Using Machine Learning Techniques for Stylometry

Ramyaa, Congzhou He, Khaled Rasheed
Artificial Intelligence Center
The University of Georgia, Athens, GA 30602 USA

Abstract

In this paper we describe our work which attempts to recognize different authors based on their style of writing (without help from genre or period). Fraud detection, email classification, deciding the authorship of famous documents like *the Federalist Papers*[1], attributing authors to pieces of texts in collaborative writing, and software forensics are some of the many uses of author attribution. In this project, we train decision trees and neural networks to "learn" the writing style of five Victorian authors and distinguish between them based on certain features of their writing which define their style. The texts chosen were of the same genre and of the same period to ensure that the success of the learners would entail that texts can be classified on the style or the "textual fingerprint" of authors alone. We achieved 82.4% accuracy on the test set using decision trees and 88.2% accuracy on the test set using neural networks.

1. Introduction

Stylometry – the measure of style – is a burgeoning interdisciplinary research area that integrates literary stylistics, statistics and computer science in the study of the "style" or the "feel" (of a document). The style of a document is typically based on a lot of parameters – its genre or the topic (which is used in text categorization), its content, its authors - to name a few. Stylometry assumes – in the context of author attribution - that there is an unconscious aspect to an author's style that cannot be consciously manipulated but which possesses quantifiable and distinctive features. These characteristic features of an author should be salient, frequent, easily quantifiable and relatively immune to conscious control. In addition, these features need to be able to distinguish authors writing in the same genre, similar topics and periods so that the classifier is not helped by differences in genre or English style which changes with time.

The origins of stylometry can be traced back to the mid 19th century, where the English logician Augustus de Morgan suggested word length could be an indicator of authorship. The real impact did not come until 1964, when two American statisticians Mosteller and Wallace decided to use word frequencies to investigate the mystery of the authorship of *The Federalist Papers*.[1] Their conclusion from statistical discrimination methods agreed with historical scholarship which gave stylometry much needed credence [7]. Today with the help of many other significant events in stylometry, its techniques are being widely applied in various areas, such as, disease detection and court trials. Examples of using stylometric techniques as a forensic tool include the appeal in London of Tommy McCrossen in July 1991, the pardon for Nicky Kelly from the Irish government in April 1992, along with others [7].

1.1 Features

There is no consensus in the discipline as to what characteristic features to use or what methodology or techniques to apply in standard research, which is precisely the greatest problem confronting stylometry. Most of the experiments in stylometry are directed to different authors with different techniques; there has not been a comparison of results on a large scale as to what features are generally more representative or what methods are typically more effective. [11] claims that there is no clear agreement on which style markers are significant. Rudman (1998) [11] notes that particular words may be used for a specific classification (like *The Federalist Papers)* but they cannot be counted on for style analysis in general. Many different kinds of tests have been proposed for use in author identification. Angela Glover (1996) [1] gives a comprehensive table of the features used in the tests.

"The following are some (tests) grouped by the degree of linguistic analysis of the data that is required for the test to be carried out.

[1] During 1787 and 1788, 77 articles were anonymously published to persuade New Yorkers to support ratification of the new US constitution. Most of the articles were later attributed credibly to their authors, except for 12 of them which were claimed to be written by both General Alexander Hamilton and President James Madison.

1) Tagged text

a. Type / token ratio
b. Distribution of word classes (parts of speech)
c. Distribution of verb forms (tense, aspect, etc)
d. Frequency of word parallelism
e. Distribution of word-class patterns
f. Distribution of nominal forms (e.g., gerunds)
g. Richness of vocabulary

2) Parsed text

a. Frequency of clause types
b. Distribution of direction of branching
c. Frequency of syntactic parallelism
d. Distribution of genitive forms (of and 's)
e. Distribution of phrase structures
f. Frequency of types of sentences
g. Frequency of topicalization
h. Ratio of main to subordinate clauses
i. Distribution of case frames
j. Frequency of passive voice

3) Interpreted text

a. Frequency of negation
b. Frequency of deixis
c. Frequency of hedges and markers of uncertainty
d. Frequency of semantic parallelism
e. Degree of alternative word use

It is reasonable that stylometers would not agree on the characteristic features to be analyzed in stylometry. Human languages are subtle, with many unquantifiable yet salient qualities. Idiolects, moreover, complicate the picture by highlighting particular features that are not shared by the whole language community. Individual studies on specific authors, understandably, rely on completely different measures to identify an author.

1.2 Classifiers

Efstathios Stamatatos, (2000) [3] states that there is no computational system as of today that can distinguish the texts of a randomly-chosen group of authors without the assistance of a human in the selection of both the most appropriate set of style markers and the most accurate disambiguation procedure.

Although there is a large variety in the methodology of stylometry, the techniques may be roughly divided into two classes: statistical methods and automated pattern recognition methods. The statistical group of methods normally features the application of Bayes' Rule in various ways to predict the probability of authorship; yet non-Bayesian statistical analyses have also been done by setting weights to different features. Statistical grouping techniques such as cluster analysis have proven to be useful by seeking to form clusters of individuals such that individuals within a cluster are more similar in some sense than individuals from other clusters [8]). The most widely used among these is the Principal Components Analysis (PCA), which arranges the observed variables in decreasing order of importance and plots the data in the space of just the first two components so that clustering is clearly visible [2,6]. Gram analysis is another commonly used technique. Text categorization, a related field, uses statistical methods like the bag of words to classify documents. Similar methods have also been tried for author attribution.

Automated pattern recognizing has also been used, though not as much [9, 12 ,13]. This involves training a neural network to learn the style of the texts. Neural networks have been used to learn the general style (and the genre) of a text and in text categorization. They have also been used to learn one particular author's style – to distinguish the one author from all others or from another particular author.

2 Components of the Problem

With automatic parsing algorithms getting more sophisticated, we believe that the role of artificial intelligence techniques in stylometry has great potential. An automatic text reader to parse the documents was written in Prolog to extract the data needed for the experiments.

2.1 Authors

Author attribution of documents – the most popular use of author recognition using stylometry – usually involves attributing an author to a document when there is some doubt about its authorship. This usually involves proving (or disproving) that the document was written by a particular author (to whom it has previously been attributed to) [10]. At times, this also would involve deciding between two authors. So, usually author recognition is done to learn the style of one (or two) author(s). This is inherently different from learning the styles of a number of authors (or in "learning the style") because in case of one (or two) authors, the list of features chosen may simply represent the authors' particular idiosyncrasies. However, these may not be in general good features that can be taken to be representative of authors' style. For example, usage of the words "no matter" may be a good feature that separates author X from author Y, but it may not be a feature that can be used to distinguish authors in general. In this paper we try to learn the styles of five authors – in hopes that the learners

would learn the general "style" of authors the rather than the idiosyncrasies of a particular author.
Also, documents can have a distinctive style without it being the style of the author. For instance, the texts from the Victorian era (written by any author) would seem to have the same style to a reader of this era. Thus, some care was needed in the selection of the texts so that they would not be easily distinguishable by their genre or period. Also, the authors should not have very different styles – or learning to differentiate between them would be trivial.

Five popular Victorian authors are chosen for the experiments: Jane Austen, Charles Dickens, William Thackeray, Emily Brontë and Charlotte Brontë.[2] The ASCII files of their works are freely downloadable at the Gutenberg Project Website.

2.2 Features Used

Hanlein's empirical research (1999) has yielded a set of individual-style features, from which the 21 style indicators in the present study are derived.

1. type-token ratio: The type-token ratio indicates the richness of an author's vocabulary. The higher the ratio, the more varied the vocabulary. It also reflects an author's tendency to repeat words.
2. mean word length: Longer words are traditionally associated with more pedantic and formal styles, whereas shorter words are a typical feature of informal spoken language.
3. mean sentence length: Longer sentences are often the indicator of carefully planned writing, while shorter sentences are more characteristic of spoken language.[3]
4. standard deviation of sentence length: The standard deviation indicates the variation of sentence length, which is an important marker of style.
5. mean paragraph length: The paragraph length is much influenced by the occurrence of dialogues.
6. chapter length: The length of the sample chapter.
7. number of commas per thousand tokens: Commas signal the ongoing flow of ideas within a sentence.
8. number of semicolons per thousand tokens: Semicolons indicate the reluctance of an author to stop a sentence where (s)he could.
9. number of quotation marks per thousand tokens: Frequent use of quotations is considered a typical involvement-feature [5].
10. number of exclamation marks per thousand tokens: Exclamations signal strong emotions.
11. number of hyphens per thousand tokens: Some authors use hyphenated words more than others.
12. number of *and*s per thousand tokens: *And*s are markers of coordination, which, as opposed to subordination, is more frequent in spoken production.
13. number of *but*s per thousand tokens: The contrastive linking *but*s are markers of coordination too.
14. number of *however*s per thousand tokens: The conjunct "however" is meant to form a contrastive pair with "but".
15. number of *if*s per thousand tokens: If-clauses are samples of subordination.
16. number of *that*s per thousand tokens: Most of the *that*s are used for subordination while a few are used as demonstratives.
17. number of *more*s per thousand tokens: 'More' is an indicator of an author's preference for comparative structure.
18. number of *must*s per thousand tokens: Modal verbs are potential candidates for expressing tentativeness [5]. *Must*s are more often used non-epistemically.
19. number of *might*s per thousand tokens: *Might*s are more often used epistemically.
20. number of *this*s per thousand tokens: *This*s are typically used for anaphoric reference.
21. number of *very*s per thousand tokens: *Very*s are stylistically significant for its emphasis on its modifiees.

These attributes include some very unconventional markers. Also, it is unconventional to use as few as twenty-one attributes in stylistic studies. Usually, a huge number of features (typically function words) are used .

In the following two sections, we discuss the learning methods used and describe the experiments and the results.

3. Decision Trees

Decision trees are rule based, explicit learners. Two decision trees were made to learn the styles of the authors. Text categorization uses decision trees extensively, but they have not been used in author identification. Due to their rule-based nature, it is easy to read and understand a decision tree. Thus, it is possible to "see the style" in the texts with these trees. Also, the results indicate that the unconventional features we used are quite useful in classification of the authors.

[2] In fact, many previous research included Austen and the Brontë sisters because of their similarity.

[3] Mean sentence length has been considered by some as unreliable, and the frequency distributions of the logarithms of sentence length have been used as well.

3.1 Experiments and Results

The See5 package by Quinlan is used in this experiment, which extends the basic ID3 algorithm of Quinlan. It infers decision trees by growing them from the root downward, greedily selecting the next best attribute for each new decision branch added to the tree. Thus, decision tree learning differs from other machine learning techniques such as neural networks, in that attributes are considered individually rather than in connection with one another. The feature with the greatest information gain is given priority in classification. Therefore, decision trees should work very well if there are some salient features that distinguish one author from the others.

Since the attributes tested are continuous, all the decision trees are constructed using the fuzzy threshold parameter, so that the knife-edge behavior for decision trees is softened by constructing an interval close to the threshold.

The decision tree constructed without taking any other options results in an error rate of 3.3% in the training set and an error rate of 23.5% in the test set, (averaged from cross validation) which is far above random guess (20%) but which is still not satisfactory. To improve the performance of the classifier, two different measures have been taken: winnowing and boosting.

3.1.1WINNOWING

When the number of attributes is large, it becomes harder to distinguish predictive information from chance coincidences. Winnowing overcomes this problem by investigating the usefulness of all attributes before any classifier is constructed. Attributes found to be irrelevant or harmful to predictive performance are disregarded (“winnowed”) and only the remaining attributes are used to construct the trees. As the relevance of features is one of the things we experiment on, winnowing was used. The package had a winnowing option which was used. This also makes building the trees a lot faster.

The result of this part of the experiment is shown below

```
Decision tree:
semicolon <= 6.72646 (8.723425): charles (18/2)
semicolon >= 8.97227 (8.723425):
:...semicolon >= 16.2095 (14.64195): charlotte (11/1)
    semicolon <= 14.6293 (14.64195):
    :...but >= 4.71113 (4.69338): jane (15/1)
        but <= 4.67563 (4.69338):
        :...quotation mark <= 12.285 (14.7294): emily (10)
            quotation mark >= 17.1738 (14.7294): william (7)

Evaluation on training data (61 cases):
             Tree Size      Errors
                 5          4 (6.6%)   <<
    (a)   (b)   (c)   (d)   (e)    ← classified as
   ----  ----  ----  ----  ----
    14                        1    (a): class jane
          16                       (b): class charles
     1     2     7                 (c): class william
                      10           (d): class emily
                            10     (e): class charlotte
Evaluation on test data (17 cases):
             Tree Size      Errors
                 5          3 (17.6%)   <<

    (a)   (b)   (c)   (d)   (e)    ← classified as
   ----  ----  ----  ----  ----
     4           1                 (a): class jane
           5           1           (b): class charles
                 2                 (c): class william
     1                 1           (d): class emily
                             2     (e): class charlotte
```

As shown, the decision tree has an accuracy of 82.4% on the 17 patterns in the test set. Also, the tree shows that ‘;’ (semicolon) and quotation marks were relevant attributes that define style (as they survived the winnowing).

3.1.2 BOOSTING

Boosting is a wrapping technique for generating and combining classifiers to give improved predictive accuracy. Boosting takes a generic learning algorithm and adjusts the distribution given to it (by removing some training data) based on the algorithm's behavior. The basic idea is that, as learning progresses, the booster samples the input distribution to keep the accuracy of the learner's current hypothesis near to that of random guessing. As a result, the learning process focuses on the currently hard data. The boosting option in the package causes a number of classifiers to be constructed; when a case is classified, all the classifiers are consulted before a decision is made. Boosting often gives higher predictive accuracy at the expense of increased classifier construction time. In this case, it yields the same error rate as in winnowing in the validation set, but has 100% accuracy in the training set. The result of this part of the experiment is as follows:

```
Decision tree that yields the best result:
semicolon <= 8.47458 (8.723425):
:...might >= 498339 (249170.4): william (2.5)
:   might <= 1.86846 (249170.4):
:   :...sen len stdev <= 23.1772 (24.8575): charles (14.9)
:       sen len stdev >= 26.5378 (24.8575): william (2.2)
semicolon >= 8.97227 (8.723425):
:...semicolon >= 14.6546 (14.64195):
    :...sen len stdev <= 10.2292 (12.68945): jane (1.9)
    :   sen len stan >= 15.1497 (12.68945): charlotte (10.4)
    semicolon <= 14.6293 (14.64195):
    :...but <= 4.67563 (4.69338):
        :...mean sen len <= 23.2097 (24.9242): emily (8.2)
        :   mean sen len >= 26.6387 (24.9242): william (6.4)
        but >= 4.71113 (4.69338):
        :...this <= 2.46609 (3.4445): jane (12.6)
            this >= 4.42291 (3.4445): william (1.9)

Evaluation on training data (61 cases):
Trial  Tree Size    Errors
   0        7        2 (3.3%)
   1        6        5 (8.2%)
   2        6       15 (24.6%)
   3        6       10 (16.4%)
   4        8        4 (6.6%)
   5        7        8 (13.1%)
   6        6        6 (9.8%)
   7        6       10 (16.4%)
   8        9        0 (0.0%)
boost                0 (0.0%)   <<

       (a)  (b)  (c)  (d)  (e)   ← classified as
       ---- ---- ---- ---- ----
       15                          (a): class jane
            16                     (b): class charles
                 10                (c): class william
                      10           (d): class emily
                           10      (e): class charlotte

Evaluation on test data (17 cases):
Trial      Tree  Size      Errors
   0        7            4 (23.5%)
   1        6            3 (17.6%)
   2        6           11 (64.7%)
   3        6           12 (70.6%)
   4        8            2 (11.8%)
   5        7            5 (29.4%)
   6        6            3 (17.6%)
   7        6           10 (58.8%)
   8        9            4 (23.5%)
boost                    3 (17.6%)   <<

      (a)  (b)  (c)  (d)  (e)   ← classified as
      ---- ---- ---- ---- ----
       4         1                (a): class jane
            5    1                (b): class charles
                 2                (c): class william
       1              1           (d): class emily
                           2      (e): class charlotte
```

As shown, the decision tree has an accuracy of 82.4% on the 17 patterns in the test set.

Noticeably, the two decision trees are constructed using different features and result in different classifications: in particular, Chapter 1 in *Tale of Two cities* is classified as written by Emily Brontë in the first case and as written by William Thakeray in the second case. Many of the conventional input attributes do not help in the classification; yet, the frequency of certain punctuation marks appears crucial in both of the decision trees, which verifies our hypothesis that a very important aspect of unconscious writing has been unduly ignored by stylometers.

4. Neural Networks

Neural networks are powerful pattern matching tools. They are basically very complex non-linear modeling equations. They are especially good in situations where the "concept" to be learned is very difficult to express as a well-defined, simple formulation, but rather is a complex, highly interrelated function of the inputs which is usually not easily transparent. This feature makes them ideal to learn a concept like "style" which is inherently intangible. The network takes all input attributes into consideration simultaneously though some are more heavily weighted than others. This differs from decision tree construction in that the style of an author is taken to be the joint product of many different features. Artificial Neural Networks has the ability to invent new features that are not explicit in the input, yet it also has the drawback that its inductive rules are inaccessible to humans.

4.1 Experiments and Results

Neuroshell – the commercial software package created by the Ward Systems Group, Inc. which is a package that creates and runs various types of neural networks, was used in this experiment.

Many structures of the multilayer network were experimented with before we came up with our best network. Kohonen Self Organizing Maps, probability networks, and networks based on statistical models were some of the architectures tried. Backpropogation feed forward networks yield the best result with the following architecture: 21 input nodes, 15 nodes on the first hidden layer, 11 nodes on the second hidden layer, and 10 output nodes (to act as error correcting codes). Two output nodes are allotted to a single author. (This increases the Hamming distance between the classifications - the bit string that is output with each bit corresponding to one author in the classification- of any two authors, thus decreasing the possibility of misclassification) 30% of the 61 training samples are used in the validation set which determines whether over-fitting has occurred and when to stop training. The remaining 17 samples were used for testing. This testing is also used to decide the architecture of the network.

After that, on the chosen network, cross validation was done by using different test sets. There was not much variation in the results obtained. On an average, the accuracy (measured on the test set) is 88.2%. There is one misclassification *Pride and Prejudice* which is misclassified as written by Charlotte Brontë; there is one misclassification in *Tale of Two Cities* - misclassified as written by William Thackeray. These are the misclassifications that were present in all the partitions used for cross validation (irrespective of whether the patterns are on testing set or in training/validation set).

The error obtained on one of the test sets is plotted in the following graph. It was observed that the network learned each author to be represented by two output nodes i.e. both the output nodes representing the same author never differed significantly in their values. So, the following graph shows only one output node for each author. It plots the error in all the output nodes for every pattern. Any deviation from zero is not a perfect classification. Spikes indicate the deviations from perfect classifications. Positive spikes indicate that non-authors are classified as authors and negative spikes indicate that authors are classified as non-authors. If a spike goes above 0.5 (or below -0.5), then a misclassification occurs.

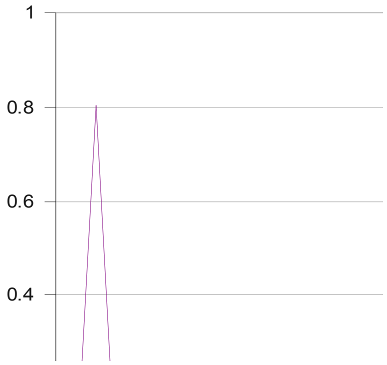

Figure 1 **error graph for Neural Networks**

As can be seen, in the test set (the last quarter), there are two misclassifications and there is one misclassification in the training set. It is interesting to note that every positive spike is accompanied by a negative spike. This means that every pattern is classified as belonging to one author and only one author. This is not built in. The network is free to predict 0 for all authors or 1 for more than one author – both meaning that it is undecided. This was not explicitly avoided. Nevertheless, the network always classified a pattern (either correctly or not).

The following table gives the classification of given by the network of the training and the validation sets

(a)	(b)	(c)	(d)	(e)	← classified as
11					(a): class jane
	9				(b): class charles
		10			(c): class william
			16		(d): class emily
1				14	(e): class charlotte

The following table gives the classification given by the network of the test set

(a)	(b)	(c)	(d)	(e)	← classified as
2					(a): class jane
	2				(b): class charles
		2			(c): class william
		2	4		(d): class emily
				5	(e): class charlotte

Also, the patterns that decision trees misclassify are not the same as those misclassified by the neural network, although the misclassification is persistent within on learning technique i.e. for more than one architecture, the same patterns get misclassified in neural network; Decision trees, with different parameters also have trouble with a same set of patterns (different from the set that troubles Artificial Neural Networks). The Neural Network is obviously looking at some different attributes from the ones that the decision trees look at.

5 Future Works

The fact that neural networks look at different features from those considered by decision trees could be put to use. A meta-learner could be trained to use both neural network and decision trees to get near perfect classifications.

Different set of features may be tried to see if there exists a set of feature which makes different learning techniques give the same results. Also feature extraction should be done in some care to train these learners with the most relevant features. Also, Automatic feature selecting algorithms (like winnowing) can be used to select the initial set of features.

The project could be extended to any number of authors. This project can classify five authors. The next step could be to generate abstract general inductive rules that can be used in all author identification problems instead of learning the ways to separate a group of n chosen authors.

Some clustering algorithms should be tried on the data to see if they perform better. Self Organizing Maps were tried. However, they did not yield good results. Other statistical methods may provide better results.

6 Conclusions

This paper attempts to recognize different authors based on their style of writing (without help from genre or time). Both decision trees and Artificial Neural Networks yield a significantly higher accuracy rate than random guess, which shows that the assumption is well justified that there is a quantifiable unconscious aspect in an author's style. We achieved 82.4% accuracy the test set using decision trees and 88.2% accuracy on the test set using neural networks. Although the neural networks yielded a better numerical performance, and are considered inherently suitable to capture an intangible concept like style, the decision trees are human readable making it possible to define style. Also, the results obtained by both methods are comparable.

Most of the previous studies in stylometry input a large number of attributes, such as the frequencies of fifty-nine function words in some recent *Federalist Papers* research. While more features could produce additional discriminatory material, the present study proves that artificial intelligence provides stylometry with excellent classifiers that require fewer input variables than traditional statistics. The unconventional features used as style markers proved to be effective. The proven efficiency of the automatic classifiers marks them as exciting tools for the future in stylometry's continuing evolution (Holmes 1998: 115). We believe that the combination of stylometry and AI will result in a useful discipline with many practical applications.

Acknowledgements

We would like to thank Dr. Covington, Associate Director, and Artificial Intelligence Center for his advice and guidance. We are grateful to Dr. Potter, graduate coordinator of Artificial Intelligence Center, UGA for his help.

References

[1] Glover A. and H. "Detecting stylistic inconsistencies in collaborative writing" in *The new writing environment: Writers at work in a world of technology*, edited by Mike Sharples and Thea van der Geest, London: Springer-Verlag, 1996.

[2] Binongo, J. N. G. & Smith, M. W. A. "The application of principal component analysis to stylometry." *Literary and Linguistic Computing*, Vol. 14, No. 4, pp445-465, 1999.

[3] Stamatatos E, Fakotakis N., and Kokkinakis G. "Automatic Text Categorization in Terms of Genre and Author" *Computational Linguistics*, 26(4), December 2000, pp. 471-495, 2000.

[4] Forsyth, R. S. "Cicero, sigonio, and burrows: investigating the authenticity of the Consolatio." *Literary and Linguistic Computing*, Vol. 14, No. 3, pp311-332, 1999.

[5] Hanlein, H. "Studies in Authorship Recognition: a Corpus-based Approach". Peter Lang, 1999.

[6] Baayen1 H., Halteren1 H., Neijt1 A., Tweedie F., "An experiment in authorship attribution" *JADT 2002 : 6th International Conference on the Statistical Analysis of Textual Data*. 2002.

[7] Holmes, D. I. "The evolution of stylometry in humanities scholarship." *Literary and Linguistic Computing*, Vol. 13, No. 3, pp111-117, 1998.

[8] Holmes, D. I. & Forsyth, R. S. "The Federalist Revisited: New Directions in Authorship Attribution." *Literary and Linguistic Computing*, Vol. 10, No. 2, pp111-127, 1995.

[9] Hoorn, J. F. "Neural network identification of poets using letter sequences." *Literary and Linguistic Computing,* Vol. 14, No. 3, pp311-332, 1999.

[10] Merriam, T. "Heterogeneous authorship in early Shakespeare and the problem of Henry V." *Literary and Linguistic Computing*, Vol. 13, No. 1, pp15-27, 1998.

[11] Rudman, J. "The state of authorship attribution studies :some problems and solutions." *Computers and the humanities,* 31:351 –365, 1998.

[12] Tweedie, F., Singh, S. and Holmes, D. I. "Neural Network Applications in Stylometry: The Federalist Papers", *Computers and The Humanities*, vol. 30, issue 1, pp. 1-10, 1996.

[13] Waugh, S. "Computational stylistics using artificial neural networks." *Literary and Linguistic Computing,* Vol. 15, No. 2, pp187-197, 2000.