

Stylometry for E-mail Author Identification and Authentication

K. Calix, M. Connors, D. Levy, H. Manzar, G. McCabe, and S. Westcott,
Seidenberg School of CSIS, Pace University, New York
{kc32446n, mc56253p, dl09253w, gm60518w}@pace.edu,
huriamanzar@hotmail.com, profwestcott@yahoo.com

Abstract

The identification of the authorship of e-mail messages is of increasing importance due to an increase in the use of e-mail for criminal purposes. An author's unique writing style can be reduced to a pattern by making measurements of various stylometric features from the written text. This paper reports on work to optimize and extend an existing C# based stylometry system that identifies the author of an arbitrary e-mail by using fifty-five writing style features. The program has been extended to provide feature vector data in a format appropriate for distribution to other project teams for subsequent data mining and classification experiments.

1. Introduction

Stylometry is the study of the unique linguistic styles and writing behaviors of individuals in order to determine authorship. Its underlying assumption is that an author displays distinctive writing habits, which are exhibited in such features as the author's core vocabulary usage, sentence complexity and phraseology. Variations in style among authors can be determined to be a result of differences in genre or content, authorial preferences and competence, communicative expression and the expectations of an intended audience [12]. An author's linguistic style is thought to have certain features that are independent of the author's will, and since these features cannot be

consciously manipulated by the author, they are considered to provide the most reliable data for a stylometric study. Stylometry attempts to define the features of an author's style and to determine those statistical methods necessary to measure those features of similarity between two or more textual sources.

Stylometry is presently entering a new era. In the past fifteen years, researchers have developed a wealth of mathematical tools, from statistical tests to Artificial Intelligence Techniques, for use in determining authorship. Researchers have applied these tools to texts from a wide range of literary genres and time periods, which include the following: The Federalist Papers; Civil War Letters; Shakespeare's plays; The New Testament; The Royal Book of Oz, and The Dialogues of Plato [14][16]. Stylometry is used for plagiaristic detection, which is presently a serious concern in many academic institutions. In 2005 Stefan Gruber and Stuart Noven, proposed a software tool that supports detection of plagiarism [10].

E-mail usage has increased dramatically in recent years as an important means of communication. Unfortunately, its potential for inappropriate usage has also increased. E-mail is being misused for

activities such as sending spam messages, hoaxes and threats. Crimes committed through e-mail are becoming commonplace. Therefore, it is important to properly identify e-mail authorship.

The purpose of this paper is to outline the methodology of developing and testing the system previously developed [9]. In addition to becoming familiar with the existing system, the researchers collected additional data in the form of plaintext e-mails from a wide range of subjects and developed a methodology for formatting the feature vector data to facilitate processing by the biometric authentication and data mining systems [2][8].

2. Relevance in the Context of Other Work

The use of stylometry to determine authorship dates to precomputer times. A number of scientists were interested in applying stylometric features to textual analysis, and in the late 1880's an American physicist, Mendenhall, suggested that authorial styles could be 'fingerprinted' by counting the numbers of letters in the words they used [15]. The application of counting the features of a text was later extended by Yule in the early 1900's to include the lengths of sentences [12]. In the mid-1960's, Morton applied sentence-lengths for testing of Greek prose authorship [12]. Zipf, a Harvard University German Professor, employed individuals as "human computers" in order to count the number of times each word appeared in a text so as to rank the frequency of specific words [4] [18].

Today, various combinations of features appear in a number of related stylometric

studies. Kacmarcik and Gamon stated: "feature selection is one of the most crucial aspects of authorship attribution." Their study is limited, however, to word frequencies because these features are generally acknowledged to be an effective way to determine authorship attribution. Feature selection allows the researcher to easily incorporate word frequencies in document verification. Kacmarcik and Gamon employed tokenization in order to separate sentences by using white spaces, new lines and punctuation marks so as to create unique tokens in the training set [13]. Tweedie and Baayen recommend the use of the developmental profiles of selected constants, rather than the isolated values of the constants for complete textual authorship attribution [17]. When more features are used in combination with one another, the discriminatory potential is increased. Gruber and Noven, produced sixty-two stylometric measurements, applied to pairs of text, out of sixty-five originally, suggested by Norton and Hilton [10].

There are currently many publications on stylometric research. The particular focus of this study is on the use of stylometry in the identification pertaining to the authorship of e-mail text messages. In 2003 Corney published a thesis entitled "Analyzing E-mail Text Authorship for Forensic Purposes." He advocated that stylometry is a valuable tool for computer forensics and investigation to determine authorship of anonymous messages. By analyzing e-mail texts he came to the conclusion that a combination of character based, word length frequency distribution, and function word attribute is an effective combination of features [6]. De Vel et al. used basic stylometric features on a set

number of authors without consideration of the authorial characteristics, such as gender, language, e-mail topic, or length [7]. Argamon, et al. used computational stylistics in ascribing authorship attribution to electronic messages. They also emphasized difficulties in identifying authorship of an e-mail text for two of the following reasons: firstly, the concise nature of e-mail messages (tens or perhaps hundreds of words comparing to thousands for articles and books) and secondly, the variation in the individual style of e-mail messages due to the fact that e-mails, as an informal and fast-paced medium, exhibit variations in an individual's writing styles due to the adaptation to distinct contexts or correspondents [1].

3. Methodology

The stylometry program was written in the C# programming language, and uses a Graphical User Interface (GUI) to simplify the tasks of determining authorship by automating the identification process. Determining authorship involves data collection, feature extraction, and classification.

Users train the program to recognize authors by initially selecting a set of sample e-mails labeled with known authors (including author demographics) and subsequently selecting a set of sample e-mails by unknown authors for comparison. Fifty-five stylistic features are considered for the program. The list of features is provided in Table 1.

Table 1. Stylistic Features

1. Number of sentences beginning with upper case
2. Number of sentences beginning with lower case
3. Number of Words
4. Average Word Length
5. Number of Sentences
6. Average Number of Words per Sentence
7. Number of Paragraphs
8. Average Number of words per Paragraph
9. Number of Exclamation Marks
10. Number of Number Signs
11. Number of Dollar Signs
12. Number of Ampersands
13. Number of Percent Signs
14. Number of Apostrophes
15. Number of Left parentheses
16. Number of Right parentheses
17. Number of Asterisks
18. Number of Plus Signs
19. Number of Commas
20. Number of Dashes
21. Number of Periods
22. Number of Forward Slashes
23. Number of Colons
24. Number of Semi-colons
25. Number of Pipe Signs
26. Number of Less than Signs
27. Number of Greater than Signs
28. Number of Equal Signs
29. Number of Question Marks

30. Number of At Signs
31. Number of Left square brackets
32. Number of Right square brackets
33. Number of Backward slashes
34. Number of Caret Signs
35. Number of Underscores
36. Number of Accents
37. Number of Left curly braces
38. Number of Right curly braces
39. Number of Vertical lines
40. Number of Tildes
41. Number of White spaces
42. Number of Multiple Question Marks
43. Number of Multiple Exclamation Marks
44. Number of Ellipsis
45. Average Number of Periods per Paragraph
46. Average Number of Commas per Paragraph
47. Average Number of Colons per Paragraph
48. Average Number of Semi-colons per Paragraph
49. Average Number of Question Marks per Paragraph
50. Average Number of Multiple Questions Marks per Paragraph
51. Average Number of White Spaces per Sentence
52. Number of times "Well" appears
53. Number of times "Anyhow" appears
54. Average Number of Times the word "Anyhow" appears
55. Average Number of Times the word "Well" appears

For comparative analysis and authentication purposes, measurements of the stylistic features are normalized in

the range 0 – 1 and classified by the k-nearest neighbor algorithm using

Euclidean distance to determine authorship of the unknown email.

3.1 Data Collection

Data, from twelve participants, was gathered from plaintext e-mails. Each participant created ten e-mails, which averaged one hundred and fifty (150) words, each on a distinct subject. Demographics for each author are recorded when training the program for more definitive identification. Figure 1 depicts the author demographic input form.

Text File Author Information

Please complete the form with the author information for the file named: 'Melissa_10'

Click on OK to accept Author Information.
Click on RESET to clear the form for a new author.

Author Name:

Author Age:

Gender:

Degree:

Computer:

Save this info for next?:

OK RESET

Figure 1: Author Demographic Form

3.2 Feature Extraction

To provide feature vector data for classification and data mining experiments, the totals of selected stylistic features were derived and the averages of such features were calculated for each author. Simple division was used to calculate each average. For example, dividing the “Number of Words” by the “Number of Paragraphs” derived the “Average Number of Words per Paragraph.” These features include those presented in Table 1.

The feature values derived were normalized into the range of 0 – 1. They were recorded in the data file with fields in a record, comma delimited and items in a field slash delimited, as depicted in Figure 2.

```

Stylometry biometric data example created december 2007
120
Andrea/F/ 31, MS/IT, Dell Computer, structured email task,11,0.48387096774
Andrea/F/ 31, MS/IT, Dell Computer, structured email task,11,0.32258064516
Andrea/F/ 31, MS/IT, Dell Computer, structured email task,11,0.77419354838
Andrea/F/ 31, MS/IT, Dell Computer, structured email task,11,0.25806451612
Andrea/F/ 31, MS/IT, Dell Computer, structured email task,11,0.22580645161
Andrea/F/ 31, MS/IT, Dell Computer, structured email task,11,0.09677419354
Andrea/F/ 31, MS/IT, Dell Computer, structured email task,11,0.06451612903
Andrea/F/ 31, MS/IT, Dell Computer, structured email task,11,0.09677419354
Andrea/F/ 31, MS/IT, Dell Computer, structured email task,11,0.29032258064
Andrea/F/ 31, MS/IT, Dell Computer, structured email task,11,0.16129032258
Claude/M/ 78, MBA/FIN, Dell Laptop, structured email task,11,0.41935483870
Claude/M/ 78, MBA/FIN, Dell Laptop, structured email task,11,0.35806451612

```

Figure 2. Example of the Feature Vector Data

3.3 Classification

Various techniques are used in pattern classification, such as the following: Decision Trees, Bayesian Theory, Neural Networks or *k*-nearest neighbor (KNN). This program uses the *k*-nearest neighbor algorithm, which classifies objects based on similarities or distance metric [5].

K-nearest neighbor classifiers are based on learning by analogy. The training samples are described by *n*-dimensional numeric attributes. Each sample represents a point in an *n*-dimensional space. All training sample are, therefore, stored in an *n*-dimensional pattern space. When an unknown sample is presented, the classifier searches the pattern space for the *k* training samples which are closest to the unknown sample, the *k* “nearest neighbors” of the unknown sample. This closeness is defined in terms of Euclidean distance [11].

The unknown author sample is assigned the most common class among its *k*-nearest neighbours. When *k* = 1, the unknown sample is assigned to the class

of the training sample that it is closest to in the pattern space [11].

Figure 3 below depicts the classification phase using this stylometric program.

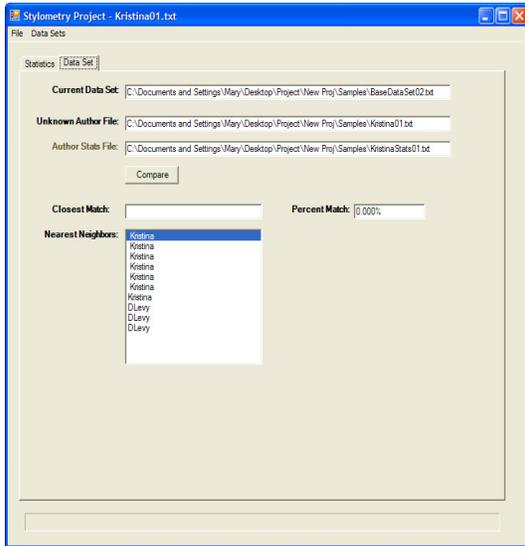


Figure 3. Illustration of Classification Phase

4. Results

A base data set of 120 data files was constructed. The data files consist of plain text e-mails, and are classified as “structured e-mail task,” where the author composed an e-mail using a desktop or laptop keyboard. Each author provided ten sample e-mails in order to train the program.

The data from these plain text emails were derived by first counting the features then displaying the average number of selected stylistic features identified. These were then normalized for use in the authentication process.

The dichotomy data [2] for the Stylometry authentication experiments contained 1770 records for each subset of six subjects. Each subset was run against the other yielding 76.72% and 66.72% accuracy.

100% accuracy was obtained from all three identification experiments performed by the data mining team on the feature vector data [8]. One test used a full data set as training and a full data set as test with the leave-one-out procedure. Second experiment used first five (five samples from each of 12 subjects) for training and the last 5 (5 samples from each of 12 subjects) as test. Third test used last 5, first 5 to yield 100% accuracy also. This higher level of accuracy compared to 80% level obtained with existing system [9] may be attributed to the larger dataset used and more substantial email samples.

4.1 Difficulties

As was stated earlier, our aims for this project included the following: familiarization with the system, the collection of additional plaintext data and the formatting of the feature-vector data, as normalized averages for ease of processing, by other project systems. These project aims have been met; however, the following difficulties may be noted:

The first difficulty concerned the attempt to input the author’s demographic information into the CSV file. Initially, only the author’s name from the demographic form was captured.

Saving the author’s information also required the researcher to input repeatedly the author’s demographic information. The program was, subsequently, amended to include a reset option, thereby eliminating the need to re-enter similar information for each sample.

Secondly, deriving normalized averages from the base data set was somewhat

challenging. Initially, the program was able to display only three required normalized features; namely, the “Average Number of Words per Sentence,” the “Average Number of Words per Paragraph,” and the “Average Word Length.” This was also later changed to display eleven (11) normalized features.

It was also difficult to create a file, which contained all required normalized averages. Therefore, the existing code was amended to both, calculate the features for the spreadsheet averages and also to create the file with respective headings, as represented in Figure 2.

Thirdly, the program code and the GUI had to be amended so as to both remove keystroke data and also to render the GUI more “user friendly.” The GUI was changed to display only information useful for this stylometric experiment of email authorial identification. A menu option has been added to assist the user to create normalized data for the authentication process. Figure 5 shows the new interface.

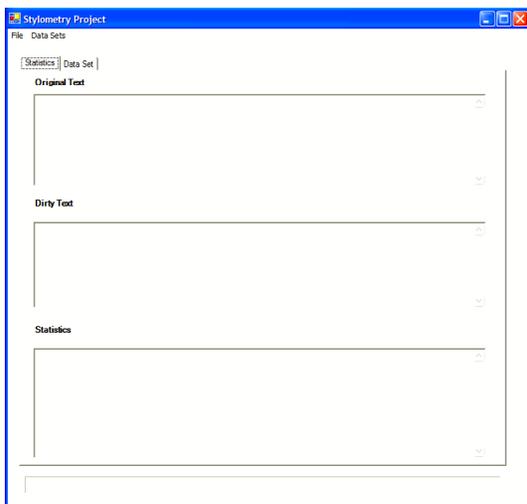


Figure 4: Old Interface

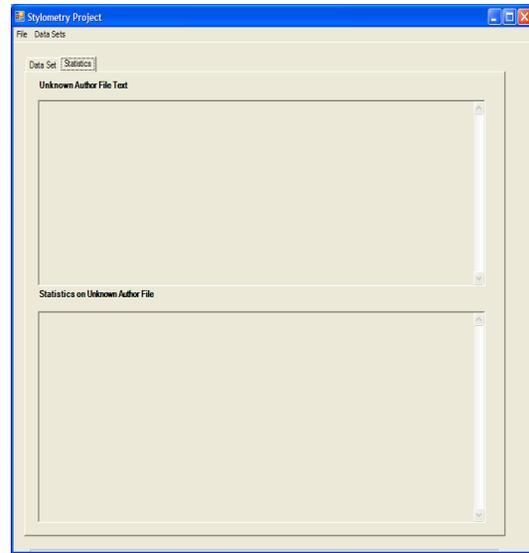


Figure 5: New Interface

5. Conclusion

This program enables the non-specialist to undertake a stylometric study so as to identify and authenticate the authorship of e-mail text messages. The program was designed for users with different levels of technical experience, from novice to expert.

The program was used to analyze, linguistically and stylistically, e-mails by identifying the commonality of symbols, word frequencies and punctuation marks. In the future, it may be helpful to extend the authentication task to identify patterns in frequently used misspelled and misused words.

Biometric and data mining techniques were utilized in the authentication process. The program has also sought to identify authorship through gender and level of education attained. Presently, this data is entered manually and hence is prone to entry errors. Therefore, additional work in this area may include enabling the program to identify the author's gender based on stylistic and linguistics habits.

8. References

- [1] Shlomo Argamon, Marin Šarić, Sterling S. Stein, "Style Mining of Electronic Messages for Multiple Authorship Discrimination: First Results." S. ACM, New York. 2003.
- [2] S. Bharati, R. Haseem, R. Khan, M. Ritzmann, and A. Wong, "Biometric Authentication System using the Dichotomy Model," Proc. CSIS Research Day, Pace Univ., May 2008.
- [3] Paul E. Black, *Euclidean distance*. U.S. National Institute of Standards and Technology. 2004. <http://www.nist.gov/dads/HTML/euclidndstnc.html>
- [4] A. Bogomolny, *Benford's Law and Zipf's Law*. http://www.cut-the-knot.org/do_you_know/zipfLaw.shtml
- [5] Ming-Syan Chen, Philip S. Yu, Bing Liu, "Advances in Knowledge Discovery and Data Mining": 6th Pacific-Asia Conference, p. 517. <http://books.google.com/books?id=k3AoWbkE-1YC&pg=PA517&dq=k+nearest+neighbor&sig=ZQMhHzCihDOa6tEhMa9IXyebMc>
- [6] M. Corney, "Analyzing E-mail Text Authorship for Forensic Purposes." March 2003. http://sky.fit.qut.edu.au/~corneym/papers/mit_the_sis.pdf
- [7] O. de Vel, A. Anderson, M. Corney, G. Mohay, "Mining E-mail Content for Author Identification Forensics." ACM, New York. 2001.
- [8] Clara Eusebi, Cosmin Gilga, Deepa John, Andre Maisonave, "A Data Mining Study of Mouse Movement, Stylometry, and Keystroke Biometric Data," Proc. CSIS Research Day, Pace Univ., May 2008.
- [9] R. Goodman, M. Hahn, M. Marella, C. Ojar, and S. Westcott, "The Use of Stylometry for Email Author Identification: A Feasibility Study", Proc. Student/Faculty Research Day, CSIS, Pace University, White Plains, NY, May 2007, pp.1-7.
- [10] Stefan Gruber & Stuart Noven, "Tool support for plagiarism detection in text documents." Symposium on Applied Computing archive Proceedings of the 2005 ACM symposium on Applied computing Pages: 776 – 781. Year of Publication: 2005.
- [11] Jiawei Han & Micheline Kamber, "Data Mining: Concepts and Techniques," pp. 314 – 315. http://books.google.com/books?id=6hkR_ixby08
- C&pg=PA314&dq=k+nearest+neighbor&sig=4C15mNAK_A4RfweZrWWc-e5UuAE
- [12] David I. Holmes, "Stylometry: Its Origins, Developments and Aspirations." <http://www.cs.queensu.ca/achallc97/papers/s004.html>
- [13] Gary Kacmarcik, & Michael Gamon, "Obfuscating Document Stylometry to preserve Author's Anonymity." Microsoft Research. 2006. <http://research.microsoft.com/nlp/publications/aclcoling06-kacmarcik-gamon.pdf>
- [14] Erica Klarreich, "Statistical tests are unraveling knotty literary mysteries." Science News Online. <http://www.sciencenews.org/articles/20031220/ob8.asp>
- [15] Mendenhall, T. C. "The characteristic curves of composition." (1887) in Science, Vol.IX no.214 (supplement) pp.237-249, and "A mechanical solution of a literary problem" (1901) in The Popular Science Monthly, Vol.LX no.7, pp.97-105 cited in "A Deception in Deptford. Christopher Marlowe's Alleged Death." Farey, Peter. 1997-2000.
- [16] S. Michaelson & A. Q. Morton, "The New Stylometry: A One-Word Test of Authorship for Greek Writers." The Classical Quarterly, New Series, Vol. 22, No. 1 (May, 1972), pp. 89-102.
- [17] Fiona J. Tweedie & R. Harald Baayen, "Lexical Constants' in Stylometry and Authorship Studies." <http://www.cs.queensu.ca/achallc97/papers/s004.html>
- [18] R.S. Wallace, "Zipf's Law." <http://www.alicebot.org/articles/wallace/zipf.html>