Principal Component Analysis for Authorship Attribution

Amir Jamak, Alen Savatić, and Mehmet Can
International University of Sarajevo, Faculty of Engineering and Natural Sciences
Hrasnićka Cesta 15, 71000 Sarajevo, Bosnia and Herzegovina
amir.jamak@bhtelecom.ba , alen@savatic.net , mcan@ius.edu.ba

Abstract: A common problem in statistical pattern recognition is that of feature selection or feature extraction. Feature selection 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. 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. Then first of the principal components is used to distinguished the books authored by a certain author.

Keywords: principal components, authorship attribution, stylometry, text categorization, function words, classification task, stylistic features, syntactic characteristics.

1. INTRODUCTION

Of all text categorization problems, that of authorship attribution is probably the oldest; however, it is also possibly the least well organized, and its history is marred with the mishandling of statistical techniques. And yet, it still promises to provide useful applications in spheres as diverse as law, security, and education.

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, Mendenhall began investigating this hypothesis, searching for a characteristic difference in the distribution of different-sized words in writings of different languages and presentation styles. 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; it seems, with Morozov proposing a model for measuring style that garnered the interest of A. Markov (Kukurushkina et al. 2002). In the West, it took 30 years or so for Mendenhall's studies to be resumed by other linguists. George 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. G. Udny 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 et. al. 2006).

A breakthrough was needed, and it came in 1963 with Mosteller and Wallace's study on the Federalist Papers. In 1787 and 1788, John Jay, Alexander Hamilton and James 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 (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.

Despite the fact that Hamilton and Madison have otherwise very similar styles—nearly identical sentence length distributions, as noted by (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 (quoted in 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 (see, for instance, Kjell 1994, Holmes & Forsyth 1995, Bosch & Smith 1998, and Fung 2003).

2 PROBLEM DEFINITION

In this paper an application of principal component analysis is presented. The authorship attribution is

considered as a classification task (Chaski, C. 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). 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. Principal components of data elicited from texts possess generalization properties that allow for the required high accuracy of classification (Hayes 2008).

2.1 Texts Used

In research texts of two famous Bosnian writers, Ivo Andrić, M. Meša Selimović, and Derviš Sušić are used. Their novels 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. If the influence is significant, then lexical and syntactic features cannot be used as the writer invariant as unreliable.

Hence all together 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.

2.2 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 five textual descriptors are used, numbers of characters, words, sentences, commas, and conjecture “and”, in Bosnian “i”, and other characteristics in paragraphs. Means and variances of the textual descriptors for the texts 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
Sentencelength	84.33	2090.92	58.71	2053.85
Wordlength	2.157	2.877	2.155	3.460
Word count	79.20	5861.724	60.362	4756.432
Sentencecount	4.395	16.886	5.012	29.411
Commacount	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	0.662	1.106	0.480	1.007
Total		8080.760		6970.200

As it is seen, there is statistical difference between the usage of textual descriptors, 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.

In the next chapter the pattern captured by principal components will be displayed.

3 PRINCIPAL COMPONENT ANALYSIS

The methods of Mosteller and Wallace have proved as enduring as the problem they investigated: they were only modestly altered when Burrows described his method of stylometric analysis in a series of papers published in the late 1980s and early 1990s (Holmes 1998; see, for instance, Burrows 1992). The Burrows method essentially involves computing the frequency of each of a list of function words (larger than that of Mosteller and Wallace), 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 (a two-dimensional plot of the first two principal components) and inspected visually for trends, which occur as clusters of points (Holmes 1998). (Later, cluster analysis would accomplish 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 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 (Binongo 2003). 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 (Holmes & Forsyth 1995, Holmes et al. 2001, and Peng & Hentgartner 2002).

4. 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, X_{ivo}, and X_{mesa}, and for other four books. These are all 400×14 matrices. Their covariance matrices C_{ivo}, C_{mesa} are 14×14 matrices. The information in the covariance matrices are used to define a set of new variables $P_{ivo} = X_{ivo}.C_{ivo}$, and $P_{mesa} = X_{mesa}.C_{mesa}$ 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 $P_{(\,)}$ is called first principal component and accounts for as much much as possible of the variation in the original data. The second column 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.

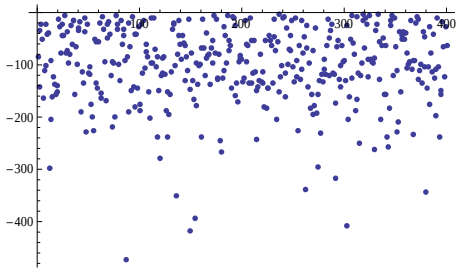
Variances and percentage variances covered by fourteen principal components of the textual descriptors for the sample texts Ivo Andrić: Na Drini Ćuprija, and M. Meša Selimović: Derviš i Smrt are shown in Table 2.

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

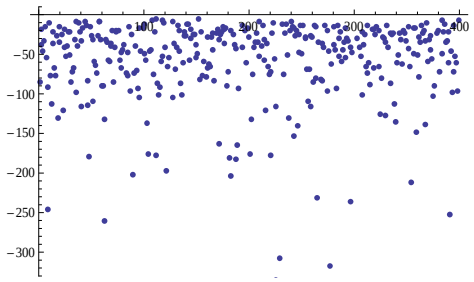
	Ivo Andrić: Na Drini Ćuprija	M. Meša Selimović: Derviš		
P. Comp.	Variance	%variance covered	Variance	%variance covered
1	7447.154	75.60063	5374.758	77.11055
2	2376.670	24.12703	1561.304	22.39971
3	8.130187	0.082534	14.21160	0.203890
4	5.310098	0.053906	6.152396	0.088267
5	3.199071	0.032475	3.335845	0.047858
6	2.811245	0.028538	2.884413	0.041382
7	2.152849	0.021854	2.027011	0.029081
8	1.569122	0.015929	1.644081	0.023587
9	1.345059	0.013654	1.530064	0.021951
10	0.830111	0.008427	1.078908	0.015479
11	0.777950	0.007897	0.700177	0.010045
12	0.477576	0.004848	0.451615	0.006479
13	0.148686	0.001509	0.116703	0.001674
14	0.074267	0.000753	0.002465	0.000035
	8080.76	100	6970.20	100

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

In Figure first principal component of each of samples from Cuprija na Drina and Derviš i Smrt data are displayed.



(a)



(b)

Figure 1. First principal components of samples from Cuprija na Drina (a) and Derviš i Smrt (b) data.

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.

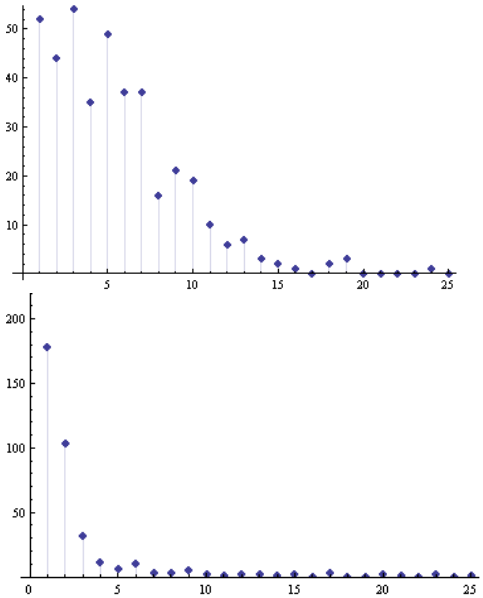


Figure 2. Frequencies of elements of first principal component vectors of random samples from Cuprija na Drina (a) and Derviš i Smrt (b) data in 25 bins.

A common range for the contents of these two vectors is the interval $[-500, 0]$. We divide this interval into 25 bins of equal length of 20

$$\{[-500, -480), [-480, -460), \ldots, [-40, -20), [-20, 0]\}$$

and count the numbers of entries of first component vectors in these bins. Figure 2 displays the data in Figure 1 in frequency domain.

It is seen that the writeprints 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.

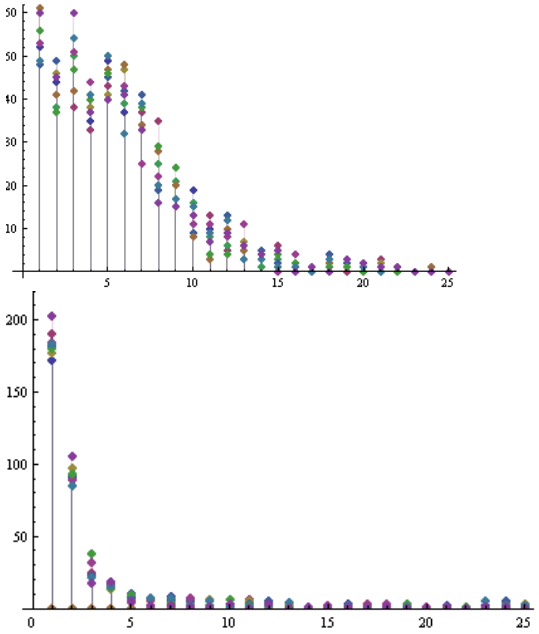
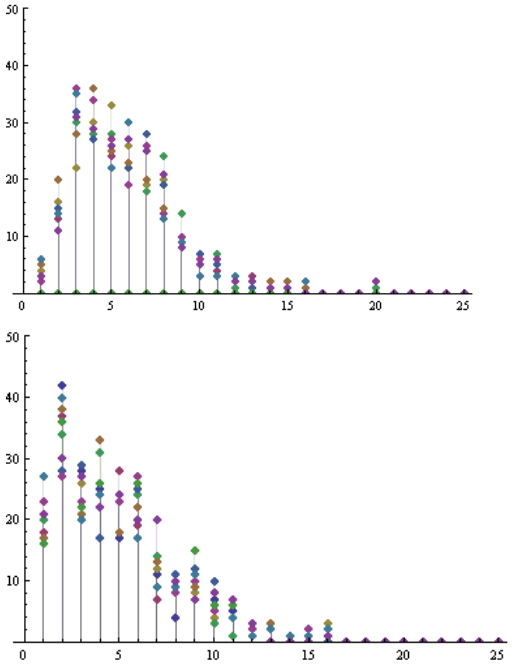


Figure 3. Frequencies of elements of first principal component vectors of ten random samples from Cuprija na Drina (a) and Derviš i Smrt (b) data in 25 bins.

To check whether these patterns are characteristic for other books of the two authors, two more books of Ivo Andrić; Znakovi Pored Puta, and Proklet Avlija, and one other book of Meša Selimović; Tvrdjeva, 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: Cuprija na Drina, Znakovi Pored Puta, Proklet Avlija are shown in Figure 3 below. The writing print of Ivo Andrić is the lower peaks – less than 70 – at the lowermost values of the principal components.

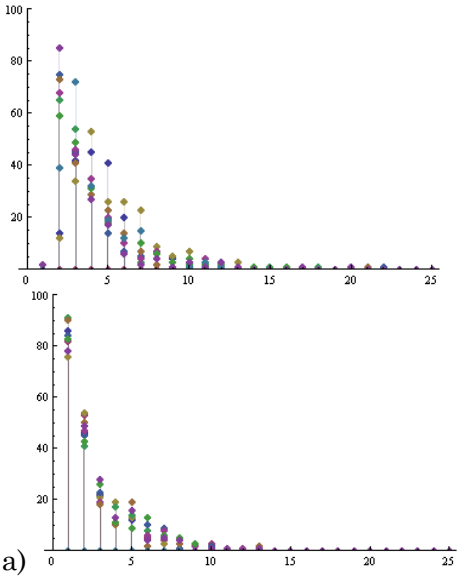


a)

b)

Figure 3. Frequencies of data in the first principal components of the two other books authored by Ivo Andrić: Znakovi Pored Puta a), and Proklet Avlija b).

The first principal components of the another book authored by Meša Selimović; Tvrdjeva displayed in Figure 4a, a third author's text Pobune (Sušić 1966) in Figure 4a. 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.



a)

b)

Figure 4. Frequencies of data in the first principal components of the book authored by Meša Selimović; Tvrdjeva a), and third author's text Pobune b).

6. CONCLUSIONS

The research described in this paper concerning author identification analysis shows that the method of principal component analysis (PCA) is an efficient a tool. Yet conclusions as to the choice of textual descriptors used as features for recognition process, based only on results presented in the previous sections 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 artificial neural networks -based methodology to 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.

Once a method for finding write prints, 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 researcher at the International University of Sarajevo follow this path (Can, Jamak, Savatić 2011, Savatić, Can, Jamak 2011, Can, Hadžiabdić, Demir 2011, Selman, Turan, Kuşakçı, 2011)

REFERENCES

Andrić, I. 1981. Na Drini Ćuprija, Svjetlost, Sarajevo.

Andrić, I. 1989. Znakovi Pored Puta,, Svjetlost, Sarajevo.

Andrić,I. 1980. Prokleta Avlija, Svjetlost, Sarajevo.

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

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

Burrows, J. 1992. Not unless you ask nicely: The interpretative nexus between analysis and information. Literary and Linguistic Computing 7(2): 91–109.

Can, M., A. Jamak, and A. Savatić. 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., K.K. Hadžiabdić, N. M. Demir. 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).

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

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

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

Holmes, D., and R. Forsyth. 1995. The Federalist revisited: New directions in authorship attribution. Literary and Linguistic Computing 10(2): 112–27.

Holmes, D., L. Gordon, and C. Wilson. 2001. A widow and her soldier: Stylometry and the American Civil War. Literary and Linguistic Computing 16(4): 403–20.

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

Juola, P., J. Sofko, and P. Brennan. 2006. A prototype for authorship attribution studies. Literary and Linguistic Computing 21: 169–78.

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

Kolman, B., and D. R. Hill. 2004. Elementary Linear Algebra, Pearson, New Jersey.

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

Peng, R., and N. Hengartner. 2002. Quantitative analysis of literary styles. The American Statistician 56(3): 175–85.

Savatić, A., A. Jamak, and M.Can. 2011. Detecting the Authors of Texts by Boosting Neural Network Committee Machines, Proceedings of the 2nd International Scientific and professional Conference of Graphic Technology and Design, 9-11 June 2011, Kselajak, BiH, pp. 223-232.

Selimović, M. M.1966 , Derviš i smrt, Svjetlost, Sarajevo.

Selimović, M. M. 1970. Tvrdjava, Svjetlost, Sarajevo.

Selman S., K. Turan, and A. O. Kuşakçı. 2011. Distingtion of the Authors of Texts Using Multilayered Feedforward Neural Networks, 9th International Conference on Knowledge, Economy & Management Proceedings, pp. 1419-1429.

Sušić, D. 1966. Pobune, Veselin Masleša, Sarajevo.

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