quo vadis"	52.6%
	"Rodzina Połanieckich"	90.5%
	other works	69.2%
	total	70.5%

Then the same set of descriptors was tested on the network with two hidden layers, with twelve and seven neurons in them respectively. The results of classification performed by this ANN are provided in the Table 2.

Table 2 Classification results for lexical descriptors, ANN with two hidden layers.

Author	Text	Classification 2
Prus	"Faraon" vol. 3	94.4%
	"Lalka" vol. 2	68.4%
	"Emancypantki"	90.9%
	"Placówka"	18.2%
	other works	60%
	total	77.6%
Sienkiewicz	"Potop" vol. 3	66.6%
	"Krzyżacy" vol. 2	88%
	"Quo vadis"	58.6%
	"Rodzina Połanieckich"	95.2%
	other works	71.1%
	total	73%

Clearly total results specified both tables do not vary in significant degree, as one network classifies better one novel while another network obtains higher accuracy for another novel. Yet the second network gives slightly better overall classification accuracy.

Set 2 of syntactic descriptors comprising punctuation marks was first tested on the network with one hidden layer comprising 20 neurons as specified by the Table 3.

Table 3 Classification results for syntactic descriptors, ANN with one hidden layer and 20 neurons.

Author	Text	Classification 3
Prus	"Faraon" vol. 3	100%
	"Lalka" vol. 2	100%
	"Emancypantki"	99%
	"Placówka"	81.8%
	other works	57.7%
	total	88.5%
Sienkiewicz	"Potop" vol. 3	100%
	"Krzyżacy" vol. 2	100%
	"Quo vadis"	97.1%
	"Rodzina Połanieckich"	100%
	other works	93.5%
	total	95.8%

Next Set 2 was tested on the network still with just one hidden layer but with only 5 neurons (results presented by the Table 4).

Accuracy of classification 3 and 4 given by the Tables 3 and 4 confirms the fact which was indicated previously with discussion of neural network properties, that fewer hidden neurons can result in better generalization ability than their high number.-

Furthermore, for this set of syntactic descriptors the classification accuracy is significantly higher than in the previous case of lexical features.

Table 4 Classification results for syntactic descriptors, ANN with one hidden layer and 5 neurons.

Author	Text	Classification 4
Prus	"Faraon" vol. 3	100%
	"Lalka" vol. 2	100%
	"Emancypantki"	100%
	"Placówka"	81.8%
	other works	60%
	total	89.6%
Sienkiewicz	"Potop" vol. 3	100%
	"Krzyżacy" vol. 2	98%
	"Quo vadis"	91.4%
	"Rodzina Połanieckich"	95.2%
	other works	93.1%
	total	94%

After checking the performance of lexical and syntactic descriptors separately, the Set 3 of mixed lexical and syntactic descriptors was tested on the neural network with one hidden layer composed of 4 neurons and the results are provided by the Table 5.

Table 5 Classification results for neural networks using mixed

descriptors.

Author	Text	Classification 5
Prus	"Faraon" vol. 3	100%
	"Lalka" vol. 2	89.5%
	"Emancypantki"	100%
	"Placówka"	27.3%
	other works	64.4%
	total	86.5%
Sienkiewicz	"Potop" vol. 3	100%
	"Krzyżacy" vol. 2	100%
	"Quo vadis"	97.1%
	"Rodzina Połanieckich"	100%
	other works	93.5%
	total	95.8%

By comparing the last table to the previous four it is quite obvious that the highest classification ratio is granted by the exploitation of syntactic textual features, while mixed ones are nearly as good, and purely lexical the worst.

When results for specified novels are studied it is clear that in cases when the last volume was used for tests while the previous one or ones worked as training, the classification ratio is higher which comes as no surprise. In the same novel the writing style is best maintained throughout volumes thus once learned by the network it is more easily recognised than in other samples.

On the other hand, there are some works which are not properly recognised, no matter which set of textual descriptors and which structure of a neural network is used and their classification results lower the overall ratio. Just from stylometry point of view the question why it happens and how to enhance results is not easily answered, at least in the latter part.

Some works are not correctly classified because textual features describing them are not precise enough for the task. The writing style can be so specific and distinctive that requires to use less typical descriptors, for example different function words.

Increasing the classification ratio is best addressed from the point of technique that was employed to the task - neural networks in the presented research. Thus the question can be posed as how to enhance generalisation property of ANN. Actually, it may happen that some data is not easily learned by neural network and when changing the structure does not help, the next approach to be checked is using for training of the network these samples that were the worst in recognition. Then the network learns "difficult" data and is tested on less difficult which results in higher classification accuracy.

VI. Conclusions

The research described in this paper concerning stylometric analysis shows beyond doubt how efficient a tool Artificial Neural Networks can be when applied in classification tasks. Yet conclusions as to the choice of textual descriptors used as features for recognition process, based only on results presented in the previous section and leading to some arbitrary statement that syntactic attributes are more effective in authorship attribution, would be much too hasty and premature. Undeniably true in the studied example, it would have to be verified against much wider corpora as for other writers other features could give better results.

Thus a series of future experiments should include application of the presented here ANN-based methodology to wider range of authors, definition of new sets of textual descriptors, and test for other types and structures of neural networks.

Acknowledgment

The experiments, the results of which are presented in this article, were performed in Silesian University of Technology, Gliwice, Poland, under supervision of K.A. Cyran by S. Cichoń in fulfillment of requirements for MSc degree [16].

References

[1] G. Bonanno, F. Moschella, S. Rinaudo, P. Pantano, and V. Talarico, "Manual and evolutionary equalization in text mining", *Proceedings of the 7th WSEAS International Conference on Simulation, Modelling and Optimization*, pp. 262–267, 2007.

[2] S. Doan, S. Horiguchi, "An efficient feature selection using multi-criteria in text categorization for naive Bayes classifier", *WSEAS Transactions on Information Science & Applications*, vol. 2, no. 2, pp. 98–103, 2005

[3] S. Argamon, J. Karlgren, and J.G. Shanahan, eds., "Stylistic analysis of text for information access", *Proceedings of the 28th International ACM Conference on Research and Development in Information Retrieval, Brasil*, 2005.

[4] P. Makvandi, J. Jassbi, and S. Khanmohammadi, "Application of genetic algorithm and neural network in forecasting with good data", *WSEAS Transactions on Systems*, vol. 4, no. 4, pp. 337–342, 2005.

[5] W. Kreesuradej, W. Kruakli, and N. Chantasut, "Clustering text data using text art neural network", *WSEAS Transactions on Systems*, vol. 3, no. 1, pp. 200–205, 2004.

[6] W.W. Aprasit, N. Laosen, and S. Chevakidagarn, "Data filtering technique for Neural Networks forecasting", *Proceedings of the 7th WSEAS International Conference on Simulation, Modelling and Optimization*, pp. 225–230, 2007.

[7] R.A.J. Matthews and T.V.N. "Merriam, Distinguishing literary styles using neural networks", in E. Fiesler and R. Beale, eds., *Handbook of neural computation*, OUP, pp. G8.1.1–6, 1997.

[8] R.D. Peng and H. Hengartner, "Quantitative analysis of literary styles, The American Statistician", vol. 56, no. 3, pp. 15–38. 2007.

[9] R. Peng, "Statistical Aspects of Literary Style", *Bachelor's Thesis, Yale University*, 1999.

[10] W. Buckland, "Forensic semiotics", *The Semiotic Review of Books*, vol. 10, no. 3, 1999.

[11] D.V. Khmelev and F.J. Tweedie, "Using Markov chains for identification of writers", *Literary and Linguistic Computing*, vol. 16, no. 4, pp. 299–307, 2001.

[12] D.T. Tran and T.D. Pham, "Markov and fuzzy models for written language verification", *WSEAS Transactions on Systems*, vol. 4, no. 4, pp. 268–272, 2005.

[13] J. Stastny and V. Skorpil, "Genetic algorithm and neural network", *Proceedings of the 7th WSEAS International Conference on Applied Informatics and Communications*, pp. 345–350, 2007.

[14] S. Ossowski, „Sieci neuronowe do przetwarzania informacji", *Wydawnictwa Naukowo–Techniczne*, Warsaw, Poland, 2000.

[15] M. Preda, "Adaptive building of decision trees by einforcement learning", *Proceedings of the 7th WSEAS International Conference on Applied Informatics and Communications*, pp. 34–39, 2007.

[16] S. Cichoń, „Sieci neuronowe w zagadnieniach stylometrycznego ustalania autorstwa tekstów literackich”, *MSc Thesis, Silesian University of Technology, Gliwice, Poland*, 2003.

Stanczyk Urszula received her MSc and PhD degrees in computer science from the Silesian University of Technology, Gliwice, Poland in 1993 and 2003 respectively.

From 1993 till 2000 she was a teaching assistant, from 2000 till 2003 a lecturer, and from 2004 till present an assistant professor at the Institute of Informatics, SUT. From 2004 she has been the Editor-in- Chief of the Activity Report for the Institute of Informatics. Her scientific research interest include digital image processing and recognition, with special emphasis on mathematical morphology methods, computational intelligence and especially rough set theory and artificial neural networks, stylometry and its tasks, elements of theory of logic circuits, their design procedures and optimisation of implementations, as well as arithmetic of digital systems.

Krzysztof A. Cyran was born in Cracow, Poland, in 1968. He received MSc degree in computer science (1992) and PhD degree (with honours) in technical sciences with specialty in computer science (2000) from the Silesian University of Technology SUT, Gliwice, Poland. His PhD dissertation addresses the problem of image recognition with the use of computer generated holograms applied as ring-wedge detectors.

He has been an author and co-author of more than 60 technical papers in journals (several of them indexed by Thomson Scientific) and conference proceedings. Dr. Cyran (in 2003-2004) was a Visiting Scholar in Department of Statistics at Rice University in Houston, US. He is currently the Assistant Professor and the Vice-Head of the Institute of Informatics at Silesian University of Technology, Gliwice, Poland. He is also a member of the Editorial Board of Journal of Biological Systems and a member of the Scientific Program Committees of many international conferences. His current research interests are in image recognition and processing, artificial intelligence, digital circuits, decision support systems, rough sets, computational population genetics and bioinformatics.