SOUTHEAST EUROPE JOURNAL OF SOFT COMPUTING
Available online at www.scjournal.com.ba

Authorship Attribution Using Principal Component Analysis and Nearest Neighbor Rule for Neural Networks

Mehmet Can[a]

[a] International University of Sarajevo, Faculty of Engineering and Natural Sciences
Hrasnićka Cesta 15, 71000 Sarajevo, Bosnia and Herzegovina
mcan@ius.edu.ba

***Abstract*— Feature extraction is a common problem in statistical pattern recognition. It refers to a process whereby a data space is transformed into a feature space that, in theory, has exactly the same dimension as the original data space. However, the transformation is designed in such a way that the data set may be represented by a reduced number of "effective" features and yet retain most of the intrinsic information content of the data; in other words, the data set undergoes a dimensionality reduction. Principal component analysis is one of these processes. In this paper the data collected by counting selected syntactic characteristics in around a thousand paragraphs of each of the sample books underwent a principal component analysis. To make a comparison, the original data is also processed. Authors of texts identified with higher success by the competitive neural networks, which use principal components. The process repeated on another group of authors, and similar results are obtained.**

Keywords— principal components, authorship attribution, stylometry, text categorization, stylistic features, syntactic characteristics, multilayer preceptor, competitive learning, artificial neural network.

INTRODUCTION

Problems of authorship have always been attacked with traditional research methods: unearthing and dating original manuscripts, for instance. But since the late 19th century, statisticians have developed "non-traditional" tools that attempt to discern quantifiable patterns within a text or corpus, with the hope that these features will help to reliably identify different authors.

The origin of non-traditional authorship attribution, or stylometry, is often said to be Augustus de Morgan's suggestion in 1851 that certain authors of the Bible might be distinguishable from one another if one used longer words (Holmes 1998). In 1887, searching for a characteristic difference in the distribution of different-sized words in writings of different languages and presentation styles, Mendenhall began investigating this hypothesis,. In 1901, he turned his methods to Shakespeare, Bacon and Marlowe, and found that while Shakespeare and Marlowe were nearly indistinguishable, they were both significantly and consistently different from Bacon (Williams 1975). The difference was mainly observed in the relative frequency of three- and four-letter words: Shakespeare used more four - letter words and Bacon more three-letter words.

Authorship studies also began independently around the same time in Russia with Morozov (Kukuskina et. al. 2002). In the West, it took 30 years or so for Mendenhall's studies to be resumed by other linguists. G. Zipf examined word frequencies and determined not a stylometric but a universal law of language, Zipf's Law: that the statistical rank of a word varies inversely to its frequency (Smith 2008). G. U. Yule devised a feature known as "Yule's characteristic K," which estimated 'vocabulary richness' by comparing word frequencies to that expected by a Poisson distribution, but like Mendenhall's word lengths, this too was later found to be an unreliable marker of style (Holmes 1998). In fact, most of the measurements proposed in this period proved unhelpful: among others, researchers tried average sentence length, number of syllables per word, and other estimates of vocabulary richness such as Simpson's D index and a simple type/token ratio, a ratio of the number of unique words, or types, to the number of total words, or tokens (Juola 2006).

The needed breakthrough came at last in 1963 with Mosteller and Wallace's study on the Federalist Papers. In 1787 and 1788, J. Jay, A. Hamilton and J. Madison collectively wrote 85 newspaper essays supporting the ratification of the constitution. Published under the pseudonym "Publius," the authors later revealed which of the Federalist Papers they had written; however, while authorship of 67 were undisputed, 12 were claimed by both Hamilton and Madison. Mosteller and Wallace hoped to characterize each author's style through their choice of function words, such as "to," "by," and so forth. Function words are regarded as good markers of style because they are assumed to be unconsciously generated and independent of semantics, the meaning, or what the author is trying to convey. That is, an author may have a preference for modes of expression, for instance, the active vs. the passive voice that emphasize certain function words, and the same broad set of function words will be used regardless of the topic at hand (Smith 2008).

Despite the fact that Hamilton and Madison have otherwise very similar styles, nearly identical sentence length distributions, as noted by Juola (Juola 2006), Mosteller and Wallace found sharp differences in their preference for different function words: for instance, the word "upon" appears 3.24 times per 1000 words in Hamilton, and just 0.23 times in Madison (Holmes 1998). Adjusting these frequencies with a Bayesian model, they showed that Madison had most likely written all 12 disputed papers. Traditional scholarship had already long come to the same conclusion, but Mosteller and Wallace's conclusion was independent, and thus a great achievement of the then quite exploratory field of stylometry. The Federalist Papers problem is still regarded as a very difficult test case, and as an unofficial benchmark it has been used to test most methods of authorship attribution developed since then (Kjell 1994, Holmes 1995, Bosh and Smith 1998, Fung 2003).

The most probable attribution can be viewed as taking paragraphs of documents as points in some space, and assigning a questioned document to the author whose paragraphs are 'closest' to it, according to an appropriate distance measure. Such distance measures continue to be used in recent studies examining the efficacy of different metrics and feature sets. A related class of techniques was developed earlier by Burrows (Burrows 1987, and 1988), who applied principal components analysis (PCA) on word frequencies to identify authorship. This method was elaborated on by Binongo and Smith (Binongo and Smith 1999), and has been used to resolve several outstanding authorship problems (Burrows 1992, Binongo 2003).

A related method is ANOVA, as applied, for example, by Holmes and Forsyth (Holmes and Forsyth 1995) to the Federalist. From the probabilistic standpoint, these methods take into account, to some extent, the statistical dependence of different words' frequencies.

Methods that model the sequencing of words in a document takes into account another form of dependence between words. This is done by using a probabilistic distance measure such as K-L divergence between Markov model probability distributions of the texts (Holmes 2003, Juola 1998, Khmelev 2001, 2002, Juola, and Baayen 2003, Sanderson, and Guanter 2006).

An important turning point in authorship attribution studies started by the emergence of text categorization techniques rooted in machine learning. The application of such methods is straightforward: training texts are represented as labeled numerical vectors and learning methods are used to find boundaries between classes that minimize some classification loss function. The nature of the learned boundaries depends on the learning method used but in any case these methods facilitate the use of classes of boundaries

that extend well beyond those implicit in methods that minimize distance.

Earliest methods were using small sets of function word as features (Matthews, and Merriam 1993, Merriam, and Matthews 1994, Kjell 1994, Lowe, and Matthews 1995). Recently, neural networks are used on a wide variety of features (Graham et al 2005, Zheng et. al 2006, Can et al 2011, 2012, Can, Hadziabdic, and Demir 2011). Some researchers used the techniques of k-nearest neighbor (Kjell et al 1995, Hoom et al 1999, Zhao, and Zobel 2005), naïve Bayes (Kjell 1994, Hoom et al 1999, Peng et al 2004), rule learners (Binongo 2003, Holmes 2003, Graham et al 2005, Argamon-Engelson et. al 1998, Koppel, and Schler 2003, Abbasi, and Chen 2005), support vector machines (Zheng et al 2008, Koppel, and Schler 2003, Abbasi, and Chen 2005 De Vel et al 2001, Diederich et al 2003, Koppel et al 2005), Winnow (Koppel et al 2002, Argamon-Engelson et. al 2003, Koppel, et al 2006), and Bayesian regression (Genkin et al 2006, Madigan et al 2006, Argamon-Engelson et. al 2009).

Comparative studies on machine learning methods for topic-based text categorization problems (Dumais et al 1998, Yang 1999) have shown that in general, support vector machine (SVM) learning is at least as good for text categorization as any other learning method. Although in (Diederich et al 2003) it is seen that SVM is able to rejects other authors and detects the target author in 60-80% of the cases, in general, it is as good as other learning methods also for authorship attribution (Zheng et al 2006, Abbasi, and Chen 2005). Some recent studies (Genkin et al 2006, Koppel, and Schler 2003) have shown that some variations of Winnow and Bayesian regression are also very promising.

Below, we compare the performance of principal components analysis, support vector machine (SVM) learning for authorship attribution.

PROBLEM DEFINITION

In this paper author attribution is considered as an application of principal component analysis, and as a classification task (Chaski 2001, 2005). Texts studied are literary works of five South East European writers, and Aleksandr Solzhenitsyn :

Ivo Andrić (1892-1975)

1. Na Drini ćuprija (Svjetlost, Sarajevo, 1945.)
2. Prokleta avlija (pripovijest, 1954.)
3. Znakovi pored puta (u okviru Sabranih djela, 1976.)

Meša Selimović (1910-1982)

1. Derviš i smrt, Svjetlost. Sarajevo, 1966; 1967;

Derviš Sušić (1925 - 1990)

1 Pobune. Veselin Masleša. Sarajevo, 1966.

Ante Kovačić (1854 - 1889)

1 U Registraturi, Mladost, 1968

2. Baruničina ljubav ; Fiškal ; Među žabari, Dom i svijet, 2004.

Aleksandar Solzenytcin (1918 - 2008)

1. Odjel za rak (Cancer Ward, 1968)
2. U prvom krugu (The First Circle, 1968; novel)

Ranko Marinković (1913 -2001)

1. Ruke, Svjetlost, Sarajevo 1964

Features 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. Principal components of data elicited from texts possess generalization properties that allow for the required high accuracy of classification (Hayes 2008).

The novels selected provide the corpora which are wide enough to make sure that characteristic features found based on the training data can be treated as representative of other parts of the texts and this generalized knowledge can be used to classify the test data according to their respective authors.

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 based 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.

Feature Selection

Establishing features that work as effective discriminators of texts under study is one of critical issues in research on authorship analysis which are lexical. In this research fourteen textual descriptors are used, average sentence length, average word length, number of words, sentences, commas, and conjecture “and”, in Bosnian “i”, and other characteristics in paragraphs listed in the first column of Table 1. Means and variances of the textual descriptors for the texts Ivo Andrić: Na Drini Ćuprija, M. Meša Selimović: Derviš i Smrt, and Derviš Sušić: Pobune are shown in Table 1 as samples for comparison.

Table 1. Paragraph averages and variances of the textual descriptors used in this research

	Na Drini Ćuprija	
Textual descrs.	Mean	Variance
Sentence length	84.331	2090.92
Word length	2.157	2.877
Word count	79.208	5861.724
Sentence count	4.395	16.886
Comma count	6.432	45.95
dots count	0.052	0.135
i count	5.375	35.072
ili count	0.250	0.514
je count	2.798	11.991
se count	1.852	4.823
pa count	0.140	0.216
da count	1.935	6.853
ne count	0.637	1.695
kao poput count	0.662	1.106
Total		8080.760

	Derviš i Smrt	
Textual descrs.	Mean	Variance
Sentence length	58.710	2053.855
Word length	2.155	3.460
Word count	60.362	4756.432
Sentence count	5.012	29.411
Comma count	7.130	87.211
dots count	0.002	0.002
i count	2.235	9.659
ili count	0.302	0.688
je count	2.552	11.531
se count	1.615	4.478
pa count	0.098	0.133
da count	2.262	9.613
ne count	0.968	2.718
kao poput count	0.480	1.007
Total		6970.200

	Pobune	
Textual descrs.	Mean	Variance
Sentence length	33.0478	1337.3416
Word length	2.5459	3.0985
Word count	24.5825	1040.4906
Sentence count	3.4843	17.0118
Comma count	2.6660	16.4196
dots count	0.2526	0.6327
i count	0.6910	1.8709
ili count	0.09390	0.1397
je count	0.6305	1.8402
se count	0.6221	1.2021
pa count	0.0731	0.0846
da count	0.8601	2.334
ne count	0.4196	0.6708
kao poput count	0.0793	0.1192
Total		2423.2562

As it is seen, there is statistical differences between the usages of textual descriptors in texts, for instance, Ivo Andrić prefers longer paragraphs. In average Ivo Andrić ‘s paragraphs contain 79 words with variance 5861.7, while Meša Selimović’s average is 62 with variance 4756.4, and Derviš Sušić’s average is 25 with variance 1040.5.

In the next chapter the pattern captured by principal components corresponding to these data will be displayed.

PRINCIPAL COMPONENT ANALYSIS

Burrows method of principle component analysis (PCA) (Burrows 1987) essentially involves computing the frequency of each of a list of function words, and performing principle component analysis (PCA) to find the linear combination of variables that best accounts for the variations in the data. Rather than analyze this result statistically, the transformed data are simply plotted. Two-dimensional plots of the first two principal components supply us with a means to inspect visually for trends, which occur as clusters of points (Binongo 2003). Later, cluster analysis may follow this step.

This simple but effective method continues to be used today, partly because of the ease with which the results are communicated and interpreted. For example, Binongo (Binongo 2003). used this method to study the problem of the authorship of L. Frank Baum’s last book, which historians had long suspected of being mostly the work of Baum’s successor, Ruth P. Thompson. He confirmed this suspicion independently, demonstrating that Thompson was much more prone to use position words such as “up,” “down,” “over,” and “back,” than Baum. This was not demonstrated using complex statistical techniques; rather, function word frequencies were tallied, the authors’ tallies compared, PCA used to reduce the dimensionality of the data, and the resulting plots inspected: the two authors’ works form obvious clusters. Similar procedures can be found in (Bosh and Smith 1998, Holmes et al 2001, Peng, and Hengartner 2002).

In this paper instead of cluster analysis of the two dimensional plots, the author attribution will be found by the use of artificial neural networks with output neurons competing on the data of first principal components.

Theory of Principal component Analysis

Multivariate statistics deals with the relation between several random variables. The sets of observations of the random variables are represented by a multivariate data matrix $\boldsymbol{X}$,

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ x_{31} & x_{32} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}. \qquad (1)$$

Each column vector $\boldsymbol{u}_k$ represents the data for a different variable. If $\boldsymbol{c}$ is an $p \times 1$ matrix, then

$$\boldsymbol{Xc} = c_1\begin{bmatrix} x_{11} \\ x_{21} \\ x_{31} \\ \vdots \\ x_{n1} \end{bmatrix} + c_2\begin{bmatrix} x_{12} \\ x_{22} \\ x_{32} \\ \vdots \\ x_{n2} \end{bmatrix} + \cdots + c_p\begin{bmatrix} x_{1p} \\ x_{2p} \\ x_{3p} \\ \vdots \\ x_{np} \end{bmatrix} \qquad (2)$$

is a linear combinations of the set of observations.

Descriptive statistics can also be applied to a multivariate data matrix $\boldsymbol{X}$, the sample mean of the *k*th variable is

$$\bar{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_{ik}\,, k = 1,2,\dots,p, \qquad (3)$$

the sample variance is defined by

$$s_k^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2\,, k = 1,2,\dots,p. \qquad (4)$$

Next we introduce a matrix that contains statistics that relate pairs of variables (x_i, x_k), sample covariance

$$s_{ik}s_{ik} = \frac{1}{n}\sum_{j=1}^{n}\left(x_{ji} - \bar{x}_i\right)\left(x_{jk} - \bar{x}_k\right), i = 1,2,\dots,p\,,$$

$$k = 1,2,\dots,p. \qquad (5)$$

It follows that $s_{ik} = s_{ki}$ and $s_{ii} = s_i^2$, the sample variance.

Matrix of sample covariances

$$\boldsymbol{S_n} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ s_{31} & s_{32} & \cdots & s_{3p} \\ \vdots & \vdots & \vdots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix} \qquad (6)$$

is symmetric.

THEOREM Let $\boldsymbol{S_n}$ be the $p \times p$ covariance matrix related to the multivariate data matrix $\boldsymbol{X}$. Let eigenvalues of $\boldsymbol{S_n}$ be $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, and corresponding orthonormal eigenvectors be $\boldsymbol{u_1}, \boldsymbol{u_2}, \dots, \boldsymbol{u_p}$. Then ith principal component $\boldsymbol{y}_i$ is given by the linear combination of the original variables in the data matrix $\boldsymbol{X}$ (Kolman 2004):

$$\boldsymbol{y}_i = \boldsymbol{X}\boldsymbol{u_i}, i = 1,2,..,p. \qquad (7)$$

The variance of $\boldsymbol{y}_i$ is λ_i, and $\mathrm{cov}(\boldsymbol{y}_i, \boldsymbol{y}_j) = 0, i \neq j$. The total variance of the data in $\boldsymbol{X}$ is equal to the sum of eigenvalues:

$$\sum_{j=1}^{p} s_{jj} = \sum_{j=1}^{p} \lambda_j. \qquad (8)$$

$$\textit{Proportion of the total variance covered by the kth principal component} = \frac{\lambda_k}{\sum_{j=1}^{p}\lambda_j}. \qquad (9)$$

If a large percentage of the total variance can be attributed to the first few components, then these new variables can replace the original variables without significant loss of information. Thus we can achieve significant reduction in data.

Principal Components of Sample Texts

Next, matrices of sample covariances for the textual descriptors for the texts are computed.

The information in the covariance matrix is used to define a set of new variables as a linear combination of the original variables in the data matrices $\boldsymbol{X_{ivo}}$, $\boldsymbol{X_{mesa}}$, etc.. The new variables are derived in a decreasing order of importance. The first of them is called first principal component and accounts for as much as possible of the variation in the original data. The second of them is called second principal component and accounts for another, but smaller portion of the variation, and so on.

If there are *p* variables, to cover all of the variation in the original data, one needs *p* components, but often much of the variation is covered by a smaller number of components. Thus PCA has as its goals the interpretation of the variation and data reduction.

In fact PCA is nothing but the spectral decomposition of the covariance matrix.

Variances and percentage variances covered by fourteen principal components of the textual descriptors for the sample texts consisting randomly chosen 400 paragraphs of six chosen works of six authors are shown in Table 2.

Table 2. Percentages of variances covered by fourteen principal components of the textual descriptors used in this research.

Princ. Comp.	Andrić Cuprija	Selimović Derviš
1	75.600	77.112
2	24.127	22.400
3	0.083	0.204
4	0.054	0.088
5	0.032	0.048
6	0.029	0.041
7	0.022	0.029
8	0.016	0.024
9	0.014	0.022
10	0.009	0.015
11	0.008	0.010
12	0.005	0.006
13	0.002	0.002
14	0.001	0.000
Total	100	100

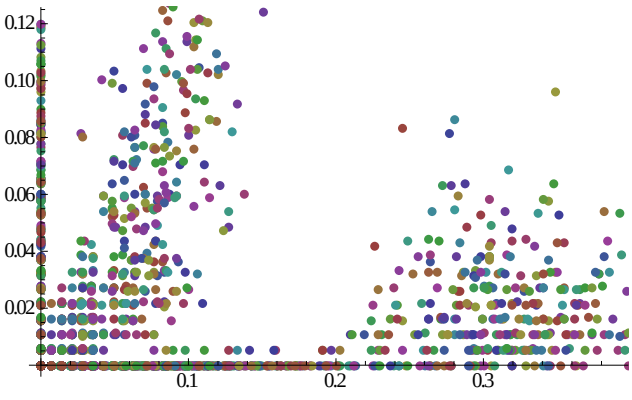
Princ. Comp.	Sušić Pobune	Kovačić Registraturi
1	74.580	63.700
2	24.845	35.424
3	0.200	0.588
4	0.154	0.104
5	0.073	0.058
6	0.040	0.035
7	0.033	0.030
8	0.024	0.024
9	0.019	0.019
10	0.015	0.012
11	0.006	0.006
12	0.005	0.001
13	0.004	0.000
14	0.002	0.000
Total	100	100

Princ. Comp.	Soljenitsin Odjel	Ranko Ruke
1	75.312	57.334
2	24.338	41.494
3	0.740	0.246
4	0.074	0.173
5	0.052	0.071
6	0.038	0.060
7	0.032	0.053
8	0.017	0.032
9	0.017	0.020
10	0.013	0.010
11	0.006	0.006
12	0.001	0.001
13	0.001	0.000
14	0.000	0.000
Total	100	100

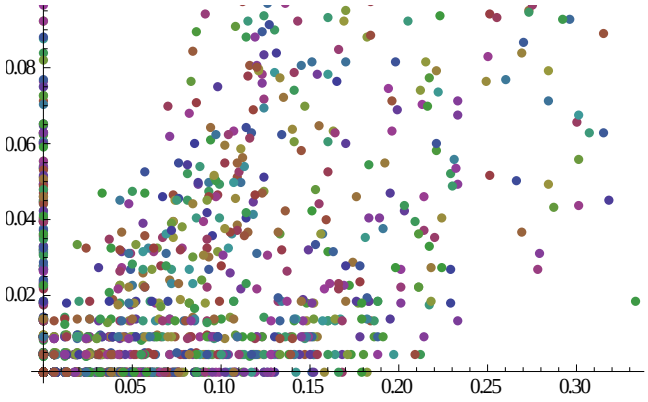
Table 2 reveals that the first two principal components cover more than %99 of variances of principal components.

It is seen that first principal component covers around 75% of the variance. Therefore in this article to classify the texts, we'll rely on only first principal components. The interval [-500, 350] covers the support of first principal components of all 400 paragraph random samples and of all texts. This interval is divided into 50 equal bins, and frequencies of the data in the principal components are counted for 500 samples for each text. The average of the frequency distributions is shown in the following figures.

Figure 1. in the below displays ListPlot of 100 sample of normalized frequencies of first (horizontal axis) and second (vertical axis) principal components of Ivo Andrić's Cuprija u Drini (a), Meša Selimović's Derviš i Smrt (b). Different colors refer to different batches of normalized frequencies. These clusters are writerprints of the authors and historically used for authorship attribution.



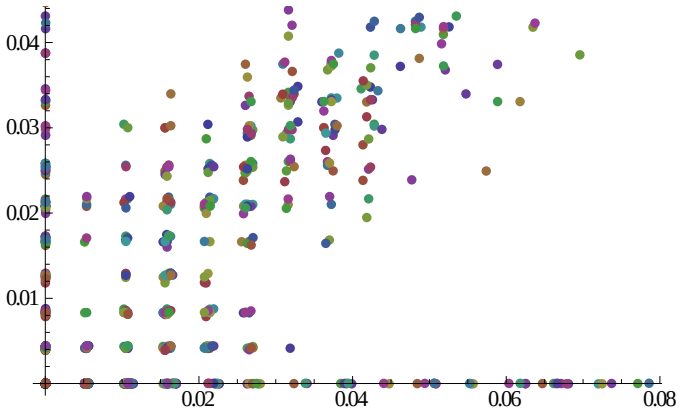
(a)



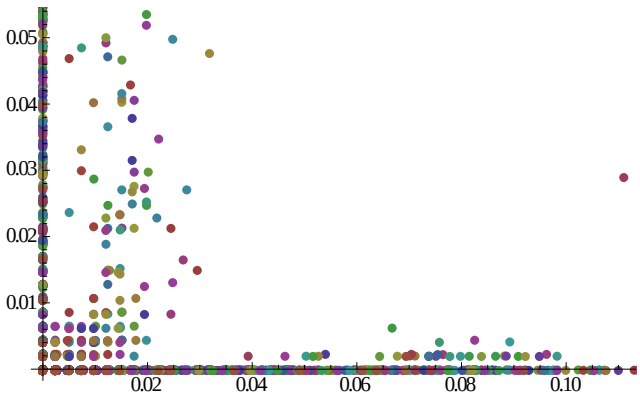
(b)

Figure 1. ListPlot of 100 sample normalized frequencies of first (horizontal axis) and second (vertical axis) principal components of Ivo Andrić's Cuprija u Drini (a), Meša Selimović's Derviš i Smrt (b).

In the below, Figure 2 displays ListPlot of 100 sample normalized frequencies of first (horizontal axis) and second (vertical axis) principal components of Derviš Sušić's Pobune (a), Ante Kovačić's U registraturi (b).



(a)



(b)

Figure 2. ListPlot of 100 sample normalized frequencies of first (horizontal axis) and second (vertical axis) principal components of Derviš Sušić's Pobune (a), Ante Kovačić's U registraturi (b).

Figure 3. in the below displays ListPlot of 100 sample normalized frequencies of first (horizontal axis) and second (vertical axis) principal components of Alexander Soljenitsin's Odjel za Rak (Cancer Ward) (a), and Ranko Maronković's Ruka (b).

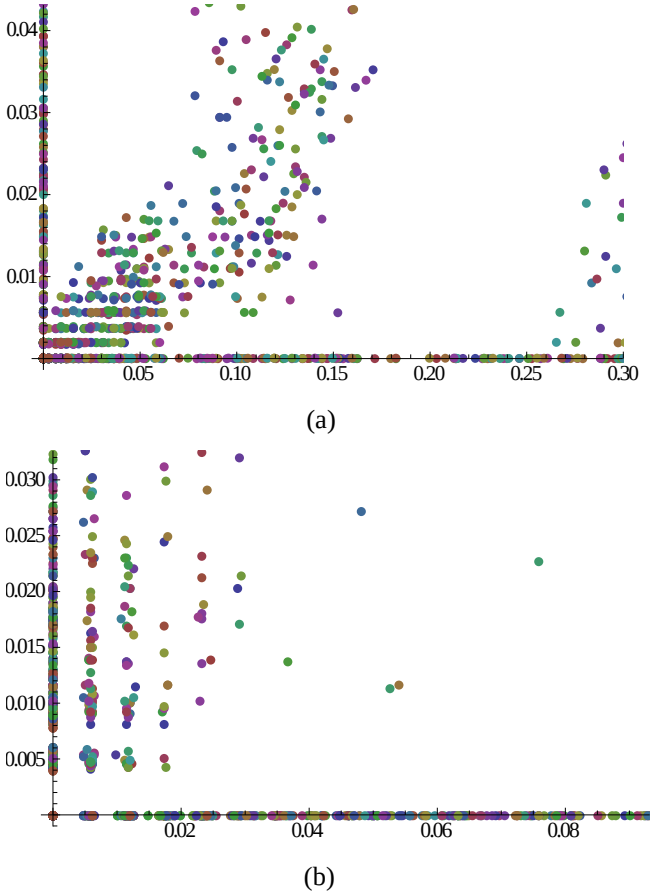


Figure 3. ListPlot of 100 sample normalized frequencies of first (horizontal axis) and second (vertical axis) principal components of Alexander Soljenitsin's Odjel za Rak (Cancer Ward) (a), and Ranko Maronković's Ruka (b).

The frequency profile of first principal components of the textual data seems to be almost invariant throughout a text. There are similarities in the frequency profiles of the text authored by the same person. In Figure 4, Ivo Andrić's Proklet Avlija, (compare with Cuprija u Drini Figure 1a) (a), Alexander Sojenitsin's Krugu (compare with Odjel Za Rak in figure 2a) (b).

Therefore these frequency profiles can be regarded as writerprints. However a visual identification of the authors of these writerprints seems to be difficult. To help the classification of these writerprints, we propose to take it as a pattern classification task, and use artificial neural networks, more specifically perceptrons with memory-based learning neurons to do the job.

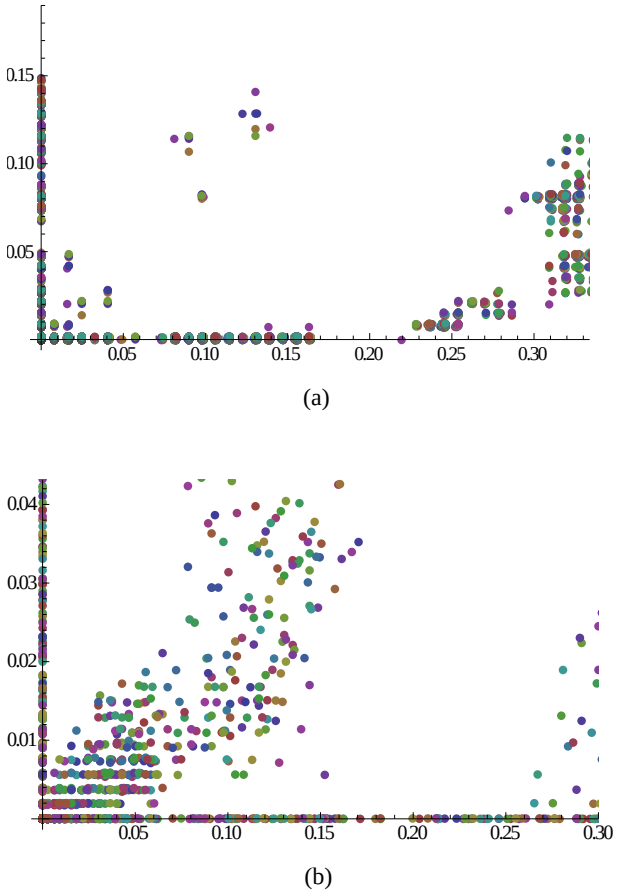


Figure 4. Ivo Andrić's Proklet Avlija, (compare with Cuprija u Drini in Figue 1a) (a), Alexander Sojenitsin's Krugu (compare with Odjel Za Rak in figure 3a) (b).

Geometry of the Problem

In the problem under consideration, each of the 50 dimensional input pattern vector x has unit Euclidean length so that we may view it as a point on an N-dimensional unit sphere where N=50 is the number of input nodes as well. Therefore, frequencies of first principal components of input data from six novels representing six authors are represented by dots in Figure 5. The mean vectors of these clusters are shown by crosses for the purpose of demonstration in the figure.

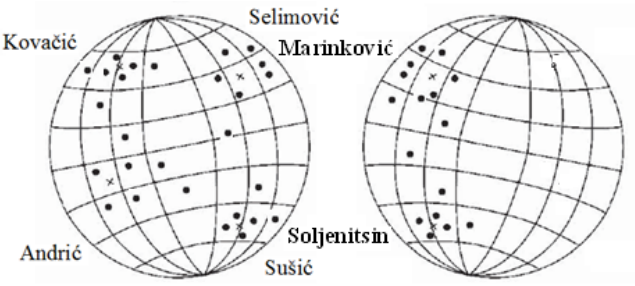


Figure 5. Geometric interpretation of the competitive learning process. The dots represent the input vectors, and the crosses represent the averages of the normalized input vectors.

Cosines of angles between normalized means of frequencies of first principal components of input data from six novels to represent six authors in Figure 5 are shown in the following table. These mean vectors are shown by crosses for the purpose of demonstration in the figure.

Table 3. Cosines of angles between normalized means of novels representing their authors.

	Andrić	Selim	Sušić	Kova	Soljen	Ranko
Andrić	1.00	0.95	0.77	0.84	0.88	0.67
Selim	0.95	1.00	0.91	0.95	0.93	0.78
Sušić	0.77	0.91	1.00	0.94	0.85	0.87
Kovač	0.84	0.95	0.94	1.00	0.94	0.80
Soljen	0.88	0.93	0.85	0.94	1.00	0.82
Ranko	0.67	0.78	0.87	0.80	0.82	1.00

According to Table 3, The nearest neighbor of Andrić is Selimović in a 0.05 cosine difference circle. While Andrić has no other neighbor in a 0.22 circle, Selimović has Kovacić in the same 0.05 circle. Sušić's nearest neighbor is Kovacić. For Kovacić, Sušić and Soljenitsin are closest neighbors with a cosine difference distance of 0.06. Marinković is an isolated island. His closest neighbor Sušić is 0.13 cosine difference away.

Let us randomly choose 500 batches of 300 data from each novel. Find 500 principal components from each novel. When frequencies of contents of these 500 principal components are distributed in 50 bins, we have 500 normalized 50×1 data vectors for each novel. The angular distances between these ten clusters can be represented by the cosine of the angles between normalized means of these clusters.

For the chosen ten novels, the mutual distances can be summarized by the scalar products of mean vectors as in Table 4.

Table 4. Cosines of angles between normalized means of all novels considered.

	cpr	*prk*	*znk*	*der*	*pob*	*reg*	*fisk*	*krg*	*odj*	*ruk*
cpr	1.	0.97	0.86	0.95	0.77	0.84	0.63	0.87	0.88	0.67
prk	0.97	1.	0.86	0.94	0.74	0.86	0.66	0.9	0.93	0.64
znk	0.86	0.86	1.	0.71	0.41	0.57	0.42	0.77	0.72	0.37
der	0.95	0.94	0.71	1.	0.91	0.95	0.76	0.87	0.93	0.78
pob	0.77	0.74	0.41	0.91	1.	0.94	0.84	0.77	0.85	0.87
reg	0.84	0.86	0.57	0.95	0.94	1.	0.81	0.85	0.94	0.8
fisk	0.63	0.66	0.42	0.76	0.84	0.81	1.	0.79	0.83	0.77
krg	0.87	0.9	0.77	0.87	0.77	0.85	0.79	1.	0.97	0.83
odj	0.88	0.93	0.72	0.93	0.85	0.94	0.83	0.97	1.	0.82
ruk	0.67	0.64	0.37	0.78	0.87	0.8	0.77	0.83	0.82	1.

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 generalizations of these concepts with mathematical model of artificial neuron due to McCuloch and Pitts described in (McCuloch, and Pitts 1943), and the first implementation of Rosenblatt's perceptron in (Rosenblatt 1958). The efficiency and applicability of artificial neural networks to computational tasks have been questioned many times, especially at the very beginning of their history the book "Perceptrons" by Minsky and Papert (Minsky, and Papert 1998) 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 artificial neural networks regained their status and proved beyond doubt to be sufficiently good approach to many problems.

Memory-Based Learning

In memory-based learning, all (or most) of the past experiences are explicitly stored in a large memory of correctly classified input-output examples: $\{(\boldsymbol{x}_i, d_i)\}_{i=1}^{N}$where $\boldsymbol{x}_i$ denotes an input vector and d_i denotes the corresponding desired response. Without loss of generality, we have restricted the desired response to be a scalar. For example, in a binary pattern classification problem there are two classes/hypotheses, denoted by $\mathfrak{C}_1$ and $\mathfrak{C}_2$, to be considered. In this example, the desired response d_i takes the value 0 (or -1) for class $\mathfrak{C}_1$ and the value 1 for class $\mathfrak{C}_2$. When classification of a test vector $\boldsymbol{x}_{test}$, (not seen before) is required, the algorithm responds by retrieving and analyzing the training data in a "local neighborhood" of $\boldsymbol{x}_{test}$.
All memory-based learning algorithms involve two essential ingredients:

- Criterion used for defining the local neighborhood of the test vector $\boldsymbol{x}_{test}$.
- Learning rule applied to the training examples in the local neighborhood of $\boldsymbol{x}_{test}$.

The algorithms differ from each other in the way in which these two ingredients are defined.

Nearest Neighbor Rule

In a simple yet effective type of memory-based learning known as the *nearest neighbor rule*, the local neighborhood is defined as the training example that lies in the immediate neighborhood of the test vector $\boldsymbol{x}_{test}$. In particular, the vector

$$\boldsymbol{x'}_N \in \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N\} \tag{10}$$

is said to be the nearest neighbor of $\boldsymbol{x}_{test}$ if

$$\min_i d(\boldsymbol{x}_i, \boldsymbol{x}_{test}) = d(\boldsymbol{x'}_N, \boldsymbol{x}_{test}) \tag{11}$$

where $d(\boldsymbol{x}_i, \boldsymbol{x}_{test})$ is the Euclidean distance between the vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_{test}$.

The class associated with the minimum distance, that is, vector $\boldsymbol{x'}_N$ is reported as the classification of $\boldsymbol{x}_{test}$. This rule is independent of the underlying distribution responsible for generating the training examples.

Cover and Hart (Cover, and Hart 1967), Dasarathy (Dasarathy 1991), and Hodges (Fix, and Hodges 1951) have

formally studied the nearest neighbor rule as a tool for pattern classification. The analysis presented therein is based on two assumptions:

- The classified examples $(\boldsymbol{x_i}, d_i)$ are *independently and identically distributed (iid)*, according to the joint probability distribution of the example ***(x, d)***.
- The sample size N is infinitely large.

Under these two assumptions, it is shown that the probability of classification error incurred by the nearest neighbor rule is bounded above by twice the Bayes probability of error, that is, the minimum probability of error over all decision rules. In this sense, it may be said that half the classification information in a training set of infinite size is contained in the nearest neighbor, which is a surprising result.
A variant of the nearest neighbor classifier is the *k*-nearest neighbor classifier, which proceeds as follows:

- Identify the *k* classified patterns that lie nearest to the test vector $\boldsymbol{x_{test}}$ for some integer *k*.
- Assign $\boldsymbol{x_{test}}$ to the class (hypothesis) that is most frequently represented in the *k* nearest neighbors to $\boldsymbol{x_{test}}$ (i.e., use a majority vote to make the classification).

Thus the k-nearest neighbor classifier acts like an averaging device. In particular, it discriminates against a single outlier, as illustrated in Figure 5. for k= 3. An outlier is an observation that is improbably large for a nominal model of interest.

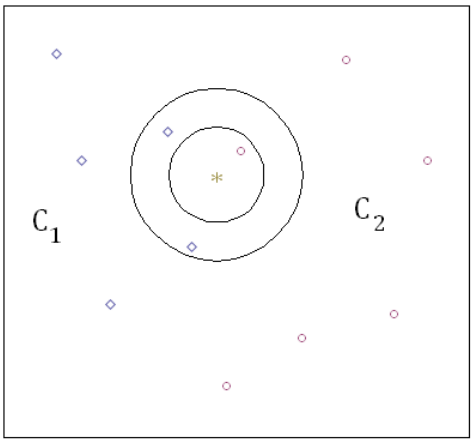


Figure 5. Elements of the set are classified into two categories C_1(diamond), and C_2 (circle). The test point (star) belongs to C_2 if classified by nearest neighbor classifier, belongs to $\boldsymbol{C_1}$ if classified by 3-nearest neighbor classifier.

APPLICATION TO AUTHOR ATTRIBUTION

Author identification 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 analyzed - *the writerprints of the authors of previously selected texts*,
- then followed the third phase of calculating characteristics for all descriptors, *calculation*,
- transform randomly chosen data matrices into matrices with principal components *principal component analysis*,
- count frequencies of principal components in bins of equal length that were later used for training of the neural network, *calculation of frequencies in bins*,
- specification of the network with its architecture and learning method can be seen as the fourth step of the whole procedure, *neural network*,
- the fifth consisted of the actual *training of the network*,
- the sixth stage is *testing*,
- and the final one corresponded to analysis of obtained results and coming up with some conclusions and possible indicators for improvement, *analysis of obtained results.*

In this paper, the training phase is simply sending training data to the neural network. All of the six output neurons memorize this data.

The input vector $\boldsymbol{x}$ is 50 dimensional with components as frequencies in corresponding bins as shown in the signal flow graph in Figure 6. Algorithm results in a decision about attribution of paragraphs whose textual description entered in the form of frequencies in bins of principal components as inputs.

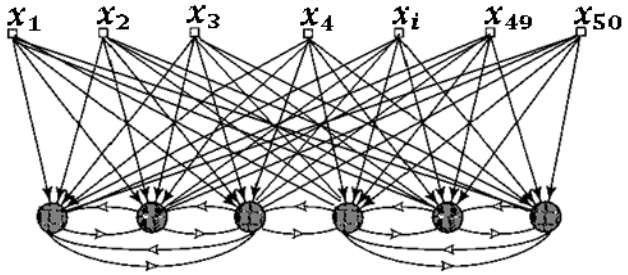


Figure 6. Architectural graph of a simple memory based learning network with feedforward (excitatory) connections from the source nodes to the neurons, and lateral (inhibitory) connections among the neurons; the lateral connections are signified by open arrows.

We have chosen 500 sets of 400 paragraphs from each of the six texts, which are not used for training. Each 400 paragraph set is transformed into its principal components, and only first principal components are taken into account. Hence we have 500 first principal components from each text. Then principal components are transformed into data vectors whose elements are frequencies in 50 uniformly specified bins. The resulting data is a 500×50 matrix for each text.

Training phase is completed simply sending 3000 data to the neural network to be memorized together with their authors.

Then the test data consisting of a random mixture of 500 test data from each text, totally 3000 unlabeled mixed data is sent to the neural network for classification. When a data vector is entered to the neural network, the angular distance of this data vector, to the 3000 training vectors sent previously are calculated by neurons collectively. The authorship of the nearest, say five training data vectors are remembered by the network. If the majority of these data points are authored by writer A, this test point attributed to author A, and the output neuron appointed to keep the records of paragraphs authored by this writer, records this result. Each of 3000 test data are classified similarly. The correct classification percentages obtained through a testing experiment are as follows.

Table 5. Results of author identification of 3000 mixed test paragraphs.

Andrić	Selim	Sušić	Kova	oljen	Marin	%Corr
499	500	503	500	501	497	99.87

When test data for ten books are sent individually to the neural network for identification, the author attributions are as in Table 6.

Table 6. The author attributions of the ten books.

Attribution	*Cupri*	*Znako*	*Prokl*	*Dervi*	*Pobu*
Andrić	500	422	497	0	0
Selimović	0	0	0	500	0
Sušić	0	0	0	0	500
Kovačić	0	0	0	0	0
Soljenitsin	0	78	3	0	0
Marinković	0	0	0	0	0
% Success	100	84.4	99.4	100	100

Attribution	*Regist*	*Fiškal*	*Odjel*	*Krug*	*Ruke*
Andrić	0	0	0	0	0
Selimović	0	0	0	0	0
Sušić	0	183	0	0	3
Kovačić	500	300	0	0	0
Soljenitsin	0	17	500	500	0
Marinković	0	0	0	0	497
% Success	100	60	100	100	99.4

As it is seen from tables above, the neural network is successful in the test data from the texts it trained for. The successes in the classification of other books of the same authors are also satisfactory in general. Authors like Ivo Andrić have writerprints that are characteristic for all novels. In Soljenitsin case the characteristics even do not disturbed by translation to foreign languages.

TESTING PCA ALGORITHM ON ANOTHER GROUP OF AUTHORS

To show that PCA algorithm is equally successful on other groups of authors, and languages. Let us consider five novels authored by Jane Austin, and six novels of Charles Dickens as shown in Table 11.

Table 7. Twelve books authored by Jane Austin, Charles Dickens.

Jane Austin		Charles Dickens	
1	Mansfield Park	7	Great Expectations
2	Sense and Sensibility	8	David Copperfield
3	Pride and Prejudice	9	Bleak House
4	Persuasion	10	Oliver Twist
5	Northanger Abbey	11	The Pickwick Papers
6	Emma	12	A Tale of two cities

From each book at least 1000 paragraphs are considered. Multilayer perceptron described in the above is trained by randomly chosen 300 batches of length 500 paragraphs from Great Expectations (Dickens), and Sense Sensibility (Austin). Then randomly chosen 300 paragraphs from all of twelve books are sent to the multilayered perceptron for classification. The result is shown in Table 12.

Table 8. The author attributions of the twelve books.

Attribution	*1*	*2*	*3*	*4*	*5*	*6.*
Jane Austin	283	300	300	289	34	299
Charles	17	0	0	11	266	1
% Success	94.3	100	100	96.3	21.3	99.7

Attribution	*7*	*8*	*9*	*10*	*11*	*12*
Jane Austin	0	0	2	261	0	0
Charles	300	300	298	39	300	300
% Success	100	100	99.3	13	100	100

Classification is successful with almost 100% accuracy. Northanger Abbey and Oliver Twist are exceptions. Charles Dickens's, and Austin's stylometric styles shifts in writing these books. For details, one needs to refer critics of these two novels.

CONCLUSIONS

The research described in this paper concerning author identification analysis shows that the method of principal component analysis (PCA), when followed by an artificial neural network is an efficient tool. Thus a series of future experiments should include wider range of authors, definition of new sets of textual descriptors, and test for other types and structures of neural networks, and search the possibility of inheritance through translation into other languages.

REFERENCES

Abbasi A, and Chen H (2005) Applying authorship analysis to extremist-group Web forum messages, IEEE Intelligent Systems.

Argamon-Engelson S, Koppel M, and Avneri G (1998) Style-based text categorization: What newspaper am I reading?, in Proc. of AAAI Workshop on Learning for Text Categorization, 1998, pp. 1-4.

Argamon S, Koppel M, Fine J, and Shimoni A (2003) Gender, Genre, and Writing Style in Formal Written Texts, Text 23(3).

Argamon S, Koppel M, Pennebaker J, and Schler J (2009) Automatically Profiling the Author of an Anonymous Text, Communications of the ACM , 52 (2), pp. 119-122.

Binongo J (2003) Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution. Chance 16(2): 9–17.

Binongo JNG, and Smith MWA (1999) The application of principal component analysis to stylometry, Lit Linguist Computing 14: 445-466.

Bosch R, and Smith J (1998) Separating hyperplanes and the authorship of the disputed federalist papers. American Mathematical Monthly 105(7): 601–8.

Burrows JF (1987) "Word Patterns and Story Shapes: The Statistical Analysis of Narrative Style", Literary and Linguistic Computing, 2, 61-70.

Burrows JF 1989) 'An ocean where each kind..': Statistical analysis and some major determinants of literary style, Computers and the Humanities 23(4), 309-321.
Burrows JF (1992) Not Unless You Ask Nicely: The Interpretative Nexus Between Analysis and Information, Literary and Linguistic Computing, 7(2):91-109.

Can M, Jamak A, and Savatić A (2011) Teaching Neural Networks to Detect the Authors of Texts Using Lexical Descriptors, 9th International Conference on Knowledge, Economy & Management Proceedings,ISBN: pp. 3607-3624.

Can M, Hadžiabdić KK, and Demir NM (2011) Teaching Neural Networks to Detect the Authors of Texts, Using Lexical Descriptors, 9th International Conference on Knowledge, Economy & Management Proceedings, pp. 1393-1402.

Chaski C (2001) Empirical evaluations of language-based author identification techniques, Journal of Forensic Linguistics. 8(1): 1–65.

Chaski C (2005) Who's at the keyboard? Authorship attribution in digital evidence investigations, International Journal of Digital Evidence 4(1).

Cover TM, and Hart PE (1967) Nearest neighbor pattern classification, IEEE Transactions on Information Theory, IT-13:21-7.

Dasarathy BV (1991) Nearest Neighbor (NN) Norms: Nn Pattern Classification Techniques, IEEE Computer Society Press.

De Vel O, Anderson A, Corney M, and Mohay GM (2001) Mining e-mail content for author identification forensics. SIGMOD Record 30(4), pp. 55-64.

Diederich J, Kindermann J, Leopold E, and Gerhard P (2003) Authorship Attribution with Support Vector Machines, Applied Intelligence, Volume 19, Numbers 1-2, 109-123.

Dumais SJ, Platt J, Heckerman D, and Sahami M (1998) Inductive learning algorithms and representations for text categorization, Proceedings of ACM-CIKM98, 148-155.

Fix E, and Hodges JL (1951) Discriminatory analysis – nonparametric discrimation: consistency properties, Report No. 4, Project No. 21-29-. 004.

Fung D (2003) The disputed Federalist Papers: SVM feature selection using concave minimization, Proceedings of the 2003 Conference on Diversity in Computing. 42–6.

Genkin A, Lewis D, and Madigan D (2006) Large-scale Bayesian logistic regression for text categorization, Technometrics.

Graham N, Hirst G, and Marthi B (2005) Segmenting documents by stylistic character. Natural Language Engineering, 11(4), December 2005, 397- 415.

Hayes JF (2008) Authorship Attribution: A Principal Component and Linear Discriminant Analysis of the Consistent Programmer Hypothesis, I. J. Comput. Appl. 15, No. 2, 79-99.

Holmes D (2003) Stylometry and the Civil War, Chance 16(2).

Holmes D, and Forsyth R (1995) The Federalist revisited: New directions in authorship attribution, Literary and Linguistic Computing 10(2): 112–127.
Holmes D, Gordon L, and Wilson C (2001) A widow and her soldier: Stylometry and the American Civil War. Literary and Linguistic Computing 16(4): 403–20.

Holmes D (1998) The evolution of stylometry in humanities scholarship. Literary and Linguistic Computing 13(3): 111–7.

Hoorn J, Frank S, Kowalczyk W, and van der Ham F (1999) Neural network identification of poets using letter sequences. Literary and Linguistic Computing, 14(3) pp. 311-338.

Jamak A, Savatić A, and Can M (2012) Principal component analysis for authorship attribution, Business Systems Research, Vol. 3. No.2, pp. 49-56.

Juola P, and Baayen H (2003) A controlled-corpus experiment in authorship identification by cross-entropy', Literary and Linguistic Computing.

Juola P (1998) Cross-entropy and linguistic typology. In Proceedings of New Methods in Language Processing 3. Sydney, Australia.

Juola P (2006) Authorship attribution, Foundations and Trends in Information Retrieval 1(3): 233–334.

Kjell B(1994) Authorship determination using letter pair frequency features with neural network classifiers. Literary and Linguistic Computing 9(2): 119–24.

Khmelev DV (2001) Disputed Authorship Resolution through Using Relative Empirical Entropy for Markov Chains of Letters in Human Language Text, Journal of Quantitative Linguistics, 7(3), 201-207.

Khmelev DV, and Tweedie FJ (2002) Using Markov chains for identification of writers. Literary and Linguistic Computing, 16(4) pp. 299-307.

Kjell B, Woods W, and Frieder O (1995) Information retrieval using letter tuples with neural network and nearest neighbor classifiers. In IEEE International Conference on Systems, Man and Cybernetics, volume 2, pp. 1222-1225, Vancouver, BC.

Kjell B (1994) Authorship attribution of text samples using neural networks and Bayesian classifiers. In IEEE International Conference on Systems, Man and Cybernetics, San Antonio, TX.

Kolman B, and Hill DR (2004) Elementary Linear Algebra, Pearson, New Jersey.

Koppel M, and Schler J (2003) Exploiting Stylistic Idiosyncrasies for Authorship Attribution, in Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis, pp. 69-72.

Koppel M, Schler J, and Zigdon K (2005) Determining an Author's Native Language by Mining a Text for Errors, Proceedings of KDD '05, Chicago IL.

Koppel M, Argamon S, and Shimoni A (2002) Automatically categorizing written texts by author gender, Literary and Linguistic Computing 17(4), pp. 401-412.

Koppel M, Akiva N, and Dagan I (2006) Feature Instability as a Criterion for Selecting Potential Style Markers, Journal of the American Society for Information Science and Technology 57(11), pp. 1519-1525.

Kukushkina O, Polikarpov A, and Khmelev D (2002), Using literal and grammatical statistics for authorship attribution. Problemy Peredachi Informatsii 37(2).

Lowe D, and Matthews R (1995) Shakespeare vs. Fletcher: A stylometric analysis by Radial Basis Functions. Computers and the Humanities, 29 pp. 449-461.

Li R, Zheng J, Chen H, and Huang Z (2006) 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, no. 3, pp. 378–393.

Madigan D, Genkin A, Lewis DD, Argamon S, Fradkin D, and Ye L (2006) Author Identification on the Large Scale, Proc. of Classification Society of N. America.

Matthews R, and Merriam T (1993) Neural computation in stylometry: An application to the works of Shakespeare and Fletcher. Literary and Linguistic Computing, 8(4), pp. 203-209.

Mcculloch WS, and Pitts W (1943) A Logical Calculus of the Ideas Immanent in Nervous Activity, Bulletin of Mathematical Biophysics, 5:115-133. Reprinted in Anderson& Rosenfeld [1988], pp. 18-28.

Merriam T, and Matthews R (1994) Neural computation in stylometry II: An application to the works of Shakespeare and Marlowe. Literary and Linguistic Computing 9, pp. 1-6.

Minsky ML, and Papert SA (1988) Perceptrons, Expanded Edition. Cambridge, MA: MIT Press. Original edition.

Peng F, Schuurmans D, and Wang S (2004) Augmenting Naive Bayes Text Classifier with Statistical Language Models , Information Retrieval, 7 (3-4), pp. 317 – 345.

Peng R, and Hengartner N (2002) Quantitative analysis of literary styles. The American Statistician 56(3): 175–85.
Rosenblatt E (1958) The Perceptron: A probabilistic model for information storage and organization in the brain," Psychological Review, vol. 65, pp. 386-408.

Sanderson C, and Guenter S (2006) Short Text Authorship Attribution via Sequence Kernels, Markov Chains and Author Unmasking: An Investigation, in Int'l Conference on Empirical Methods in Natural Language Processing, pp. 482-491.

Smith J (2008) A review of authorship attribution, MUMT 611: Music Information Acquisition, Preservation, and Retrieval, Report.

Williams C (1975) Mendenhall's studies of word-length distribution in the works of Shakespeare and Bacon, Biometrika 62(1): pp. 207–12.

Yang Y (1999) An evaluation of statistical approaches to text categorization. Journal of Information Retrieval, 1 (1-2), pp 67-88.

Zhao Y, and Zobel J (2005) Effective authorship attribution using function word, in 'Proc. 2nd AIRS Asian Information Retrieval Symposium', Springer, pp. 174-190.