Evaluating Text Visualization: An Experiment in Authorship Analysis

Victor Benjamin[1], Wingyan Chung[2], Ahmed Abbasi[3], Joshua Chuang[4], Catherine A. Larson[1], Hsinchun Chen[1]
[1]MIS Dept., University of Arizona, Tucson, AZ USA (contact author: vabenji@email.arizona.edu); [2]Dept. of Management, Fayetteville State University, Fayetteville, NC USA; [3]McIntire School of Commerce, University of Virginia, Charlottesville, VA USA; [4]HC Analytics, Tucson, AZ USA

***Abstract*—Analyzing authorship of online texts is an important analysis task in security-related areas such as cybercrime investigation and counter-terrorism, and in any field of endeavor in which authorship may be uncertain or obfuscated. This paper presents an automated approach for authorship analysis using machine learning methods, a robust stylometric feature set, and a series of visualizations designed to facilitate analysis at the feature, author, and message levels. A testbed consisting of 506,554 forum messages, in English and Arabic, from 14,901 authors was first constructed. A prototype portal system was then developed to support feasibility analysis of the approach. A preliminary evaluation to assess the efficacy of the text visualizations was conducted. The evaluation showed that task performance with the visualization functions was more accurate and more efficient than task performance without the visualizations.**

Keywords-terrorism; text visualization; online forum; authorship analysis.

I. INTRODUCTION

Authorship analysis can be useful in any application area when attribution is uncertain (e.g., in academic fields such as the humanities or history) or obfuscated (e.g., plagiarism, cybercrime). In the security context, manual analysis of authorship is extremely difficult due to both the volume of criminal activity and the exponentially increasing amounts of online text (whether email, forums, blogs, websites, IRC channels, etc.). The Arizona Authorship Analysis Portal (AzAA) was conceived as a platform on which automated approaches to analyzing authorship identities could be developed and assessed. The portal integrates machine learning methods, a robust stylometric feature set, and a series of visualizations designed to facilitate analysis at the feature, author, and message levels. It allows identification of "extreme" jihadi authors, and also supports analysis and comparison of author writing styles.

In this paper, we describe the task of authorship analysis, examine the previous works that contributed to the portal's development, provide an overview of the system components, and describe a recent preliminary evaluation of the portal's text visualization function. We conclude with next steps and future development.

II. BACKGROUND

The AzAA portal was initially built as an extension of the Dark Web (DW) Forum Portal, a large archive of international terrorist and extremist Web forums containing over 15M messages in several different languages [Zhang et al. 2010, Chen 2011, Chen at al. 2011]. The DW portal supports retrieval and analysis of forum postings through searching and browsing functions, multilingual translation, and social network analysis. The most recent version of the DW portal did not include the ability to perform authorship analysis, which can be important both for cybercrime investigation [Zheng et al. 2003; Frantzeskou and Gritzali 2004] and counter-terrorism [Abbasi and Chen 2006]. The AzAA portal was built to address this important gap and provide additional support for terrorism research and intelligence analysis.

III. PREVIOUS WORK

Relevant previous work in semi-automated data collection, and in text analysis including stylometry and authorship analysis, sentiment and affect analysis, and text visualization are briefly reviewed here.

Manual collection of Web-based pages is time-consuming and clumsy, and often results in unusable data. Previously developed techniques for semi-automated data collection that support fast and accurate "spidering" methods allowed large amounts of forum text to be collected relatively quickly [Fu et al. 2010; Zhang et al. 2010].

Stylometric analysis techniques based on previous research serve as the foundation of the analysis tasks performed on the testbed. Stylometry can be defined most simply as the analysis of literary style [McHenry & Oakes 2000; Zheng et al. 2006]. Authorship analysis evolved from the application of computer-based statistical techniques for textual analysis and depends on the extraction of literary style and other features for analysis.

Features are the characteristics that help distinguish one text or body of related texts from another. They are categorized in a variety of ways depending on the discipline. This study incorporated text style features, HTML features, and content-specific features. Text style features generally include structural features, which characterize the overall organization and arrangement of the text as a whole; syntactic features, which relate to the sentence-level structure; and lexical features, which are associated with word-level characteristics such as frequencies or distributions of characters, lemmas, vocabulary, etc. (Abbasi & Chen 2006). HTML features may be sometimes categorized as a structural feature (de Vel et al. 2001) or as a non-structural feature (Layton et al. 2012). They include those features specific to the HTML-encoded text, such as tags, links, fonts, etc. (Urvoy et al. 2008). Content-specific features are the keywords relevant to the topics inherently

contained in the text (Abbasi & Chen 2006).

The authorship analysis task generally has one of three purposes: *authorship identification*, *authorship characterization* or profiling, and *similarity detection*. *Authorship identification* examines an author's known set of writings in order to determine the level of possibility that he or she is the author of a particular unattributed or possibly mis-attributed work. Li et al. [2006] investigated the key features important to identifying authorship of online texts. Argamon et al. [2009] developed a means of constructing a "profile" of an author's characteristics (such as age, gender, personality) and determined which features were most effective for profiling each characteristic type. Abbasi and Chen [2008a] focused on stylistic features for the *similarity detection* task, where authors' identities are not known ahead of time but writings can be compared for their similarity/dissimilarity to each other.

Sentiment analysis, a variant of text mining, has its roots in natural language processing and also relies on computational linguistics techniques. Its purpose is to classify the attitudes and polarity inherent in a document or other text [Abbasi et al. 2008; Zimbra & Chen 2012]. Similarly, affect analysis represents the specific emotions, moods, or attitudes of an author [Balahur et al. 2010]. Both sentiment and affect analysis can be used to scrutinize a text to potentially reveal the author's opinions and affect state concerning multiple items. Automated techniques such as machine learning classifiers are commonly used to implement such analyses.

Text visualization is a broad field that aims to graphically represent the different types of content and stylometric features within a particular piece of writing. Such visualizations can help develop various perspectives for better understanding or future analysis of text [Abbasi & Chen 2008b; Chuang et al. 2012]. Typically, such visualizations are intended to help analysts better understand large textual datasets by providing graphical representations of various aspects of them; statistical tests and feature counts are often times the outputs included [Chuang et al. 2012]. Text visualizations can be used to help compare writing samples between authors, as emphasis can be placed on the differences and similarities between text data.

IV. The Authorship Portal

A. Research Testbed and Feature Set

The purpose of the prototype system is to support the identification of "extreme" Jihadi authors of postings in forums from the Dark Web Forum Portal, and to allow analysis and comparison of author writing styles. The design framework is shown in Figure 1.

A testbed was first constructed using a semi-automated spidering method that collected Web forum postings in English and Arabic. After collection, text parsers were created to sift through all downloaded forum web pages, extract relevant data, and save it to a local database.

Feature extraction was performed by a multi-class classifier using a J48 Decision Tree. Text style and HTML features were extracted for stylometric analysis. The text style features included lexical, syntactic, and structural features. The HTML features, commonly used to add additional context, information, and style to messages, included the images, links, fonts, alignment, and HTML tags. During machine learning training, 670 English features and 392 Arabic features were extracted from the forums, respectively.

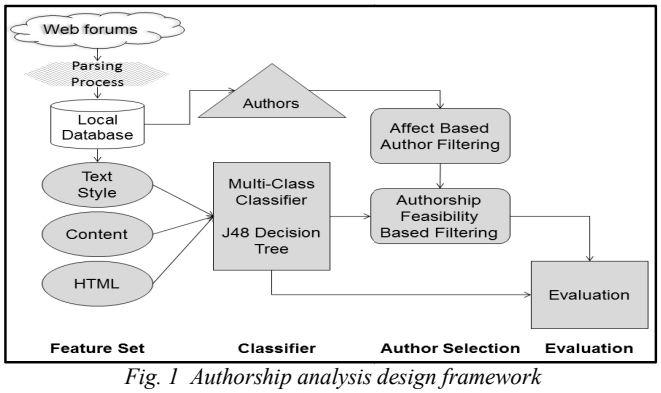

Fig. 1 Authorship analysis design framework

Sentiment and affect analysis was performed using machine learning classification. We extracted content-specific features based on feature frequency and classifier information gain. Such features included religious/cultural terms, sentiment cues, and words associated with violence, anger, hate, and racism.

These techniques allow the authors' styles, and authors with the highest sentiment and affect intensities for anger, violence, hate, etc. to be identified and selected through filters. Thirty of the most extreme authors were identified by the classifier.

All authors do not necessarily have a unique pattern, however; some exhibit writing patterns that are erratic and/or include considerable reposting, quoting, plagiarism, non-sequiturs, short responses, etc. To evaluate the author selection, feasibility analysis was conducted to determine which authors were detectable. The latest 100 posts for each of the 30 identified authors were used for evaluation purposes. Recent postings were selected to avoid writing style changes that may occur naturally over time. For each author, 50 messages were used for classifier training and the other 50 are for testing. Using all three feature sets together (text style, HTML, and content-specific features) for authorship identification yielded good performance (90% accuracy).

B. System Architecture

The system backend was built with Java EE running on Apache Tomcat. The data is pre-processed and stored within an MSSQL Server. The Struts2 framework Framework and Spring framework, two popular enterprise-level open-source frameworks, were adopted for better scalability, flexibility, compatibility, and extendibility, thus allowing the portal to be more easily integrated into other local projects sharing the same frameworks. The front-end design and implementation were through JSP, Javascript, JQuery, HTML, CSS, and Bootstrap.

C. Use of the Authorship Portal

The portal provides users with multiple perspectives for viewing data. From the welcome screen a user may start with the author- or message-based perspective (Figure 2).

In the Author-Based Perspective, for example, users can

view which authors use particular stylometric features the most (Figure 3). Ranked lists of feature usage are organized by affect words, text style features, HTML features, and content-specific features (1). The data can be viewed in summary form (2; shown) or as a heat map. The stylometric features between two authors can also be compared (3)

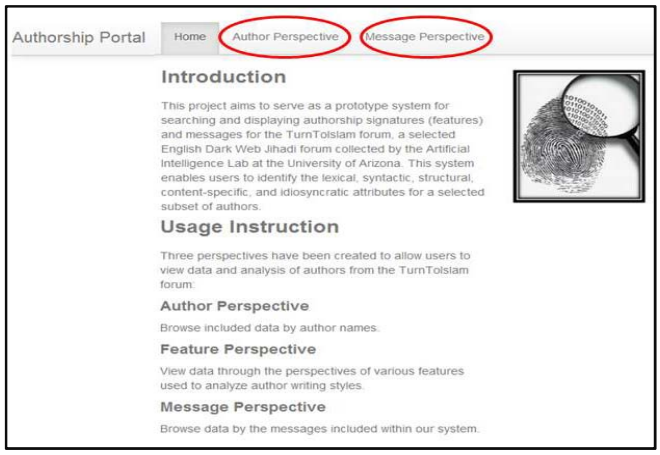

Fig. 2. On the authorship analysis welcome screen, users can start with the Author or Message Perspective.

Fig. 3. A portion of the Author Perspective screen showing which authors use which stylometric features the most.

Visualizations provide quick context on author stylometric features. Heat maps (Figure 4) and radar charts (Figure 5) can be invoked for comparing authorship styles. Figure 4 shows a portion of the heat map, where in this instance authors are sorted by their usage of racist terminology. The darker and more intense shades indicate greater usage. Racism appears to be a prominent feature in the writing of user muharram23, the first author listed.

Figure 5 shows a radar chart highlighting the similarities and differences in stylometric feature usage between two different authors. In this example, the user Muharram23 is compared against Brother4ever across various stylometric features. As shown in the previous example, Muharram23 uses a great deal of racist terminology within his messages; conversely, Brother4ever appears to discuss a wider range of topics, particularly religion and culture.

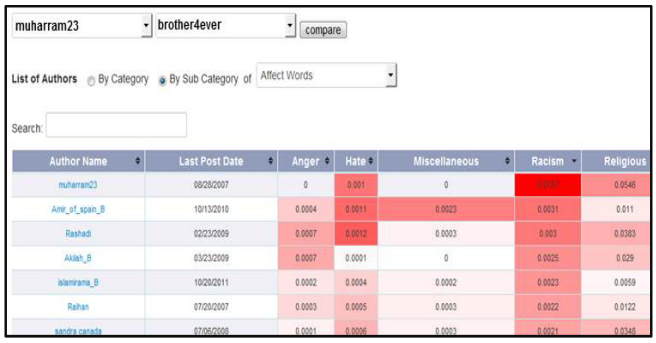

Fig. 4 The"Heat map view" (here showing usage of racist terms) provides a comprehensive overview of which features are the most distinctive for each author, listed in rows. The darker and more intense shades indicate greater usage.

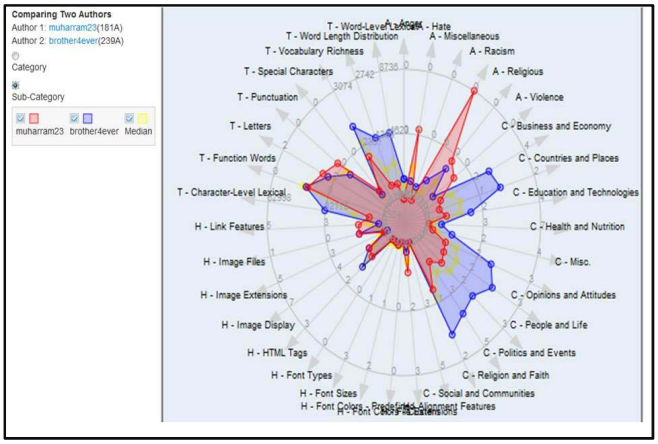

Fig. 5 The radar chart visualization

Users may also browse the raw content enhanced by feature highlighting in the text. In Figure 6, affect words reflecting hate, violence, anger, etc. are highlighted in the message with "warm" tones (red, orange, pink). Content-specific features indicating topical associations (such as "Religion and Faith," "Education and Technologies," etc.) are highlighted in "cooler" tones (blues, greens, purples). In addition, if a user is interested in a particular topic, messages can be filtered with search terms.

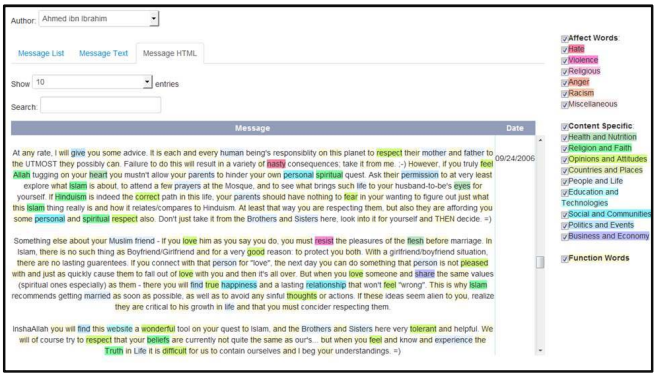

Fig. 6 Message feature highlighting

The data representation types selected for the portal each have advantages and disadvantages. Heat maps are useful for representing complex data and various data types, and can impart legibility to "unordered" ordinal data. They also use humans' natural ability to discern pattern violations. A gradient, as opposed to "rainbow," heat map was chosen to avoid the luminance effect: the fact that colors next to each other can affect or interfere with perception. Radar charts are also suitable for representing complex multivariate data. The ability to "layer" data (e.g., include multiple properties) allows rapid visual comparison and ready identification of outliers and clusters. They can be less suitable for discerning fine or exact distinctions between data points or making precise comparisons.

V. Experimental Evaluation

We conducted an experiment to evaluate the performance of the Authorship Portal in identifying and comparing "extreme" Jihadi authors listed on the TurnToIslam online forum. The goal of the experiment was to understand the performance of the portal's visualization functionality: feature highlighting, authorship comparison spider chart, and stylometric heat map.

A. Experimental Setup

We adopted a one-factor repeated-measures approach in our experimental design, which gives a greater precision than designs that employ only between-subjects factors [Myers & Well 1995]. Each subject used the Authorship Portal to perform two sets of tasks. One set of tasks used the portal's visualization functionality. The subject used the portal's functions to answer two or three questions in each of the three parts that required the use of feature highlighting, the authorship comparison spider chart, and the stylometric heat map, respectively. A sample task in Part 1 (using feature highlighting) was "How many times do 'Opinions and Attitudes' words appear in the same sentences or adjacent sentences that also have 'Religion and Faith' words?" A sample task in Part 2 (using the authorship comparison spider chart) was "Between BintMuhammad and Raihan, which author has a higher usage of racism features?" A sample task in Part 3 (using the stylometric heat map) was "Which author has the highest usage of 'Politics and Events' content in their authoring of forum messages?" The other set of tasks, also divided into three parts, used only ordinary message browsing. The questions used in these tasks are similar to those aforementioned, and subjects were not allowed to use the visualization functionality described above.

The whole experiment took approximately 50 minutes and was divided into three sections. The first section (10 minutes) was a tutorial in which the experimenter guided the subjects to use the portal's functionality. The second and third sections (20 minutes each) contained the tasks as described above. The order of the two sets of tasks was assigned randomly to each subject to remove any bias on the results due to a learning effect. Upon finishing all three sections, the subject filled out a short questionnaire asking them to rate (on a five-point Likert Scale) their perception on the usability of the portal's visualization functionality and to provide demographic data.

Thirteen subjects participated in the study. These subjects were students (six males and seven females) enrolled in a business software application course offered by a regional public university in the United States.

B. Performance Measures

1) Accuracy: The accuracy of the task performance was measured by how close the subject's answer was to the correct answer (for tasks in in Part 1 where a number was expected), as shown in the following formula. The accuracies in all tasks in Part 1 were averaged to obtained an overall accuracy of that part.

$$\text{Accuracy}=1-\min\left(\left|\frac{\text{Correct Answer - Subject's Answer}}{\text{Correct Answer}}\right|,1\right)$$

For tasks in Parts 2 and 3 where written responses were expected, the accuracy was calculated by averaging the correctness of each task's performance (correct response = 1, incorrect response = 0).

2) Efficiency: The efficiency was measured by the time elapsed (in minutes) between the beginning of a task and the completion of the task.

3) User Rating: The user rating was measured in a five-point Likert Scale, where 5 indicated "strongly agree" and 1 indicated "strongly disagree."

C. Experimental Results

The accuracy and efficiency of task performance using the Authorship Portal's visualization functions were generally higher than those without using the visualization functions. Table I shows the detailed performance levels and the mean differences in all three parts of each section (using or not using visualization). The figures show that the subjects achieved higher efficiency in all three parts of the study, and obtained higher accuracy in Parts 2 and 3 when they used the visualization functions of the Authorship Portal.

Subjects rated the Authorship Portal very highly. Table II shows their ratings on three statements related to the three visualization functions of the portal. All these ratings were close to the maximum of 5 (strongly agree) along a Likert Scale. In particular, the mean rating of the Authorship Comparison Spider Chart was the highest (4.38) among the three, showing subjects' preference toward a novel visualization of the different writing feature values.

TABLE I. Accuracy and Efficiency of Task Performance

Part	A(V)	A(~V)	Diff_A	E(V)	E(~V)	Diff_E
1	0.79	0.79	0.00	**1.31**	1.50	- 0.19
2	**0.90**	0.74	0.16	**1.46**	2.13	- 0.67
3	**0.77**	0.54	0.23	**1.27**	1.88	- 0.61

Note:
A(V) = Mean accuracy of task performance using visualization
A(~V) = Mean accuracy of task performance without using visualization
Diff_A = Mean difference of A(V) – A(~V)
E(V) = Mean efficiency of task performance using visualization
E(~V) = Mean efficiency of task performance without using visualization
Diff_E = Mean difference of E(V) – E(~V)

TABLE II. SUBJECTS' RATING OF AUTHORSHIP PORTAL

Statement	Mean Rating	S.D.
I find the feature highlighting of the Authorship Portal to be more useful in identifying message features than manually reading the forum messages.	4.15	0.90
I find the authorship comparison spider chart of the Authorship Portal to be more useful in comparing authors' writing than manually reading and comparing the authors' messages.	4.38	0.87
I find the stylometric heat map of the Authorship Portal to be more useful in feature usage than manually reading the forum messages.	4.31	0.85

D. Implications

The positive results shown in the experimental findings illustrate the power of the Authorship Portal's visualization functions. Using these functions, subjects were able to complete the tasks with higher accuracy and less time than using only ordinary message browsing. The results demonstrate the usability of the portal in supporting authorship analysis. The tools may possibly save analysts' time and enhance accuracy in understanding online messages relevant to the task at hand.

VI. CONCLUSIONS AND FUTURE WORK

The Arizona Authorship Analysis Portal was developed to support authorship analysis for terrorism research, cybercrime investigation, and intelligence analysis. The portal presently supports two tasks: identification of "extreme" jihadi authors, and analysis/comparison (similarity detection) of author writing styles. Sentiment analysis incorporating content-specific features with a machine learning classifier was integrated with web-based visualizations for graphical representations of authors' styles. System functions such as searching, browsing, and feature highlighting were incorporated with an interface to allow users to explore authors' styles from a variety of perspectives and contexts.

An evaluation of the text visualizations showed that they supported greater task efficiency and accuracy for task performance.

Future efforts will include a larger testbed of data with investigation of a scalable classifier [e.g., Zimbra and Chen 2012], alternative visualizations, and further evaluation of the system's efficacy for authorship analysis tasks beyond identification of "extreme" authors.

ACKNOWLEDGEMENTS

This work was supported in part by the National Science Foundation (CBET-0730908) and the Defense Threat Reduction Agency (HDTRA1-09-1-0058).

REFERENCES

Abbasi, A., and Chen, H. (2005). "Identification and Comparison of Extremist-Group Web Forum Messages using Authorship Analysis," *IEEE Intelligent Systems* 20:5, pp. 67-75.

Abbasi, A., and Chen, H. (2006). "Visualizing Authorship for Identification." *In* Proceedings of the 4th IEEE Symposium on Intelligence and Security Informatics (ISI 2006), San Diego, CA. Springer, pp. 60-71, May 23-24.

Abbasi, A. and Chen, H. (2008a). "Writeprints: A Stylometric Approach to Identify-Level Identification and Similarity Detection in Cyberspace," *ACM Transactions on Information Systems* 26:2.

Abbasi, A. and Chen, H. (2008b). "CyberGate: A Design Framework and System for Text Analysis of Computer Mediated Communication," *MIS Quarterly*, (32:4), pp. 811-837.

Abbasi, A. Chen, H. and Salem, (2008) "Sentiment Analysis in Multiple Languages : Feature Selection for Opinion Classification in Web Forums," *ACM Transactions on Information Systems,* 26:3.

Argamon, S., Koppel, M., Pennebaker, J., and Schler, J. (2009). "Automatically Profiling the Author of an Anonymous Text," *Communications of the ACM* **52** (2): 119–123 (virtual extension).

Balahur, A. Hermida, J. M. Montoyo, Andres. (2010) Detecting implicit expressions of emotion in text: A comparative analysis. *Decision Support Systems*. 53:4, pp. 742-753.

Chen, H. (2011). *Dark Web: Exploring and Data Mining the Dark Side of the Web*. Integrated Series in Information Systems. Springer.

Chen, H., Denning, D., Roberts, N., Larson, C.A., Yu, X., and Huang, C.-N. (2011). "Dark Web Forum Portal." *In* Proceedings of the 9th IEEE Symposium on Intelligence and Security Informatics (ISI 2011),

Chuang, J. Ramage, D. Manning, C. D. Heer, J. (2012) Interpretation and Trust: Designing Model-Driven Visualizations for Text Analysis. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 473-482.

Chung, W., Chen, H., & Nunamaker, J. F. (2005). A visual framework for knowledge discovery on the Web: An empirical study on business intelligence exploration. *Journal of Management Information Systems, 21*(4), 57-84.

Chung, W. (2008). Visualizing E-Business Stakeholders on the Web: A Methodology and Experimental Results. *International Journal of Electronic Business*, 6(1), 25-46.

De Vel, O., Anderson, A., Corney, M., & Mohay, G. (2001). "Mining e-mail content for author identification forensics." *SIGMOD Record* 30:4, pp. 55-64.

Frantzeskou, G. and Gritzalis, Stefanos. (2004). "Source Code Authorship Analysis for Supporting the Cybercrime Investigation Process". *In* Proceedings of the 1st International Conference on E-Business and Telecommunication Networks, Setúbal, Portugal, INSTICC Pr., pp. 85-92.

Fu, T.J., Abbasi, A. and Chen, H. (2010). "A focused crawler for Dark Web forums," *Journal of the American Society for Information Science and Technology*, 61:6, pp. 1213-1231.

Layton, R., Watters, P., Dazeley, R. (2012). "Unsupervised authorship analysis of phishing webpages," *In* Communications and Information Technologies (ISCIT), 2012 International Symposium on, Oct. 2-5, pp. 1104-1109.

McEnery, A., and Oakes, M. (2000). "Authorship Studies and Computational Stylometry" in Dale, R., Marcel Dekker.

Myers, J., and A. Well (1995) *Research Design and Statistical Analysis*. Hillsdale, NJ, USA: Lawrence Erlbaum Associates, Publishers.

Urvoy, T., Chauveau, E., Filoche, P., and Lavergne, T. (2008). "Tracking Web spam with HTML style similarities." ACM Transactions on the Web, 2:1.

Wise, J. A., Thoma, J. J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., & Crow, V. (1995). Visualizing the non-visual: spatial analysis and interaction with information from text documents. Paper presented at the IEEE, Proceedings of Information Visualization.

Zhang, Y. and Zeng, S. et al., (2010). "Developing a Dark Web collection and infrastructure for computational and social sciences," Intelligence and Security Informatics (ISI), 2010 IEEE International Conference on Vancouver, BC, Canada, pp. 59 - 64.

Zheng, R., Li, J., Chen, H., and Huang, Z. (2006). "A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques." *Journal of the American Society for Information Science and Technology*, 57:3, pp. 378-393.

Zimbra, D., and Chen, H. (2012). "Scalable Sentiment Classification Across Multiple Dark Web Forums." *In* Proceedings of the 2012 IEEE International Conference on Intelligence and Security Informatics (ISI 2012), Washington, D.C., pp. 78-83.