Principal component analysis for authorship attribution

Amir Jamak, Alen Savatić and Mehmet Can
International University of Sarajevo, Faculty of Engineering and Natural Sciences, Sarajevo, Bosnia and Herzegovina

Abstract

Background: To recognize the authors of the texts by the use of statistical tools, one first needs to decide about the features to be used as author characteristics, and then extract these features from texts. The features extracted from texts are mostly the counts of so called function words. **Objectives:** The data extracted are processed further to compress as a data with less number of features, such a way that the compressed data still has the power of effective discriminators. In this case feature space has less dimensionality then the text itself. **Methods/Approach:** In this paper, the data collected by counting words and characters in around a thousand paragraphs of each sample book, underwent a principal component analysis performed using neural networks. Once the analysis was complete, the first of the principal components is used to distinguish the books authored by a certain author. **Results:** The achieved results show that every author leaves a unique signature in written text that can be discovered by analyzing counts of short words per paragraph. **Conclusions:** In this article we have demonstrated that based on analyzing counts of short words per paragraph authorship could be traced using principal component analysis. Methodology could be used for other purposes, like fraud detection in auditing.

Keywords: principal components, authorship attribution, stylometry, text categorization, function words, classification task, stylistic features, syntactic characteristics.
JEL classification: C65
Paper type: Research article

Recieved: 18, November, 2011
Revised: 19, June, 2012
Accepted: 22, July, 2012

Citation: Jamak, A., Savatić, A., Can, M. (2012). "Principal component analysis for authorship attribution", Business Systems Research, Vol. 3, No. 2, pp. 49-56.
DOI: 10.2478/v10305-012-0012-2

Introduction

Although the author identification problem is the oldest of the all text classification problems, it is still not well organized. Throughout its history, it is mishandled by statisticians. At the date it has wide application in disperse areas such as law, education and internet security. The most simple version of it, the stylometry dates back to Augustus de Morgan's suggestion in 1851 that authors of the various Bibles might be distinguishable from the length of the words they use (Holmes, 1998).

The length of words used as a stylistic feature was worthy of investigation. Mendenhall (1887) started the research on the plays of Shakespeare. To his findings while Shakespeare and Marlowe were nearly indistinguishable, they were both significantly and consistently different from Bacon (Williams, 1975) in their use of long words: Shakespeare used more four - letter words and Bacon more three-letter words.

In Russia, first attempts in the search of specific quantitative structures for individual author style took place with the works of Morozov (Morozov, 1915). A famous mathematician Markov (Markov 1916) argued that characteristics of writer's style through individual word frequencies, like the distribution of negative word "ne" offered by Morozov are unstable. A.A. Markov had given an example of good statistical approach in his work by using distribution of vowels and consonants. In those early years, the

main problem was the lack of objective evaluation of the proposed methods. In most of the cases, the testing ground was literary works of unknown or disputed authorship, so the estimation of attribution accuracy was not even possible.

Linguist George Kingsley Zipf proposed an empirical law using statistics. He claims that many types of data studied in the physical and social sciences can be approximated with a Zipfian distribution. Zipfian distribution is a member of the family of discrete power law probability distributions (Zipf 1935). In linguistics, Zipf defined statistical rank of a word which is inversely proportional to its frequency. "Yule's characteristic K," is found by G. Udny Yule which defines 'vocabulary richness' by comparing word frequencies which distributes according to a Poisson distribution (Yule 1944). During the following decades, this future also found to be an unreliable marker of style just like Mendenhall, Morozov approaches (Holmes, 1998), average sentence length, number of syllables per word, and other estimates of vocabulary richness such as Simpson's D index (Simpson 1949) and a ratio of the number of rare or unique words, or types, to the number of total words, or phrases (Juola et al., 2006).

In United States, in the years 1787 and 1788, three authors John Jay, Alexander Hamilton and James Madison collectively wrote 85 newspaper essays supporting the ratification of the constitution. All of these 85 assays signed by the pseudonym "Publius,". The three authors later revealed which of the 67 Federalist Papers they had written; however 12 of these assays were claimed by both Hamilton and Madison. The ground breaking contribution to the field of stylometric analysis for authorship attribution basically came by the joint works of Mosteller and Wallace in 1964 on 12 disputed Federalist Papers (Mosteller, and Wallace 1964).

Function words like but, and, upon, then, for are regarded as characteristics of authors style. They are assumed to be unconsciously generated and used independent of the meaning and context. An author may have a preference for modes of expression. Hamilton and Madison have nearly identical sentence length distributions (Juola, 2006), Mosteller and Wallace found sharp differences in their preference for different function words: for instance, as quoted by Holmes (Holmes, 1998) the word "upon" is fifteen times more frequent in Hamilton. Mosteller and Wallace adjusted these frequencies with a Bayesian model and showed that Madison is most likely the author of all 12 disputed papers, supporting the historians.

The problem of the identification of 12 disputed Federalist Papers is one of the most challenging issues of the field, and has been used as a benchmark to test most new methods (see, for instance, Kjell, 1994, Holmes and Forsyth, 1995; Bosch and Smith, 1998; Fung, 2003).

The goal of this research is to propose a method to distinguish books authored by a certain author among the other books written by various authors. The proposed method could be used in fraud detection and similar areas.

In the following sections we will explain the conducted research. The next section will give the problem definition; the third section describes principal component analysis methodology. The fourth section explains the application of the method to samples of text and the results of the research are given in the fifth section. The final section brings the discussion and concludes the article.

Problem Definition

In this paper, an application of principal component analysis is presented. The authorship attribution is considered as a classification task (Chaski, 2001, 2005). Texts studied are literary works of three Bosnian writers, Ivo Andrić (1892-1975), M. Meša Selimović (1910-1982), and Derviš Sušić (1925–1990). Authors of the texts use lexical and syntactical components rather subconsciously. Therefore for author identification purposes, they are selected to characterize styles of writers. These features are difficult to repeat by others, and they can be used as author invariants. Principal components of data elicited from texts possess generalization properties, which allow for the required high accuracy of classification (Hayes, 2008).

Texts Used

This research focused on and used the texts of three famous Bosnian writers, Ivo Andrić, M. Meša Selimović and Derviš Sušić. Sections of the novels authored, supply large structured sets of data enough to capture features that are characteristic to this writer. This data is used as training data to represent

other parts of the text, even other books written by the same author. This information then can be used to identify the author of the test text.

Written texts, for example novels, may differ in length. Also the same author may write in different styles during his life time. The same author may exhibit different styles when the type of the text changed. To overcome the length problem, we choose equal number of paragraphs from each text. Sometimes from a long novel we choose several smaller parts, and during the classification process we use these extra parts to get a majority vote decision.

When we try to classify several books of the same author, it is seen that mostly the genre of the book affects the style dramatically. A theatre scenario has more short sentences then an assay. Genre dependence of features may also be studied first. If a feature differs in the theatre scenarios, travelling notes, diaries, and novels, this feature cannot be used in classification. Another solution is to perform the classification in each genre separately.

In our research, we have selected thousands of paragraphs from "Na Drini Ćuprija", "Znakovi Pored Puta", "Prokleta Avlija" by Ivo Andrić, "Derviš i Smrt", "Tvrdjava" by M. Meša Selimović, and "Pobune" by Derviš Sušić.

Feature Selection

On lexical author identification research, one of the important issues is to decide about features that will serve as characteristics to distinguish the author of text. In this research, five textual descriptors are used: numbers of characters, words, sentences, commas and conjectures "and" (in Bosnian "i"), along with other paragraph characteristics. Means and variances of the textual descriptors for the texts of Ivo Andrić: "Na Drini Ćuprija", and M. Meša Selimović: "Derviš i Smrt" are shown in Table 1.

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

	Ivo Andrić: Na Drini Ćuprija		**M. Meša Selimović: Derviš**	
Textual descriptors	**Mean**	**Variance**	**Mean**	**Variance**
Sentence length	84.331	2090.92	58.710	2053.855
Word length	2.157	2.877	2.155	3.460
Word count	79.208	5861.724	60.362	4756.432
Sentence count	4.395	16.886	5.012	29.411
Comma count	6.432	45.95	7.130	87.211
dots count	0.052	0.135	0.002	0.002
i count	5.375	35.072	2.235	9.659
ili count	0.250	0.514	0.302	0.688
je count	2.798	11.991	2.552	11.531
se count	1.852	4.823	1.615	4.478
pa count	0.140	0.216	0.098	0.133
da count	1.935	6.853	2.262	9.613
ne count	0.637	1.695	0.968	2.718
kao poput count	0.662	1.106	0.480	1.007
Total		8080.760		6970.200

As can be seen, there is a statistical difference between the usage of textual descriptors - for instance, Ivo Andrić prefers longer paragraphs. On average, Ivo Andrić's paragraphs contain 79 words with a variance of 5861.7, while Meša Selimović's average is 62 with a variance of 4756.4. In the next chapter, the pattern captured by principal components will be displayed.

Methodology

The method of stylometric analysis for authorship attribution basically founded by Mosteller and Wallace in their 1964 work on the analysis of the twelve federalist papers of questionable authorship using the function words in those essays with known writings of John Jay, Alexander Hamilton, and James Madison (Mosteller, and Wallace 1964). Based on the evidence such as a higher usage of the word "upon", they concluded that the twelve unattributed papers were most likely authored by Madison, supporting the conclusion of the historians.). In the late 1980s and early 1990s a series of papers appeared by Burrows and Holms (Holmes, 1998; Burrows, 1992). Burrows method, is the method exploited in this paper. First the function words are counted from selected text, then this data is transferred into principal components world. Principal components method compresses data without losing its descriptive power. In the work of Holmes (Holmes 1998), the resulting data is not underwent a analytical investigation, rather they are plotted as cluster plots, as in this article. The patterns in these cluster plots are taken as author's writeprints.

Although the method is very simple to use, it is still used by some researchers. For Example Binongo, (Binongo, 2003) used this method to find the real author of the L. Frank Baum's last book: The Royal Book of Oz. Although Mrs. Baum explained that the book was based on "some unfinished notes" her husband had left three decades later, it is clarified through the PCA analysis performed by Binongo that the 15th book was, in reality, Ruth P. Thompson's own work. Thompson did not slavishly imitate what Baum had done, but instead built on his style. Similar methods are used in other works as well (Holmes and Forsyth, 1995; Holmes et al., 2001; Peng and Hentgartner, 2002).

Principal Components of Sample Texts

Next, random samples of 400 data are chosen from data sets for the textual descriptors for the texts authored by Ivo Andrić: "Na Drini Ćuprija", and M. Meša Selimović: "Derviš i Smrt", , and , and for other four books. These are all 400×14 matrices. Their covariance matrices are 14×14 matrices. The information in the covariance matrices is used to define a set of new variables , and as a linear combination of the original variables in the data matrices. The new variables are derived in a decreasing order of importance. The first column of is called first principal component and accounts for as much as possible of the variation in the original data. The second column is called the 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.

Results

Variances and percentage variances covered by fourteen principal components of the textual descriptors for the sample texts Ivo Andrić: "Na Drini Ćuprija" and M. Meša Selimović: "Derviš i Smrt" are shown in Table 2.

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

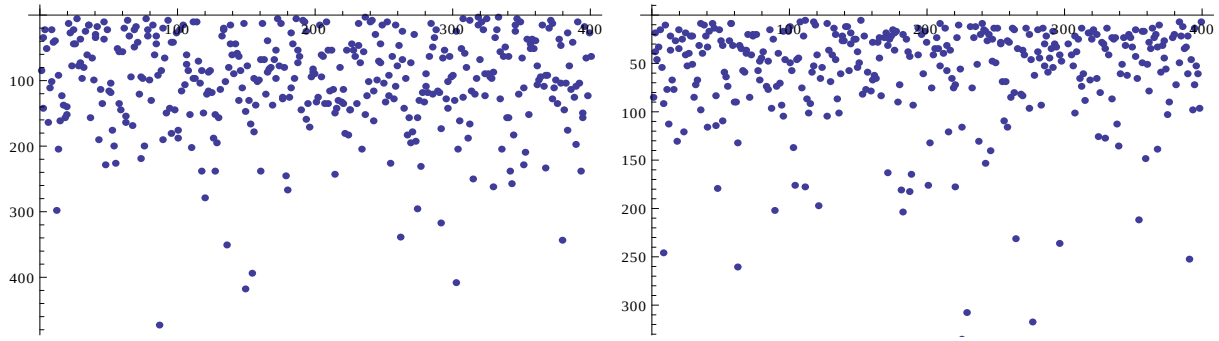
In Figure 1, one first principal component of each of samples from "Na Drini Ćuprija" and "Derviš i Smrt" data are displayed.

These figures are similar and do not seem to be used as writeprints of authors. It is the same for the second principal components. To search for a writerprint, we transform this information into the frequency domain. A common range for the contents of these two vectors is the interval. We divide this interval into 25 bins of equal length of 20, and count the numbers of entries of first component vectors in these bins. Figure 2 displays the data from Figure 1 in frequency domain.

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

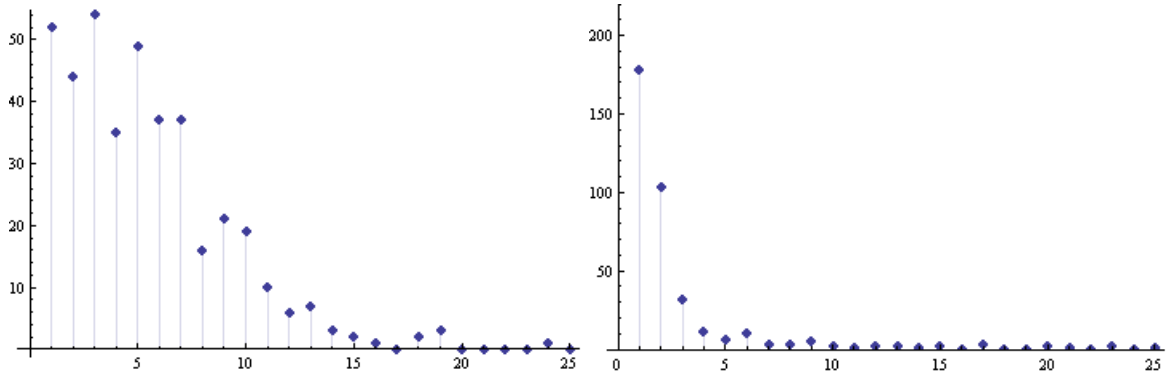
	Ivo Andrić: Na Drini Ćuprija		M. Meša Selimović: Derviš	
P. Comp.	**Variance**	**% Variance covered**	**Variance**	**% Variance covered**
1	7447.154220	75.6006325	5374.758432	77.1105536
2	2376.67024	24.1270381	1561.304509	22.3997146
3	8.130187	0.0825345	14.211600	0.2038909
4	5.310098	0.0539061	6.152396	0.0882672
5	3.199071	0.0324757	3.335845	0.0478587
6	2.811245	0.0285387	2.884413	0.0413821
7	2.152849	0.0218549	2.027011	0.0290811
8	1.569122	0.0159291	1.644081	0.0235873
9	1.345059	0.0136545	1.530064	0.0219515
10	0.830111	0.0084270	1.078908	0.0154789
11	0.777950	0.0078975	0.700177	0.0100453
12	0.477576	0.0048482	0.451615	0.0064792
13	0.148686	0.0015094	0.116703	0.0016743
14	0.074267	0.0007539	0.002465	0.0000354
	8080.76	100	6970.20	100

Figure 1
First principal components of samples from "Na Drini Ćuprija" (a) and "Derviš i Smrt" (b) data



(a) (b)

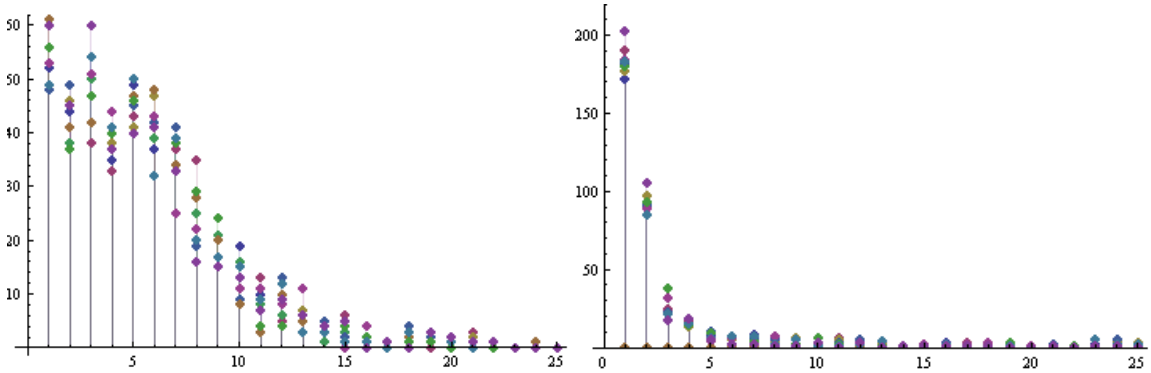
Figure 2
Frequencies of elements of first principal component vectors of random samples from "Na Drini Ćuprija" (a) and "Derviš i Smrt" (b) data in 25 bins.



(a) (b)

It is seen that the write prints of the two authors are distinguishable. To see whether the captured features remains similar through random samplings from data sets, we sketch together the frequencies of ten different samples in Figure 3.

Figure 3
Frequencies of elements of first principal component vectors of ten random samples from "Na Drini Ćuprija" (a) and "Derviš i Smrt" (b) data in 25 bins

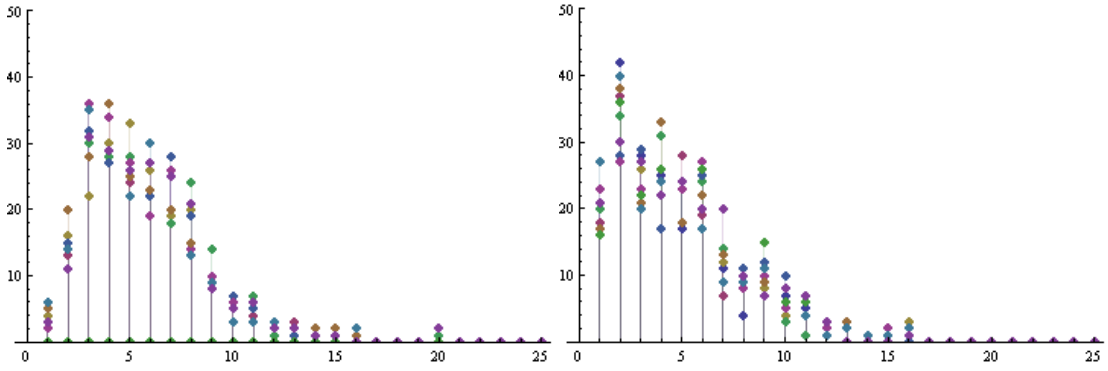


(a) (b)

To check whether these patterns are characteristic for other books of the two authors, two more books of Ivo Andrić; "Znakovi Pored Puta", and "Prokleta Avlija", and one other book of Meša Selimović; "Tvrdjava", as well as "Pobune" authored by a third novelist Derviš Sušić are investigated.

The comparison of the frequencies in the first principal components of the three books authored by Ivo Andrić: "Na Drini Ćuprija", "Znakovi Pored Puta", "Prokleta Avlija" are shown in Figure 4 below. The writing print of Ivo Andrić is the lower peaks – less than 70 – at the lowermost values of the principal components.

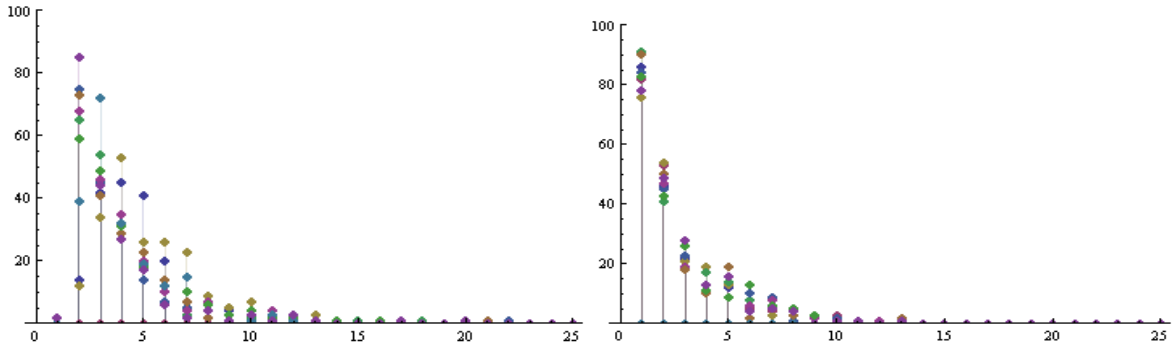
Figure 4
Frequencies of data in the first principal components of the two other books authored by Ivo Andrić: "Znakovi Pored Puta" (a), and "Prokleta Avlija" (b).



(a) (b)

The first principal components of the other book authored by Meša Selimović; "Tvrdjava" displayed in Figure 5a, a third author's text "Pobune" (Sušić, 1966) in Figure 5a. The writing print of Meša Selimović is revealed as twice higher peaks compared to the corresponding Ivo Andrić peaks, and differs significantly from pattern for Derviš Sušić.

Figure 5
Frequencies of data in the first principal components of the book authored by Meša Selimović; "Tvrdjava" (a), and third author's text "Pobune" (b).



Discussion and Conclusion

In this article we have proposed a method of authorship detection using principal component analysis for texts written in Bosnian language, which has proven as an efficient tool. The textual descriptors chosen proved to be successful for author identification. Although syntactic features are effective in classification, when we use test data from other books of the same author, we confronted by some difficulties. In Figure 4, the frequencies of data in the first principal components of the two books authored by Ivo Andrić: "Znakovi Pored Puta" (a), and "Prokleta Avlija" (b) is shown. The difference of the appearance of these charts shows that it is early to deduce that syntactic features are enough to make successful author identification. Therefore we concluded that for better classification results, features other then syntactic ones must also accompany. At the same time we must study the invariance of textual descriptors under translation into other languages.

Once a method for finding write prints is established, it is not difficult to deal with the author attribution problems, simply by the use of perceptrons of artificial neural networks. Indeed in a series of articles, the authors of this article, with a group of researchers at the International University of Sarajevo follow this path (Can et al., 2012; Savatić, Can and Jamak, 2012; Selman et al., 2011)

The method is applicable to other similar southern Slavic languages as well. We have tested the method against the texts of well-known novel writers. The achieved results show that every author leaves a unique signature in written text that can be discovered by analyzing counts of short words per paragraph. Future research should focus on the short words, and answer the following questions: "What short words are most significant for the authorship detection?" Furthermore the research should test the behavior of the method among larger amount of text samples.

References

1. Andrić, I. (1981). Na Drini Ćuprija, Svjetlost, Sarajevo.
2. Andrić, I. (1989). Znakovi Pored Puta, Svjetlost, Sarajevo.
3. Andrić, I. (1980). Prokleta Avlija, Svjetlost, Sarajevo.
4. Binongo, J. (2003), "Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution", Chance, Vol. 16, No. 2, pp. 9–17.
5. Bosch, R, Smith J. (1998), "Separating hyperplanes and the authorship of the disputed federalist papers", The American Mathematical Monthly, Vol. 105, No. 7, pp. 601–608.
6. Burrows, J. (1992), "Not unless you ask nicely: The interpretative nexus between analysis and information", Literary and Linguistic Computing, Vol. 7, No. 2, pp. 91–109.
7. Can, M, Jamak, A, Savatić, A. (2012), "Teaching Neural Networks to Detect the Authors of Texts Using Lexical Descriptors", Southeast Europe Journal of Soft Computing, Vol. 1, No. 1, pp. 57-72.
8. Chaski, C. (2001), "Empirical evaluations of language-based author identification techniques", Journal of Forensic Linguistics, Vol. 8, No. 1, pp. 1–65.
9. Chaski, C. (2005), "Who's at the keyboard? Authorship attribution in digital evidence investigations", International Journal of Digital Evidence, Vol. 4, No. 1, pp. 14.
10. Fung, G. (2003), "The disputed Federalist Papers: SVM feature selection using concave minimization", Proceedings of the 2003 Conference on Diversity in Computing, Tapia Companion,Atlanta,pp. 42–46.

11. Hayes, J. F. (2008), "Authorship Attribution: A Principal Component and Linear Discriminant Analysis of the Consistent Programmer Hypothesis", International Journal of Computers and Their Applications, Vol. 15, No. 2, pp. 79-99.
12. Holmes, D. (1998), "The evolution of stylometry in humanities scholarship", Literary and Linguistic Computing, Vol. 13, No. 3, pp. 111–117.
13. Holmes, D, Forsyth R. (1995), "The Federalist revisited: New directions in authorship attribution", Literary and Linguistic Computing, Vol. 10, No. 2, pp. 111–127.
14. Holmes, D, Gordon L, Wilson C. (2001), "A widow and her soldier: Stylometry and the American Civil War", Literary and Linguistic Computing, Vol. 16, No. 4, pp. 403–420.
15. Juola, P. (2006), "Authorship attribution", Foundations and Trends in Information Retrieval, Vol. 1, No. 3, pp. 233–334.
16. Juola, P, Sofko J, Brennan P. (2006), "A prototype for authorship attribution studies", Literary and Linguistic Computing, Vol. 21, No. 2, pp. 169–178.
17. Kjell, B. (1994), "Authorship determination using letter pair frequency features with neural network classifiers", Literary and Linguistic Computing, Vol. 9, No. 2, pp. 119–124.
18. Markov, A.A. (1916). Ob odnom primenenii statisticheskogo metoda (On some application of statistical method). In: Izvestia Akademii Nauk. (Russia). Ser.6, vol.X, N4, p.239 (in Russian).
19. Mendenhall, T. C. (1887). The characteristic curves of composition. Science, IX, 237–49.
20. Morozov, N.A. (1915). Lingvisticheskie spektry (Linguistic spectrums). In: Izvestia Akademii Nauk (Russia), (Section of Russian Language), Books 1-4, vol.XX, (in Russian).
21. Mosteller, F, Wallace, DL. (1964), Inference and Disputed Authorship: The Federalist, Addison Wesley, Reading, MA.
22. Peng, R, Hengartner N. (2002), "Quantitative analysis of literary styles", The American Statistician, Vol. 56, No. 3, pp. 175–185.
23. Savatić, A, Jamak A, Can M. (2012), "Detecting the Authors of Texts by Boosting Neural Network Committee Machines", Southeast Europe Journal of Soft Computing, Vol. 1, No. 1, pp. 81-92.
24. Selimović, M. M. (1966). Derviš i smrt, Svjetlost, Sarajevo.
25. Selimović, M. M. (1970). Tvrdjava, Svjetlost, Sarajevo.
26. Selman S, Turan K, Kuşakçı A. O. (2011), "Distingtion of the Authors of Texts Using Multilayered Feedforward Neural Networks", S. Europe Journal of Soft Computing, Vol. 1, No. 1, pp. 128-138.
27. Simpson, E. H. (1949) "Measurement of diversity". Nature 163, 688-688.
28. Sušić, D. (1966). Pobune, Veselin Masleša, Sarajevo.
29. Williams, C. (1975), "Mendenhall's studies of word-length distribution in the works of Shakespeare and Bacon", Biometrika, Vol. 62, No. 1, pp. 207–212.
30. Yule, G.U. (1944) The Statistical Study of Literary Vocabulary, Cambridge University Press.
31. Zipf, G. K. (1935) The Psychobiology of Language. Houghton-Mifflin.

About the authors

Amir Jamak is a PhD candidate at Faculty of Engineering and Natural Sciences, International University of Sarajevo. Main research interests are flexible computing, dynamically reconfigurable FPGA, spatial computing, EDA tools, ASIC, intelligent algorithms, data mining and authorship detection. Author can be contacted at ***amir.jamak@bhtelecom.ba***

Alen Savatić is a PhD candidate at Faculty of Engineering and Natural Sciences, International University of Sarajevo. Main research interests are E-learning, Computer Science, Computer Forensics, Social Network Analysis (SNA), and Artificial Neural Network. Author can be contacted at ***alen@savatic.net***

Mehmed Can is Professor at International University of Sarajevo, Faculty Engineering and Natural Sciences. Research interests are Integrability theory of differential equations, Lie group analysis of differential equations, Local existence of solutions of quasilinear parabolic and hyperbolic partial differential, Equations, Global nonexistence of solutions of quasilinear parabolic and hyperbolic equations, Painlevé analysis of differential equations, Lie-Backlund transformations and conservation laws of differential equations, Nonlinear Dynamical systems, bifurcations, fractals and chaos, Fuzzy sets, fuzzy linear programming and fuzzy decision making, and Mathematical modeling. Author can be contacted at ***mcan@ius.edu.ba***