Distinction of The Authors of Texts Using Multilayered Feedforward Neural Networks

Suvad Selman[1]
Kemal Turan[2]
Ali Osman Kuşakçi[3]
International University of Sarajevo
Faculty of Engineering and Natural Sciences
Hrasnicka Cesta 15, 71000 Sarajevo
Bosnia and Herzegovina

Abstract

This paper proposes a means of using a multilayered feedforward neural network to identify the author of a text. The network has to be trained where multilayer feedforward neural network as a powerful scheme for learning complex input-output mapping have been used in learning of the average number of words and average characters of words in a paragraphs of an author. The resulting training information we get will be used to identify the texts written by authors. The computational complexity is solved by dividing it into a number of computationally simple tasks where the input space is divided into a set of subspaces and then combining the solutions to those tasks. By this, we have been able to successfully distinguish the books authored by Leo Tolstoy, from the ones authored by George Orwell and Boris Pasternak.

Keywords—Machine learning, author identification, artificial neural networks

[1] International University of Sarajevo, Faculty of Engineering and Natural Sciences, Electrical and Electronics Program
[2] International University of Sarajevo, Faculty of Engineering and Natural Sciences, Mechanical Engineering Program
[3] International University of Sarajevo, Faculty of Engineering and Natural Sciences, Industrial Engineering Program

INTRODUCTION

Individuals have distinctive ways of speaking and writing, and there exists a long history of linguistic and stylistic investigation into author identification. In recent years, practical applications for author identification have grown in areas such as intelligence, criminal law, civil law, and computer security. This activity is part of a broader growth within computer science of identification technologies, including, cryptographic signatures, intrusion detection systems, and others. Automating author identification [2, 3] promises more accurate results and objective measures of reliability, both of which are critical for legal and security applications. Recent research has used techniques from machine learning [4, 5] and natural language processing author identification.

Author identification is the task of identifying the author of a given text. It can be considered as a typical classification problem, where a set of documents with known authors are used for training and the aim is to automatically determine the corresponding author of an anonymous text. In contrast to other classification tasks, it is not clear which features of a text should be used to classify an author. Consequently, the main concern of computer-based author identification is to define an appropriate characterization of documents that captures the writing style [5, 6] of authors.

Author identification has a long history that includes some famous disputed authorship cases and also has forensic applications. The advent of non-traditional author identification techniques can be traced back to 1887, when Mendenhall [10] first created the idea of counting features such as word length. His work was followed by work from Yule and Morton [7] with the use of sentence lengths to judge authorship. Brainerd [7] concentrated on syllables per word. Moreover, Holmes [7] developed a function to relate the frequency of used words and the text length. Karlgren-Cutting [15] figured out the style marker of the text. Biber [9, 11] added the syntactic and lexical style markers. In the recent improvements on author identification we can see Kessler [3], who developed a fairly simple and reliable method. Twedie and Baayen [10] showed that the proportion of the different word count to the total word count could be a fair measurement and the results for the texts which are shorter than 1000 word in length could be inconsistent. Burrows [13] used principal components analysis (PCA) to find combinations of style markers that can discriminate between a particular pair (or small set) of authors. Another related class of techniques that have been applied are machine learning algorithms categorization and other stylistic discrimination tasks. Often, studies have relied on intuitive evaluation of results, based on visual inspection of scatter plots and cluster analysis trees, though recent work has begun to apply somewhat more rigorous tests of statistical significance and cross validation accuracy. Other stylometric features that have been applied include various measures of vocabulary richness and lexical repetition, based on Zipf's [18] studies on word frequency distributions. Most such measures, however, are strongly dependent on the length of the text being studied, and so are difficult to apply reliably. Many other types of features have been applied, including word class frequencies, syntactic analysis, word collocations, grammatical errors, and word, sentence, clause, and paragraph lengths. Many studies combine features of different types using multivariate analysis techniques.

Author identification can be used in a broad range of applications, to analyze anonymous or disputed documents/books. In Plagiarism detection which can be used to establish whether claimed authorship is valid. In criminal investigation as Ted Kaczynski [19] was targeted as a primary suspect in the Unabomber case, because author identification methods determined that he could have written the Unabomber's manifesto. In forensic investigations where verifying the authorship of e-mails and newsgroup messages, or identifying the source of a piece of intelligence.

In this paper an application to artificial neural networks is presented to authorship attribution is considered as a classification task [5]. Texts studied are literary works of worldwide known writers, Leo Tolstoy, George Orwell and Boris Pasternak. 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. Properly trained neural networks possess generalisation properties that allow for the required high accuracy of classification.

1. OBJECTIVES

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

Analysts claim that each writer possesses some unique characteristic, called the authorial or writer invariant, which keeps constant for all texts written by this author and perceivably different for texts of other authors [20]. To find writer invariants there are used style markers which are based on textual properties belonging to either of four categories: lexical, syntactic, structural, and content-specific.

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

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

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

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

3 NEURAL NETWORKS

There are a number of different answers possible to the question of how to define

neural networks. At one extreme, the answer could be that neural networks are simply a class of mathematical algorithms, since a network can be regarded essentially as a graphic notation for a large class of algorithms. Such algorithms produce solutions to a number of specific problems. At the other end, the reply may be that these are synthetic networks that emulate the biological neural networks found in living organisms [21]. In light of today's limited knowledge of biological neural networks and organisms, the more plausible answer seems to be closer to the algorithmic one.

In search of better solutions for engineering and computing tasks, many avenues have been pursued. There has been a long history of interest in the biological sciences on the part of engineers, mathematicians, and physicists endeavouring to gain new ideas, inspirations, and designs. Artificial neural networks have undoubtedly been biologically inspired, but the close correspondence between them and real neural systems is still rather weak [22]. Vast discrepancies exist between both the architectures and capabilities of artificial and natural neural networks. Knowledge about actual brain functions are so limited, however, that there is little to guide those who would try to emulate them. No models have been successful in duplicating the performance of the human brain. Therefore, the brain has been and still is only a metaphor for a wide variety of neural network configurations that have been developed [19].

3.1 TOPOLOGY

From topology point of view neural networks can be divided into two categories: feed-forward and recurrent networks. In feed-forward networks the flow of data is strictly from input to output cells that can be grouped into layers but no feedback interconnections can exist. On the other hand, recurrent networks contain feedback loops and their dynamical properties are very important. The most popularly used type of neural networks employed in pattern classification tasks is the feedforward network which is constructed from layers and possesses unidirectional weighted connections between neurons [19]. The common examples of this category are Multilayer Perceptron or Radial Basis Function networks, out of which the former will be addressed in more detail.

Multilayer Perceptron (MLP) type is more closely defined by establishing the number of neurons from which it is built, and this process can be divided into three parts, the two of which, finding the number of input and output units, are quite simple, whereas the third, specification of the number of hidden neurons can become crucial to accuracy of obtained classification results [25].

The number of input and output neurons can be actually seen as external specification of the network and these parameters are rather found in a task specification. For classification purposes as many distinct features are defined for objects which are analysed that many input nodes are required. The only way to better adapt the network to the problem is in consideration of chosen data types for each of selected features. For example instead of using the absolute value of some feature for each sample it can be more advantageous to calculate its change as this relative value should be smaller than the whole range of possible values and thus variations could be more easily picked up by Artificial Neural Network [26]. The number of network outputs typically reflects the number of classification classes.

The third factor in specification of the Multilayer Perceptron is the number of

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

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

Apart from the number of layers there is another issue of the number of neurons in these layers. When the number of neurons is unnecessarily high the network easily learns but poorly generalises on new data. This situation reminds autoassociative property: too many neurons keep too much information about training set rather "remembering" than "learning" its characteristics. This is not enough to ensure good generalization that is needed [27].

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

3.2 ACTIVATION FUNCTIONS

All neural networks take numeric input and produce numeric output. The transfer function of a unit is typically chosen so that it can accept input in any range, and produces output in a strictly limited range (it has a squashing effect). Although the input can be in any range, there is a saturation effect so that the unit is only sensitive to inputs within a fairly limited range. The illustration below shows one of the most common transfer functions, the logistic function (also sometimes referred to as the sigmoid function, although strictly speaking it is only one example of a sigmoid - S-shaped - function). In this case, the output is in the range (0, 1), and the input is sensitive in a range not much larger than (-1, +1). The function is also smooth and easily differentiable, facts that are critical in allowing the network training algorithms to operate (this is the reason why the step function is not used in practice) [28].

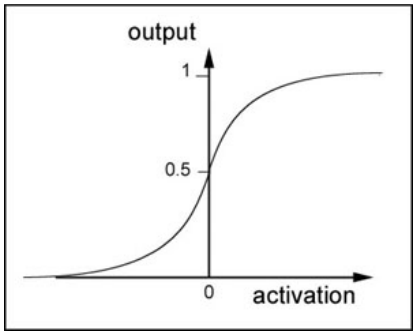

Fig. 1. Binary sigmoid activation function

The limited numeric response range, together with the fact that information has to be in numeric form, implies that neural solutions require pre-processing and post-processing stages to be used in real applications [19].

3.3 LEARNING ALGORITHMS

In order to produce the desired set of output states whenever a set of inputs is presented to a neural network it has to be configured by setting the strengths of the interconnections and this step corresponds to the network learning

procedure [31]. Learning rules are roughly divided into three categories of supervised, unsupervised and reinforcement learning methods.

In supervised learning, we are given a set of example pairs and the aim is to find a function in the allowed class of functions that matches the examples. In other words, we wish to infer the mapping implied by the data; the cost function is related to the mismatch between our mapping and the data and it implicitly contains prior knowledge about the problem domain. A commonly used cost is the mean-squared error, which tries to minimize the average squared error between the network's output, *f(x)*, and the target value y over all the example pairs [29]. When one tries to minimize this cost using gradient descent for the class of neural networks called multilayer perceptrons, one obtains the common and well-known backpropagation algorithm for training neural networks [33]. Tasks that fall within the paradigm of supervised learning are pattern recognition (also known as classification) and regression (also known as function approximation). The supervised learning paradigm is also applicable to sequential data (e.g., for speech and gesture recognition). This can be thought of as learning with a "teacher", in the form of a function that provides continuous feedback on the quality of solutions obtained thus far.

In unsupervised learning, some data is given and the cost function to be minimized, that can be any function of the data and the network's output. The cost function is dependent on the task (what we are trying to model) and our a priori assumptions (the implicit properties of our model, its parameters and the observed variables). Tasks that fall within the paradigm of unsupervised learning are in general estimation problems; the applications include clustering, the estimation of statistical distributions, compression and filtering [30].

In reinforcement learning, data are usually not given, but generated by an agent's interactions with the environment. At each point in time, the agent performs an action and the environment generates an observation and an instantaneous cost, according to some (usually unknown) dynamics. The aim is to discover a policy for selecting actions that minimizes some measure of a long-term cost; i.e., the expected cumulative cost. The environment's dynamics and the long-term cost for each policy are usually unknown, but can be estimated [31].

In reinforcement learning, data are usually not given, but generated by an agent's interactions with the environment. At each point in time, the agent performs an action and the environment generates an observation and an instantaneous cost, according to some (usually unknown) dynamics. The aim is to discover a policy for selecting actions that minimizes some measure of a long-term cost; i.e., the expected cumulative cost. The environment's dynamics and the long-term cost for each policy are usually unknown, but can be estimated [34].

4 DATASET DESCRIPTION

In research there were used texts of famous writers, Leo Tolstoy, George Orwell and Boris Pasternak. Their novels provide the corpus which is wide enough to make sure that characteristic features found based on the training data can be treated as representative of other texts and this generalized knowledge can be used to confirm or discount the possibility of either of considered writers being recognised as the author of a text of unknown origin.

Obviously literary texts can greatly vary in length and 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 [32].

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. On the other hand, this can be rectified by including within the training data set fragments of texts being representatives of not only one but several genres. For intended implementation of the classifier with Artificial Neural Networks, which efficiently deal with large amount of data, adding samples to the training set simply means better coverage of the input space that is important in continuous case [35].

Hence in the training set there were included samples coming from "War and Peace"[40] by Leo Tolstoy, "Animal Farm"[39] by George Orwell and "Doctor Zivago"[41] by Boris Pasternak.

4.1 FEATURES EXTRACTION

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

In the research only two textual descriptors are used, number of words, and average length of words in paragraphs.

Words and characters in 200 paragraphs from each book are counted. The descriptive statistics for these two textual descriptors are as follows:

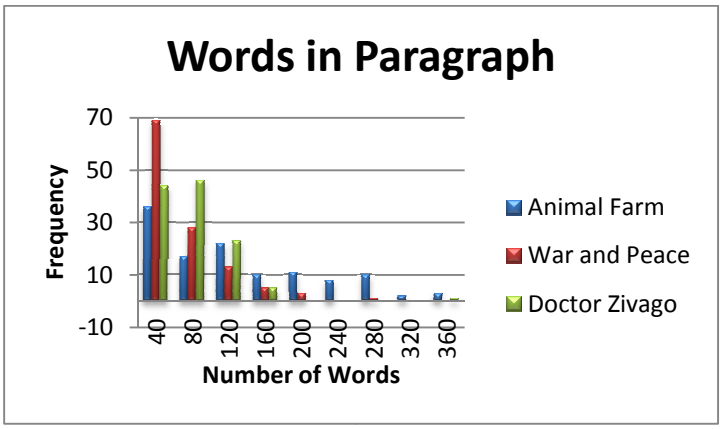

Figure 1: Frequencies of average number of words in paragraphs

As it is seen, Boris Pasternak prefers longer paragraphs. In average Orwell's paragraphs contain 111 words with standard deviation 86.5, Leo's paragraphs in average contains 51 words with standard deviation of 41 and Pasternak's paragraphs are long 58 words in average while having standard deviation of 44.

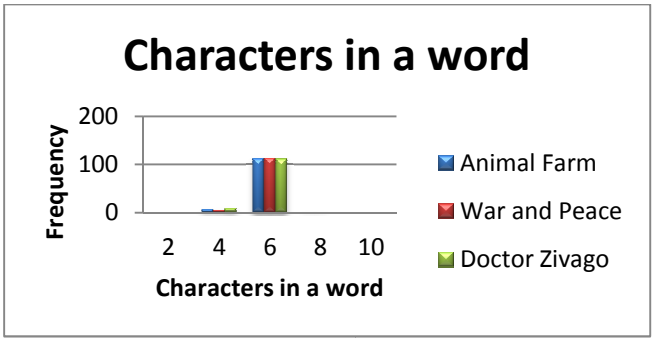

Figure 2: Frequencies of the average number of characters in words

The average characters in words are the second textual descriptor we are interested in. In this aspect, averages do not change much between authors. Orwell's words per paragraph are in average 4.56 characters long with

standard deviation 0.38. On the other hand Leo's words in a paragraph are containing 4.75 characters in average with standard deviation of 0.51 and these figures for Pasternak are 4.65 and 0.43 respectively.

5. RESULTS AND DISCUSSION

For validation purposes samples are used from some other works of all writers that were available, consisting of other parts of the same novels used previously during training and from different novels. Equal numbers of paragraphs are chosen from the mentioned works of authors.

As lexical descriptors, the number of words in paragraphs and paragraph average of characters in words are chosen.

Set 1 of data consists of lexical descriptors from 50 paragraphs chosen from two novels, one from Leo and the other one from Orwell. N1=100 is the number of data to train the neural network which has two input terminals, three hidden neurons in one hidden layer. The results of classification performed at the end of training by this network machine are given in the Table 1 below.

	Data Number	Correct Classification
Leo	50	42
Orwell	50	45

Table 1 Classification results for lexical descriptors for the first combination of two authors

Then another set, N2 descriptors sent to the same machine. Set 2 of data consists of lexical descriptors from 50 paragraphs chosen from two novels, one from Leo and the other one from Pasternak. N2=100 is the number of data to train the neural network which has two input terminals, three hidden neurons in one hidden layer. The results of classification performed at the end of training by this network machine are given in the Table 2 below.

	Data Number	Correct Classification
Leo	50	42
Pasternak	50	40

Table 2 Classification results for lexical descriptors for the second combination of two authors

Then another set of N3 descriptors sent to the same machine. Set 3 of data consists of lexical descriptors from 50 paragraphs chosen from two novels, one from Orwell and the other one from Pasternak. N3=100 is the number of data to train the neural network which has two input terminals, three hidden neurons in one hidden layer. The results of classification performed at the end of training by this network machine are given in the Table 3 below.

	Data Number	Correct Classification
Orwell	50	43
Pasternak	50	42

Table 3 Classification results for lexical descriptors for the third combination of two authors

As it is seen from Tables, the success is satisfactory; paragraphs authored by Leo Tolstoy are correctly identified in more than 80% in both trials. Same could be said for the other two authors, their correct classification is always more than 80% while in one instance George Orwell's paragraphs are correctly classified 90%. Overall correct classification probability is high enough.

6. CONCLUSIONS

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

It seems to interesting to replace neural network with support vector machines in the system. It would be also important to automatize the process of features selection on the multistage of the system. Estimating length of the Cosine Representation depending on recognized symbol is also a good direction for future work. Finally, it should fine to find other recognition fields, where multistage recognition systems can be used.

References

[1] A. Genkin, D. D. Lewis, and D. Madigan, Large-scale bayesian logistic regression for text

categorization, 2004.

[2] B.Diri, M. F. Amasyalı, Automatic Author Detection for Turkish Text, ICANN/ICONIP'03 13th International Conference on Artificial Neural Network and 10th International Conference on Neural Information Processing, 2003.

[3] B.Kessler, G. Nunberg, H.Schutze, Automatic Detection of Text Genre, Proc. of 35th Annual

Meeting of the Association for Computational Linguistics (ACL/EACL'97), 32-38 1997.

[4] Chris Callison-Burch, Co-training for Statistical Machine Translation, Master's thesis, University of Edinburgh, 2002.

[5] Christopher D. Manning and Hinrich Schutze, Foundations of Statistical Natural Language

Processing, The MIT Press, 1999.

[6] D. Biber, Variations Across Speech and Writing, Cambridge University Press, 1988.

[7] D. I. Holmes, Stylometry: Its Origins, Development and Aspirations, presented to the Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing, Queen's University, Kingston, Ontario, 1997.

[8] D. Khmelev, Disputed authorship resolution using relative entropy for markov chain of letters in a text, In R. Baayen, editor, 4th Conference Int. Quantitative Linguistics Association, Prague, 2000.

[9] E. Stamatatos, N. Fakotakis, G. Kokkinakis, Automatic Text Categorization in Terms of Genre and Author, Computational Linguistics, pages 471-495, 2000.

[10] F. J. Tweedie, S. Singh, D. I. Holmes, Neural Network Applications in Stylometry: The Federalist Paper, Computers and the Humanities, Vol. 30, pages 1-10, 1996.

[11] H. Baayen, H. van Halteren, and F. Tweedie, Outside the cave of shadows:

Using syntactic annotation to enhance authorship attribution. Literary and Linguistic Computing, Vol.11(3), pages 121-131, 1996.

[12] J. Allen, Natural Language Understanding, Benjamin/Cummings Pub. Co., Redwood City,

California, 1995.

[13] J. Burrows, Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method, Clarendon Press, Oxford, 1987.

[14] J. Goldsmith, Unsupervised learning of the morphology of a natural language, Computational Linguistics, Vol. 27(2), pages 153–198, 2001.

[15] J. Karlgren, and D. Cutting, Recognizing Text Genres with Simple Metrics using Discriminant Analysis, Proceedings of the 15th. International Conference on Computational Linguistics, Kyoto, 1994.

[16] Jill M. Farringdon, Analyzing for Authorship: A Guide to the Cusum Technique. University of Wales Press, 1996.

[17] J. Rissanen, Stochastic Complexity in Statistical Inquiry, Volume 15. World Scientific Series in Computer Science, Singapore, 1989.

[18] Mathias Creutz, Unsupervised segmentation of words using prior distributions of morph length and frequency, In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL), pages 280–287, Sapporo, Japan, 2003.

[19] R. A. Bosch, J. A. Smith, Separating Hyperplanes and the Authorship of the Disputed

[20] Durbin, R. 1989. "On the Correspondence Between Network Models and the Nervous System," in The Computing Neuron, ed. R. Durbin, C. Miall, G.Mitchison, Reading, Mass.: Addison-Wesley Publishing Co.

[21] Bishop, C. (1995). Neural Networks for Pattern Recognition. Oxford: University Press.

[22] D. Khmelev, Disputed authorship resolution using relative entropy for markov chain of letters in a text, In R. Baayen, editor, 4th Conference Int. Quantitative Linguistics Association, Prague, 2000.

[23] E. Stamatatos, N. Fakotakis, G. Kokkinakis, Automatic Text Categorization in Terms of Genre and Author, Computational Linguistics, pages 471-495, 2000.

[24] Chris Callison-Burch, Co-training for Statistical Machine Translation, Master's thesis, University of Edinburgh, 2002.

[25] Christopher D. Manning and Hinrich Schutze, Foundations of Statistical Natural Language Processing, The MIT Press, 1999.

[26] F. J. Tweedie, S. Singh, D. I. Holmes, Neural Network Applications in Stylometry: The Federalist

[27] Paper, Computers and the Humanities, Vol. 30, pages 1-10, 1996.

[28] H. Baayen, H. van Halteren, and F. Tweedie, Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. Literary and Linguistic Computing, Vol. 11(3), pages 121-131, 1996.

[29] J. Allen, Natural Language Understanding, Benjamin/Cummings Pub. Co., Redwood City, California, 1995.

[30] J. Burrows, Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method, Clarendon Press, Oxford, 1987.

[31] J. Goldsmith, Unsupervised learning of the morphology of a natural language,

Computational Linguistics, Vol. 27(2), pages 153–198, 2001.

[32] J. Karlgren, and D. Cutting, Recognizing Text Genres with Simple Metrics using Discriminant Analysis, Proceedings of the 15th. International Conference on Computational Linguistics, Kyoto, 1994.

[33] Jill M. Farringdon, Analyzing for Authorship: A Guide to the Cusum Technique. University of Wales Press, 1996.

[34] J. Rissanen, Stochastic Complexity in Statistical Inquiry, Volume 15. World Scientific Series in Computer Science, Singapore, 1989.

[35] Mathias Creutz, Unsupervised segmentation of words using prior distributions of morph length and frequency, In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL), pages 280–287, Sapporo, Japan, 2003.

[36] R. A. Bosch, J. A. Smith, Separating Hyperplanes and the Authorship of the Disputed Federalist Papers, American Mathematical Monthly, Volume 105, pages 601-608, 1998.

[37] S. Argamon-Engelson, M. Koppel, and G. Avneri, Style-based text categorization: What newspaper am I reading?, In Proc. AAAI Workshop on Learning for Text Categorization, pages 1-4, 1998.

[38] T. Mendenhall, The characteristic curves of composition, Science, 214:237249, 1887.

[39] George Orwell, Animal Farm

[40] Leo Tolstoy, War and Peace

[41] Boris Pasternak, Doctor Zivago