Data Mining Instant Messaging Communications to Perform Author Identification for Cybercrime Investigations

Angela Orebaugh and Dr. Jeremy Allnutt

MS 1G5, 4400 University Drive, Fairfax, VA 22030-4444
angela@securityknox.com, jallnutt@ece.gmu.edu

Abstract. Instant messaging is a form of computer-mediated communication (CMC) with unique characteristics that reflect a realistic presentation of an author's online stylistic characteristics. Instant messaging communications use virtual identities, which hinder social accountability and facilitate IM-related cybercrimes. Criminals often use virtual identities to hide their true identity and may also supply false information on their virtual identities. This paper presents an IM authorship analysis framework and feature set taxonomy for use in cyber forensics and cybercrime investigations. We explore authorship identification of IM messages to discover the parameters with the highest accuracy for determining the identity of a cyber criminal.

Keywords: Cybercrime investigations, cyber forensics, authorship analysis, forensic data mining.

1 Introduction

The Internet has evolved from a resource of simple information sharing and exchange to a mecca of virtual communications and e-commerce activities. One increasingly popular use of the Internet is computer-mediated communication (CMC). CMC includes any communicative transaction, which occurs through the use of two or more networked computers. [McQ2005] Instant messaging (IM) is a form of synchronous CMC that occurs in real time and requires the simultaneous participation of users. Examples of IM services include AOL Instant Messenger (AIM), I Seek You (ICQ), Skype, MSN Messenger, Google Talk, Jabber, and Yahoo! Messenger. Users register themselves with the service provider and download a compatible client for use on the service provider network. By registering, the user creates an account that consists of a unique identifier such as a name or number, also called a *screen name*. The screen name and its associated information become the user's virtual identity.

1.1 IM and Cybercrime

IM is growing in popularity and is becoming increasingly used for both personal and business communications, often replacing e-mail in certain environments and situations. This explosive growth in the use of IM communication in both personal and

S. Goel (Ed.): ICDF2C 2009, LNICST 31, pp. 99–110, 2010.

professional environments has resulted in an increased risk to proprietary, sensitive, and personal information and safety due to the influx of IM-related cybercrimes, such as phishing, social engineering, threatening, cyber bullying, hate speech and crimes, child exploitation, sexual harassment, and illegal sales and distribution of software. [MD2000] IM is also used as a communication channel for gangs, terrorists, and cyber intruders. [AC2005] The anonymous nature of the Internet and use of virtual identities hinder social accountability and present a critical challenge for cybercrime investigations. Criminals use IM virtual identities to hide their true identity or impersonate other users and may also supply false information on their virtual identities. They can use multiple screen names or impersonate other users with the intention of harassing or deceiving unsuspecting victims. Criminals may also supply false information on their virtual identities, for example a male user may configure his virtual identify to appear as female. New cyber forensics methods are needed to identify cyber criminals, discover criminals who supply false information in their virtual identities, and collect digital evidence for cybercrime investigation. The objective is to collect sufficient and accurate digital evidence for prosecution.

Individuals often leave behind textual identity traces in cyberspace. [AC2008] Determining an IM user's real identity relies on the fact that humans are creatures of habit and have certain persistent personal traits and patterns of behavior, known as behavioral biometrics. [Rev2008] Behavioral biometrics are measurable traits that are acquired over time (versus a physiological characteristic or physical trait) and used to recognize or verify the identity of a person. [Bio2006] For example, handwriting style is consistent throughout a person's life, even though it may vary with age. As with handwriting, user's have certain online writing habits that are unconscious and deeply ingrained. [TLML2004] Even if a criminal made a conscious effort to disguise his/her style, it would be difficult to achieve. Online writing habits include composition syntax and layout, vocabulary patterns, unique language usage, and other stylistic traits. Thus, certain user behavior characteristics that remain relatively constant may be used to identify an author of a particular piece of work. [DACM2001b]

Cyber forensic investigations for instant messaging rely on instant messaging exchanges, or conversations, as digital evidence. The accurate identification of an author of the IM messages in a cybercrime is essential for a successful prosecution. Authorship analysis techniques can be used to discover behavior biometrics to identify an author, as well as certain characteristics of the author, of IM messages.

1.2 Authorship Analysis

Authorship analysis is the process of examining the characteristics of a document to find or validate the document's author. Authorship analysis is based on a linguistic research area called stylometric analysis and is sometimes referred to as authorship attribution. [BVT1996, Hol1994, Rud1998]. Authorship analysis can be divided into three categories. [ZLCH2006] ***Authorship identification*** attempts to determine the author of a piece of text by examining other text samples that have been authenticated as having been produced by that author. [Cha2005] ***Authorship characterization*** categorizes an author's text according to sociolinguistic attributes such as gender, age, educational background, income, linguistic background, nationality, profession, psychological status, and race. These attributes are aimed at inferring an author's background characteristics rather than identity. ***Similarity detection*** compares multiple sample texts and determines whether they were produced by a single author without actually identifying the author. [ZLCH2006] This is used mostly for the purpose of detecting plagiarism [GHM2005]. This paper is focused in authorship identification techniques.

Authorship analysis uses a variety of writing style features that can be derived from a particular piece of work to facilitate authorship analysis. Four feature categories have evolved for CMC. ***Lexical*** features include the total number of words, number of words per sentence, word length distribution, *vocabulary richness*, average characters per sentence, average characters per word, and character usage frequency. [AC2006] ***Syntactic*** features include patterns used for the formation of sentences, including punctuation and *function words*. [AC2006] ***Structural*** features include the organization and layout of text including the use of greetings and signatures and the number of paragraphs and average paragraph length. [AC2006] ***Content specific*** features include key words that are used within a specific topic area. [AC2006] For example, a criminal selling illegal merchandise will use certain terms related to the items. This paper uses lexical, syntactic, and structural features.

1.3 IM Authorship Analysis and Cyber Forensics

Researchers have begun to use authorship analysis as a cyber forensics tool, with recent application to e-mail [DACM2001b], [ASS2003], online forums (i.e. discussion groups or newsgroups) [ZLCH2006], [AC2005], [ZLCH2006], program code [GSM1997], and online chat [AC2008]. As a cyber forensics tool, authorship analysis may be used to identify the most plausible author of an IM conversation from a group of suspects and to gather convincing digital evidence to support the finding.

The style of IM messages is very different than that of any other text used in traditional literature or other forms of computer-mediated communication. The real time nature of IM messages produces unedited text that reflects the author's true writing style and vocabulary. [KCAC2008] The textual nature of IM also creates a unique need to exhibit emotions. Emotion icons, called emoticons, are sequences of punctuation marks commonly used to represent feelings within computer-mediated text. [KCAC2008] IM also includes other special linguistic elements such as abbreviations, and computer and Internet terms, known as netlingo. These characteristics of IM allow authors to have a unique online writing style, known as a writeprint, which can assist researchers and investigators in discovering the author's true identity. Because IM conversations are relatively casual compared with formal text, an author's IM writeprint is more likely to represent an authors true stylistic habits. An author's writing style is usually consistent across his/her writings, thus an author may be identified by matching his/her writeprint with the writeprint of the text in question. [LZC2006]

Writeprints provide computer forensic experts a unique tool for identifying criminals in CMC. The principle challenge with writeprint analysis is the creation of a set of features that represent an author's stylistic traits with the highest accuracy. Certain IM specific features such as message structure, unusual language usage, and unusual stylistic markers are useful in forming a suitable writeprint feature set for authorship analysis. [ZLCH2006] Thus far, the research community has performed a number of studies in authorship analysis for e-mails and other CMC, but none have studied IM authorship analysis in a comprehensive, systematic way.

1.4 Related Works

There has been significant research in identifying authors of literary texts, such as Shakespeare's works and The Federalist Papers. [Men1887, MW1964] Some of the earliest research dates back to the fourth century BC, when librarians in the library of Alexandria studied the authentication of texts attributed to Homer. [Lov2002] Other early known research dates back to the 18th century when English logician Augustus de Morgan theorized that authorship can be determined by examining if one text contains more longer words than another.

Recent research has introduced authorship analysis to CMC including e-mail, online newsgroups, and chat groups, with promising results. Recent applications of authorship analysis in computer-mediated communications include several exceptional publications by Olivier de Vel [DeV2000, DACM2001a, DACM2001b] that studied classification of e-mail documents for the purpose of authorship identification. Another important contribution to the study of authorship analysis for cyber forensics includes the research by Rong Zheng, Yi Qin, Zan Huang, Hsinchun Chen [ZQHC2003, ZLCH2006]. The authors presented a comparison of techniques to automate author identification by using several classification algorithms to extract features such as style markers, structural features, and content-specific features. A unique visualization authorship analysis study includes the research by Ahmed Abbasi and Hsinchun Chen [AC2006, AC2008]. This research introduced an authorship writeprint visualization technique that can assist in identifying online authors based on their writing style. Another writeprint authorship analysis study includes the research by Jiexun Li, Rong Zheng, Hsinchun Chen [LZC2006]. This research introduced "a method of identifying the key writeprint features for authors of online messages to facilitate identity tracing in cybercrime investigation." [LZC2006] A study that performs both authorship identification and authorship characterization includes the research by Tayfun Kucukyilmaz, B. Barla Cambazoglu, Cevdet Aykanat, and Fazli Can [KCAC2008]. This research performs classification of online chat messages to determine both author attributes (gender, age, educational environment, and Internet connection domain) and message attributes (author, receiver, and time of day). A large research gap exists in applying authorship analysis techniques to instant messaging communications to determine the author identity. There are no known studies that present a comprehensive examination of IM authorship analysis.

2 Research Methodology

We propose an IM authorship analysis framework, shown in Figure 1, that extracts features from the messages to create author writeprints and applies several data mining algorithms to build classification models to perform automated authorship analysis. The parameters are systematically modified in an iterative process to assess their impact on the prediction accuracy of the classification methods. The goal of the framework is to identify the set of features, classification algorithm, number of authors, and number of messages with the highest prediction accuracy.

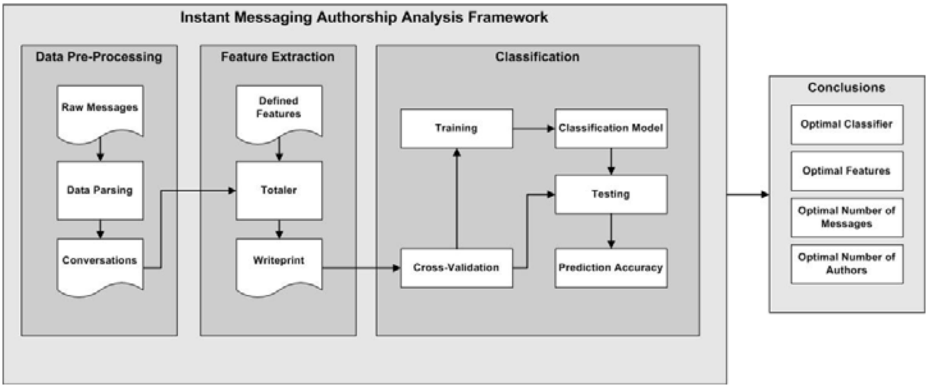

Fig. 1. IM Authorship Analysis Framework

The framework consists of three stages: data pre-processing, feature extraction, and classification. The data pre-processing stage parses the messages to extract the data for each particular author and to remove metadata and noise, such as timestamps, usernames, and automatic away message responses. Next, the logs are input into the extractor.pl Perl program. The extractor first splits the logs into a configurable conversation size, such as 50 messages per conversation.

The feature extraction stage inputs conversations and pre-selected features to the extractor.pl program totaler module to create totals for each feature, resulting in the output of a writeprint (W_n) for each set of messages {M_1,…,M_n} of each supplied author (A_n). A writeprint is an *n*-dimensional vector, where *n* represents the total number of features. Each writeprint is assigned a class, which is the author (A_n) of the writeprint (W_n). The extractor.pl program then outputs a writeprint in comma-separated value (CSV) format for input into the WEKA data mining toolkit.

The classification stage uses the writeprints as input to build classification models and resulting prediction accuracies. This stage uses the WEKA data mining toolkit to perform classification on the writeprints using the C4.5, k-nearest neighbor, Naïve Bayes, and SVM classifiers. The cross-validation module supplies training and testing instances to the training module and testing module, respectively. The training module creates a classification model and the testing module calculates the prediction accuracy of the classification model. To evaluate the effectiveness of the prediction, the framework uses the average prediction accuracy produced by the testing module.

There are a number of parameters that impact the prediction accuracy performance of the data mining classifiers. To assess the impact of the parameters, the framework is repeated with varying numbers of authors and messages, features, and classification models.

2.1 Feature Set Taxonomy

A feature set is composed of a predefined set of measurable writing style attributes. This research presents a feature set taxonomy of instant messaging writing style characteristics for performing authorship analysis for the purpose of cyber forensics. The proposed feature set is a 356-dimensional vector including lexical, syntactic, and

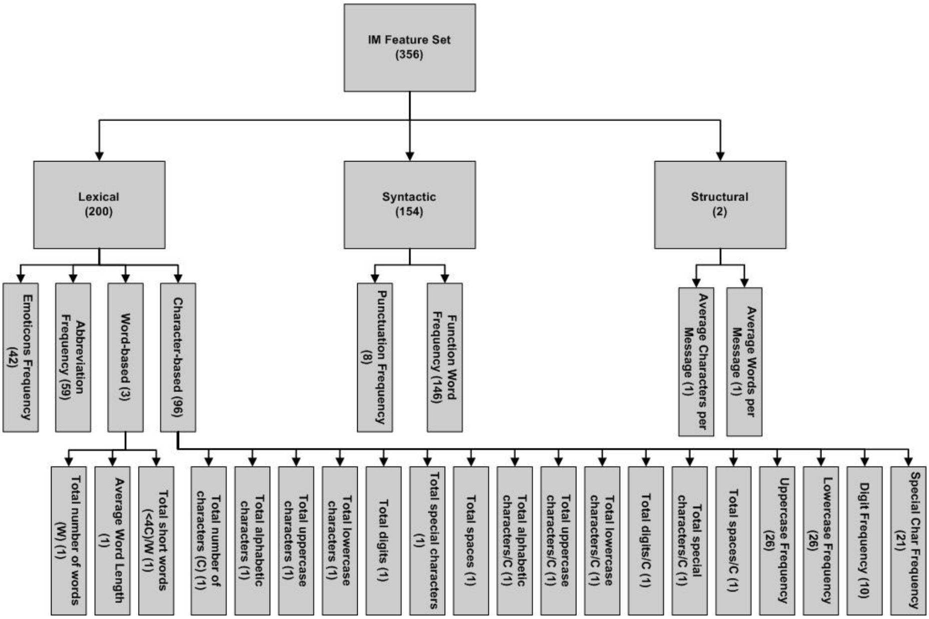
IM Feature Set (356)
Lexical (200)
Syntactic (154)
Structural (2)
Emoticons Frequency (42)
Abbreviation Frequency (59)
Word-based (3)
Character-based (96)
Punctuation Frequency (8)
Function Word Frequency (146)
Average Characters per Message (1)
Average Words per Message (1)
Total number of words (W) (1)
Average Word Length (1)
Total short words (<4C)/W (1)
Total number of characters (C) (1)
Total alphabetic characters (1)
Total uppercase characters (1)
Total lowercase characters (1)
Total digits (1)
Total special characters (1)
Total spaces (1)
Total alphabetic characters/C (1)
Total uppercase characters/C (1)
Total lowercase characters/C (1)
Total digits/C (1)
Total special characters/C (1)
Total spaces/C (1)
Uppercase Frequency (26)
Lowercase Frequency (26)
Digit Frequency (10)
Special Char Frequency (21)

Fig. 2. IM Feature Set Taxonomy

structural features, shown in Figure 2. Content specific features are highly dependent on the topic of the messages; therefore the feature set taxonomy does not include content specific features in order to achieve generic authorship identification across various applications.

Instant messaging communications have several characteristics that are useful in forming a well-rounded feature set, which may help reveal the writing style of the author. The IM feature set taxonomy includes several new stylistic features, such as abbreviations and emoticons, which are frequently found in instant messaging communications. Lexical features mainly consist of frequency totals and are further broken down into emoticons, abbreviations, word-based, and character-based features. Syntactic features include punctuation and function words in order to capture an author's habits of organizing sentences. Structural features capture the way an author organizes the layout of text. With IM communications there are no standard headers, greetings, farewells, or signatures, leaving simply the average characters and words per message in terms of structural layout. Each feature in the taxonomy was selected for its relevance to IM communications. The goal of the IM feature set taxonomy is to develop a streamlined set of features that best reveal the true writing style of the author.

3 Experiment Results and Analysis

There are no standard data sets used for instant messaging research. In addition, due to privacy and legal reasons, obtaining IM communication logs is very difficult. This research uses two datasets: a personally collected dataset of known authors and a publicly available dataset, with 19 and 105 authors respectively. Both datasets contain unedited messages communicated between two users. Dataset #1 contains personal IM conversation logs collected by the Gaim and Adium clients over a three-year period. It includes complete IM logs from 19 authors. Dataset #2 contains publicly available data from U.S. Cyberwatch[1]. U.S. Cyberwatch aims to assist law enforcement with the interception, apprehension, and prosecution of online child predators. U.S. Cyberwatch data was collected from April 2004 to March 2007. The dataset includes complete IM logs between undercover agents and 100 different child predators. The U.S. Cyberwatch data is an example of real world cybercrime digital evidence that cyber forensics investigators are obtaining, analyzing, and presenting in court proceedings.

Dataset #1 experiments include 19 classes (authors) from which to determine identification. The data was undersampled to 500 messages per author to create a balanced dataset. Table 1 shows the highest accuracy and associated parameters for each feature category for Dataset #1. The highest overall accuracy (88.42%) for all 19 authors was achieved using the entire set of 356 features. The optimal algorithm is SVM and the optimal number of messages per each instance is 50. In Dataset #1, digits were rarely used and were not useful for authorship identification. These results indicate that when using the entire features set along with the SVM algorithm that an author of a given set of messages can be identified with a sufficiently high degree of accuracy.

Table 1. Dataset #1 Top Results for Feature Categories

Features	Number of Features	Highest Accuracy	Algorithm	Number of Messages
All Features	**356**	**88.42%**	**SVM**	**50**
Lexical	**200**	**78.26%**	**SVM**	**50**
Syntactic	**154**	**68.32%**	**SVM**	**100**
Structural	**2**	**34.21%**	**k-NN**	**250**
Emoticon Frequency	**42**	**38.95%**	**NB**	**100**
Abbreviation Frequency	**59**	**47.37%**	**NB**	**50**
Word-based Lexical	**3**	**47.37%**	**k-NN**	**125**
Character-based Lexical	**96**	**70.00%**	**SVM**	**50**
Uppercase Frequency	**26**	**46.32%**	**NB**	**100**
Lowercase Frequency	**26**	**44.74%**	**k-NN**	**250**
Digit Frequency	**10**	**11.05%**	**k-NN**	**50**
Special Character Frequency	**21**	**34.21%**	**C4.5**	**125**
Punctuation Frequency	**8**	**67.37%**	**NB**	**100**
Function Word Frequency	**146**	**60.00%**	**SVM**	**100**
CHI2 Top Ten	**10**	**80.46%**	**k-NN**	**125**
CHI2 Top Five	**5**	**64.70%**	**k-NN**	**125**

[1] For more information please refer to: http://www.uscyberwatch.com

Additional experiments were performed on various combinations of feature categories for Dataset #1. The feature combination of Character-based Lexical and Punctuation performs fairly well at 81.81% with only 104 features. In this case the optimal algorithm is SVM and the optimal number of messages per each instance is 100.

The next set of experiments evaluates the size of the set of author suspects $\{A_1,\ldots,A_n\}$ by incrementing the number of authors by one for each classification test. Each test used 50 messages per author and performed classification with each of the four different algorithms. The C4.5 algorithm performed best with an author suspect set of 5 or less. SVM performed best on author suspect sets greater than 5. Figure 5 shows a graphical comparison of each classification algorithm's prediction accuracy when incrementing authors. The SVM algorithm's accuracy remained rather constant as the number of authors increased. Overall there was a 13% drop in accuracy when increasing the number of authors from 2 to 19.

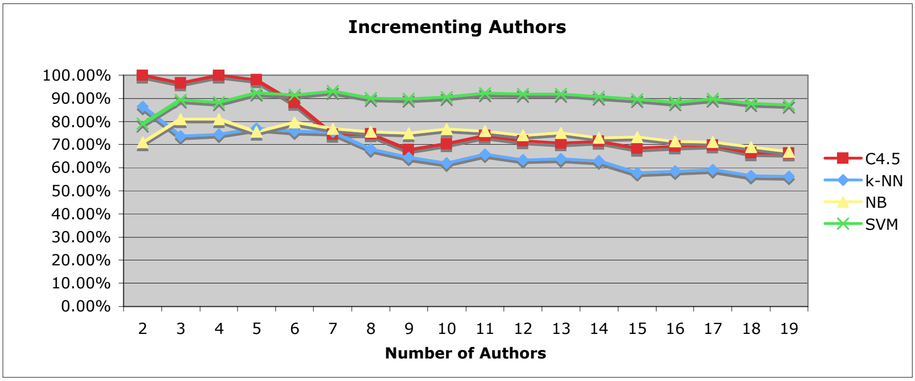

Fig. 3. Dataset #1 Incrementing Authors Accuracy Results Graph

Fig. 4 depicts a graphical comparison between several experiments with varying numbers of authors and messages. It shows the accuracy results for experiments on all 100 authors, the Top 50 authors, and the Top 25 authors. By creating data subsets, a larger message set $\{M_1,\ldots,M_n\}$ can be used when balancing the data. The graph shows accuracy comparisons for author sets $\{A_1,\ldots,A_n\}$ of various sizes. The results show that as the size of the message set $\{M_1,\ldots,M_n\}$ per author increases, the accuracy significantly increases for author suspect sets $\{A_1,\ldots,A_n\}$ of equal size. Overall the best results are achieved with datasets of 25 authors or less and 500 messages or more for each author.

Table 2 shows the highest accuracy and associated parameters for each feature category for Dataset #2 Top 25 authors. The highest overall accuracy (84.44%) for all 25 authors was achieved using the entire set of 356 features. The optimal algorithm is SVM and the optimal number of messages per each instance is 50. For Dataset #2, emoticons performed poorly because they were recorded as blank lines in the raw data and were removed. Thus, we were unable to test the true use of emoticons in this dataset. These results indicate again, that when using the entire feature set along with the SVM algorithm that an author of a given set of messages can be identified with a sufficiently high degree of accuracy.

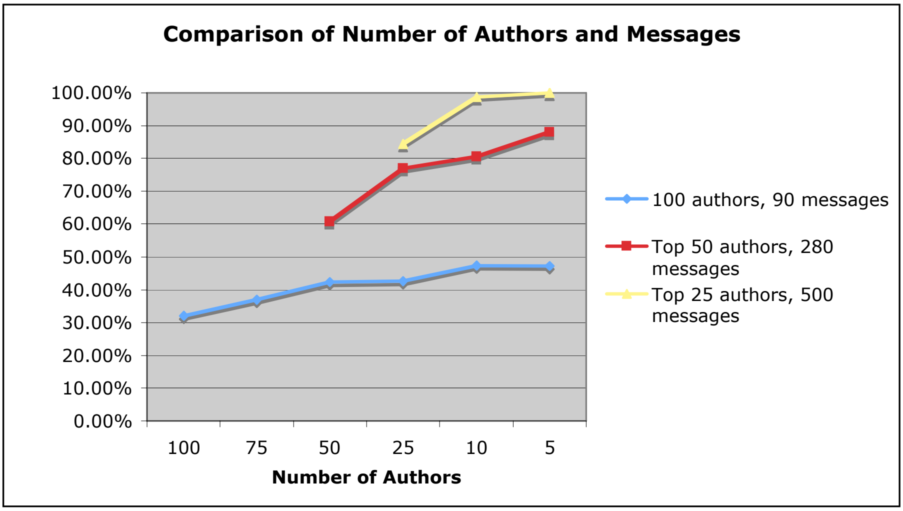

Fig. 4. Dataset #2 Accuracy Comparison Table

Table 2. Dataset #2 Top Results for Feature Categories

Features	Number of Features	Highest Accuracy	Algorithm	Number of Messages
All Features	356	84.44%	SVM	50
Lexical	200	74.84%	SVM	50
Syntactic	154	64.56%	NB	50
Structural	2	22.72%	NB	100
Emoticon Frequency	42	8.00%	C4.5/k-NN	50
Abbreviation Frequency	59	39.80%	NB	125
Word-based Lexical	3	32.98%	k-NN	100
Character-based Lexical	96	64.80%	NB	50
Uppercase Frequency	26	34.10%	k-NN	125
Lowercase Frequency	26	47.12%	NB	50
Digit Frequency	10	16.09%	NB	100
Special Character Frequency	21	21.60%	NB	125
Punctuation Frequency	8	62.50%	NB	125
Function Word Frequency	146	54.20%	SVM	50
CHI2 Top Ten	10	71.40%	NB	50
CHI2 Top Five	5	46.84%	NB	50

Additional experiments were performed on various combinations of feature categories for Dataset #2 Top 25 authors. The feature combination of Lexical and Syntactic performs best at 84.32%. In this case the optimal algorithm is SVM and the optimal number of messages per each instance is 50. Character-based Lexical and Punctuation performs fairly well at 76.44% with only 104 features. In this case the optimal algorithm is NB and the optimal number of messages per each instance is 50.

The next set of experiments evaluates the size of the set of author suspects $\{A_1,\ldots,A_n\}$ by incrementing the number of authors by one for each classification test. Each test used 50 messages per author and performed classification with each of the four different algorithms. The NB algorithm performed best with 2 author suspects and SVM performed best on author suspect sets greater than 2. Fig. 5 shows a graphical comparison of each classification algorithm's prediction accuracy when incrementing authors. The SVM algorithm's accuracy remained rather constant as the number of authors increased. Overall there was a 15% drop in accuracy when increasing the number of authors from 2 to 25.

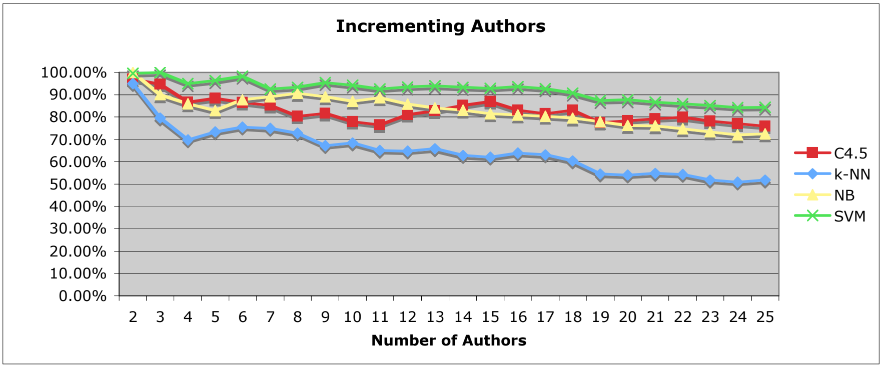

Fig. 5. Dataset #2 Incrementing Authors Accuracy Results Graph

4 Conclusions and Future Work

In this paper we have applied authorship analysis techniques to the CMC medium of instant messaging to perform authorship identification of IM messages to assist cyber forensics and cybercrime investigation. We have developed a formalized IM authorship analysis framework to analyze the prediction accuracy of four data mining algorithms with varying parameter settings in a systematic way. We have also created a holistic IM-specific feature set taxonomy that may be easily utilized in future research. Our experiments achieved authorship identification prediction accuracies of 88.42% and 84.44% for Dataset #1 (19 authors) and Dataset #2 (25 authors) respectively. We are continuing this research by expanding the datasets, exploring other classification techniques, and including author characterization techniques to narrow the field of suspects in a cybercrime investigation.

References

1. Abbasi, A., Chen, H.: Applying Authorship Analysis to Extremist-Group Web Forum Messages. IEEE Intelligent Systems 20(5), 67–75 (2005)
2. Abbasi, A., Chen, H.: Visualizing Authorship for Identification. Proceedings of the Intelligence and Security Informatics. In: IEEE International Conference on Intelligence and Security Informatics (2006)
3. Abbasi, A., Chen, H.: Writeprints: A Stylometric Approach to Identify-Level Identification and Similarity Detection in Cyberspace. ACM Transactions on Information Systems 26(2) (2008)
4. Argamon, S., Saric, M., Stein, S.S.: Style mining of electronic messages for multiple authorship discrimination: First results. In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2003)
5. BioPassword.: Authentication Solutions Through Keystroke Dynamics (2006)
6. Baayen, R.H., van Halteren, H., Tweedie, F.: Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution. Literary and Linguistic Computing 11(3) (1996)
7. Chaski, C.E.: Who's At The Keyboard? Authorship Attribution in Digital Evidence Investigations. International Journal of Digital Evidence 4(1) (2005)
8. de Vel, O., Anderson, A., Corney, M., Mohay, G.: Multi-Topic E-mail Authorship Attribution Forensics. In: ACM Conference on Computer Security - Workshop on Data Mining for Security Applications, Philadelphia, PA, USA (2001)
9. de Vel, O., Anderson, A., Corney, M., Mohay, G.: Mining E-mail Content for Author Identification Forensics. SIGMOD Record Web Edition 30(4) (2001)
10. de Vel, O.: Mining E-mail Authorship. In: KDD 2000 Workshop on Text Mining, Boston, Massachusetts, USA, pp. 21–27 (2000)
11. Graham, N., Hirst, G., Marthi, B.: Segmenting documents by stylistic character. Natural Language Engineering 11(4), 397–415 (2005)
12. Gray, A., Sallis, P., Macdonnel, S.: Software forensics: Extended authorship analysis techniques to computer programs. In: Proceedings of the 3rd Biannual Conference on the International Association of Forensic Linguists (1997)
13. Holmes, D.I.: Authorship Attribution. Computers and the Humanities 28(2) (1994)
14. Kucukyilmaz, T., Cambazoglu, B.B., Aykanat, C., Can, F.: Chat mining: predicting user and message attributes in computer-mediated communication. Information Processing & Management 44(4), 1448–1466 (2008)
15. Love, H.: Attributing authorship: an introduction. Cambridge University Press, Cambridge (2002)
16. Li, J., Zheng, R., Chen, H.: From fingerprint to writeprint. Commun. ACM 49(4), 76–82 (2006)
17. McQuail, D.: McQuail's Mass Communication Theory, 5th edn. SAGE Publications, London (2005)
18. Moores, T., Dhillon, G.: Software Piracy: A View from Hong Kong. Communications of the ACM 43(12), 88–93 (2000)
19. Mendenhall, T.C.: The Characteristic Curves of Composition. Science 11(11), 237–249 (1887)
20. Mostellar, F., Wallace, D.: Inference and Disputed Authorship: The Federalis. Addison-Wesley, Reading (1964)
21. Revett, K.: Behavioral Biometrics: A Remote Access Approach. John Wiley & Sons, Ltd., Chichester (2008)

22. Rudman, J.: The state of authorship attribution studies: some problems and solutions. Computers and the Humanities 31(4) (1998)
23. Teng, G., Lai, M., Ma, J., Li, Y.: E-mail Authorship Mining Based on SVM for Computer Forensic. In: Proceedings of the Third International Conference on Machine Learning and Cybernetics, Shanghai (2004)
24. Zheng, R., Li, J., Chen, H., Huang, Z.: 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), 378–393 (2006)
25. Zheng, R., Qin, Y., Huang, Z., Chen, H.: Authorship Analysis in Cybercrime Investigation. In: Chen, H., Miranda, R., Zeng, D.D., Demchak, C.C., Schroeder, J., Madhusudan, T. (eds.) ISI 2003. LNCS, vol. 2665, pp. 59–73. Springer, Heidelberg (2003)