Stylometric Identification in Electronic Markets: Scalability and Robustness

AHMED ABBASI, HSINCHUN CHEN, AND JAY F. NUNAMAKER JR.

AHMED ABBASI is a Professor in the Sheldon B. Lubar School of Business at the University of Wisconsin–Milwaukee. He received his Ph.D. in Management Information Systems from the University of Arizona and an MBA and B.S. in Information Technology from Virginia Tech. His research interests include application of text mining and information visualization techniques for improved online trust and analysis within electronic markets and computer-mediated communication. His research has appeared in *IEEE Intelligent Systems, ACM Transactions on Information Systems,* and various conferences, including the ACM/IEEE Joint Conference on Digital Libraries.

HSINCHUN CHEN is McClelland Professor of Management Information Systems at the University of Arizona. He received a B.S. from the National Chiao-Tung University in Taiwan, an MBA from SUNY Buffalo, and a Ph.D. in Information Systems from New York University. Dr. Chen is a Fellow of IEEE and AAAS. He received the IEEE Computer Society 2006 Technical Achievement Award. He is author/editor of 13 books, 17 book chapters, and more than 130 Science Citation Index journal articles covering digital library, intelligence analysis, biomedical informatics, data/text/Web mining, knowledge management, and Web computing. He serves on ten editorial boards and has served as a scientific counselor/advisor of the National Library of Medicine. He has been an advisor for major research programs in digital library, digital government, medical informatics, and national security research.

JAY F. NUNAMAKER JR. is Regents and Soldwedel Professor of MIS, Computer Science and Communication, and Director of the Center for the Management of Information at the University of Arizona, Tucson. He received his Ph.D. in systems engineering and operations research from Case Institute of Technology, an M.S. and B.S. in engineering from the University of Pittsburgh, and a B.S. from Carnegie Mellon University. Dr. Nunamaker received the LEO Award from the Association of Information Systems at ICIS in Barcelona, Spain, December 2002. This award is given for a lifetime of exceptional achievement in information systems. He was elected as a fellow of the Association of Information Systems in 2000. Dr. Nunamaker has over 40 years of experience in examining, analyzing, designing, testing, evaluating, and developing information systems. He has served as a test engineer at the Shippingport Atomic Power facility, as a member of the ISDOS team at the University of Michigan, and as a member of the faculty at Purdue University, prior to joining the faculty at the University of Arizona in 1974. His research on group support systems addresses behavioral as well as engineering issues and focuses on theory as well as implementation. He has been a licensed professional engineer since 1965.

ABSTRACT: Online reputation systems are intended to facilitate the propagation of word of mouth as a credibility scoring mechanism for improved trust in electronic

Journal of Management Information Systems / Summer 2008, Vol. 25, No. 1, pp. 49–78.

0742–1222 / 2008 $9.50 + 0.00.
DOI 10.2753/MIS0742-1222250103

marketplaces. However, they experience two problems attributable to anonymity abuse—easy identity changes and reputation manipulation. In this study, we propose the use of stylometric analysis to help identify online traders based on the writing style traces inherent in their posted feedback comments. We incorporated a rich stylistic feature set and developed the Writeprint technique for detection of anonymous trader identities. The technique and extended feature set were evaluated on a test bed encompassing thousands of feedback comments posted by 200 eBay traders. Experiments conducted to assess the scalability (number of traders) and robustness (against intentional obfuscation) of the proposed approach found it to significantly outperform benchmark stylometric techniques. The results indicate that the proposed method may help militate against easy identity changes and reputation manipulation in electronic markets.

KEY WORDS AND PHRASES: anti-aliasing, electronic markets, online trust, similarity detection, stylometry.

ELECTRONIC MARKETS HAVE SEEN UNPRECEDENTED GROWTH in recent years. Online auction marketplaces such as eBay are one type of electronic market that has become especially popular. However, the lack of physical contact and prior interaction makes such places more susceptible to opportunistic member behavior [40]. While reputation systems attempt to alleviate some of the troubles with electronic markets, these systems suffer from two problems—easy identity changes and reputation manipulation. Easy identity changes stem from the fact that online traders can create new identities, thereby refreshing their reputation [10]. Reputation manipulation allows online market traders to inflate their reputations using multiple identities or to sabotage competitors' reputation scores. Consequently, fraud and deception are highly prevalent in electronic markets, particularly in online auctions, which account for 50 percent of Internet fraud [9].

The aforementioned problems stem from online anonymity. However, individuals leave behind textual traces of their identity in the feedback comments posted to other traders. Stylometric similarity detection techniques applied to reputation system feedback comments can help minimize problems stemming from anonymity abuses in reputation systems. These techniques attempt to assess the degree of similarity between individuals based on writing style. Since text traces are often the only identity cues left behind in cyberspace, researchers have begun to use online stylometric analysis techniques as a forensic tool. They have recently been applied to e-mail, Web forums, and program code [11, 19, 49], as well as group support system comments [21, 22].

Despite significant progress, online stylometry has several current limitations. Most previous work focused on the identification task (where potential authorship identities are known in advance). There has been limited evaluation of similarity detection techniques where no identities are known a priori, and are clustered based on their similarity scores. Similarity detection is more practical for cyberspace applications, such as reputation systems. Furthermore, there has been a lack of evaluation of the scalability of stylometric analysis in terms of number of authors and identities per author for reputation systems. Moreover, there has been a lack of assessment of ro-

bustness against intentional stylistic alteration and message copycatting or forging. In this study, we propose a system that can provide stylometric analysis scalability and robustness for identifying traders in online reputation systems based on their feedback comments posted for others. The proposed system is highly accurate at differentiating across hundreds of identities based on stylistic tendencies inherent in feedback comments, and is also fairly robust against intentional stylistic alteration. The system uses an extended feature set consisting of several static and dynamic feature categories and also includes the Writeprint technique, which assesses the degree of stylistic similarity and dissimilarity between authors. Writeprint uses Karhunen–Loeve transforms to assess the degree of similarity between traders and a pattern disruption mechanism to determine stylistic dissimilarity. The system can be used for similarity detection in reputation systems to alleviate the identity change and rank manipulation problems.

Related Work

Reputation Systems/Online Feedback Mechanisms

Reputation systems are online feedback mechanisms where users rate other members and provide textual comments describing the quality of service (i.e., transaction experience). Such systems are intended to provide "soft security" for electronic markets and online auctions [43]. In contrast to "hard security" systems (e.g., access control/authentication), these systems are designed to offer social control mechanisms. They are meant to allow social translucence for improved accountability [13]. Online markets rely on such information provided via reputation systems in order to promote trust [5]. While recommender systems are designed to support collaborative filtering, reputation systems are intended to support "collaborative sanctioning" [38]. As Josang et al. pointed out, "the purpose is to sanction poor service providers, with the aim of giving an incentive for them to provide quality services" [25, p. 10].

The perceived effectiveness of online feedback mechanisms plays a critical role in the amount of member trust in the community [40]. Reputation scores are often synonymously referred to as "trust scores." An important class of trust is "identity trust," which describes the belief that an identity is who they claim to be [18]. Trustworthiness is an important factor affecting online market outcomes [6]. Identity trust is especially crucial to the success of reputation systems. However, the anonymous nature of the Internet makes "identity trust" difficult to ensure in online settings. This has resulted in two critical problems pertaining to reputation systems [10, 25]: identity changes and reputation manipulation.

Identity Changes

Easy identity changes allow con artists and fraudulent buyers and sellers to thrive in electronic markets by constantly reappearing under different aliases. As Josang et al. noted, identity changes allow parties to "cut with the past and start from fresh" [25, p. 639]. Community members can build up a reputation, use it to deceive unsuspecting

members, and start over under a new identity [10, 16]. Friedman and Resnick [16] refer to this identity change characteristic as "cheap pseudonyms." Cheap pseudonyms stemming from easy identity changes allow online auction traders to circumvent the collaborative sanctioning mechanisms critical to the success of reputation systems.

Reputation Manipulation

Reputation scores in electronic markets are important because they influence product prices and traders' perceived credibility. There has been a plethora of work done to evaluate the correlation between reputation scores and product prices. Often enhanced seller reputation scores result in premium sales prices [36]. Resnick et al. [45] observed an 8.1 percent increase in the buyer's willingness-to-pay price when transacting with a highly reputable identity as compared with a nonestablished trader. Thorough reviews of literature evaluating the impact of reputation scores on selling price can be found in Dellarocas [10] and Resnick et al. [45]. Enhanced reputation also increases the willingness of other members to engage in transactions, which may be partially responsible for the enhanced selling prices. This is particularly important for fraudulent members attempting to "bait" unsuspecting members based on the fraudulent traders' false credibility.

Reputation manipulation can take two forms—rank inflation and discrimination. A common form of rank inflation involves using additional (fake) identities to inflate one's reputation [10]. This is also referred to as ballot box stuffing [25]. Corroborating with other members to create deceitful groups can further amplify the impact of such score inflation or stuffing [44]. Discrimination entails blackmailing or threatening to post negative feedback about fellow traders [44]. Posting dishonest comments to tarnish a competitor's reputation is a common ploy in online markets [10].

Reputation Systems and Stylometry

Rank manipulation and easy identity changes have facilitated numerous forms of fraud in electronic markets [9], including failure to ship, failure to pay, fencing, shell auctions, and so on. Consequently, many researchers have stated the need for techniques to mitigate the impact of identity change and rank manipulation [10, 25], both of which stem from online anonymity. Some have proposed using social network analysis for anti-aliasing; however, these techniques have had limited success on real-world data, with accuracies around 2 percent for matching e-mail aliases [23].

Systemic functional linguistic theory states that language has three kinds of meaning—ideational, textual, and interpersonal [20]. *Ideational* means that language consists of ideas. *Textual* indicates that language has organization, structure, and style. *Interpersonal* refers to the fact that language is a medium of exchange. The textual dimension of computer-mediated communication indicates that individuals convey their ideas using varying stylistic elements [14, 20]. Authorial style is influenced by education, gender, and vocabulary [28, 49] as well as subconscious factors described in the psycholinguistics literature [17]. Reputation rank systems entail users/traders

posting text comments. The traders often leave behind potential textual traces of their identity [37]. Keselj et al. [30] refer to an author's unique writing style tendencies as an "author profile." Ding et al. [12] describe such identifiers as "text fingerprints" that can discriminate authorship. Juola and Baayen [28] call them "stylistic fingerprints."

Stylometric/authorship identification techniques that can discriminate authorship in cyberspace could help alleviate the anonymity-related problems pervasive in electronic markets. Comparing trader feedback comments could help detect identity changes. Such methods may also help detect reputation score manipulation attributable to fake identities. Furthermore, comparing known fraudulent identities' comments against active members could help prevent further scamming. Many Web sites have begun to post archives and databases containing names, aliases, and text from fraudulent buyers and sellers [9]. For example, some documented fraudulent online auction individuals listed on www.traderlist.com have as many as 30–40 known fake identities. Clustering such "cheap pseudonyms" based on writing style tendencies could dramatically reduce the effectiveness of recurring deceptive behavior attributable to reappearing under different aliases.

Stylometric Analysis

Stylometry (also referred to as authorship analysis) is defined as the "statistical analysis of writing style." Four important characteristics of stylometric analysis [49] are the tasks, stylistic features, classification techniques, and parameters (i.e., factors influencing authorship analysis performance, such as number of classes, amount of text, noise).

Stylometric Analysis Tasks

Two major stylometric analysis tasks are identification and similarity detection [11, 19]. Identification entails comparing anonymous texts against those belonging to identified entities, where the anonymous text is known to be written by one of those entities. However, this "known class" assumption is not practical [28], especially for online settings. In cyberspace, author classes are rarely known in advance, and hence require the use of unsupervised clustering-based approaches. Such a similarity detection task requires the comparison of anonymous texts against other anonymous texts in order to assess the degree of similarity. For instance, in online forums, where there are numerous anonymous identities (i.e., screen names, handles, e-mail addresses), one can only use unsupervised stylometric analysis techniques because no class definitions are available. Similarly, in an online auction setting, 100 trader identities could represent anywhere between one and 100 actual traders.

Stylometric Analysis Features

Stylistic features are the attributes or writing style markers that are the most effective discriminators of authorship. The vast array of stylistic features includes lexical, syntactic, structural, content-specific, and idiosyncratic style markers.

Lexical features are word- or character-based statistical measures of lexical variation. These include style markers such as sentence/line length [3], vocabulary richness [11], and word length distributions [11, 49]. *Syntactic* features include function words [1, 37], punctuation, and part-of-speech tag n-grams [4, 34]. *Structural* features, which are especially useful for online text, include attributes relating to text organization and layout [11, 49]. *Content-specific* features are important key words and phrases pertaining to certain topics. For example, content-specific features on a discussion of computers may include "laptop" and "notebook." *Idiosyncratic* features include misspellings, grammatical mistakes, and other usage anomalies. Such features are extracted using spelling and grammar checking tools [8, 34].

Over 1,000 different features have been used in previous authorship analysis research with no consensus on a best set of style markers [46]. However, this could be attributable to certain feature categories being more effective at capturing style variations in different contexts. This necessitates the use of larger feature sets comprised of several categories of features spanning various feature groups (i.e., lexical, syntactic, etc.). For instance, the use of feature sets containing lexical, syntactic, structural, and syntactic features has been shown to be more effective for online identification than feature sets containing only a subset of these feature groups [1, 49].

Stylometric Analysis Techniques

Several techniques have been used for stylometric identification. These can broadly be classified as supervised and unsupervised methods. However, only unsupervised techniques are suitable for online settings, such as reputation system feedback comments, because class definitions are unknown a priori [39]. We discuss previous unsupervised methods useful for online similarity detection. These techniques include principal component analysis (PCA), n-gram models, Markov models, cross entropy, and K–L similarity. Previous stylometric analysis studies using these techniques are summarized in Table 1.

Principal Component Analysis. PCA is a popular stylometric identification technique that has been used in numerous previous studies [2, 4, 7, 33]. PCA's ability to capture essential variance across large amounts of features in a reduced dimensionality makes it attractive for text analysis problems, which typically involve large feature sets. The essence of PCA can be described as follows: given a feature matrix with each column representing a feature and instance vector rows for the various authors' texts, project the matrix into a lower dimensional space by plotting principal component scores (which are the product of the component weights and instance feature vectors). The similarity between authors can be compared based on visual proximity of patterns [33] or computation of average distance [2]. Given a set of n text instance vectors and p eigenvectors, the average distance can be used to compute authorship dissimilarity as follows:

$$\text{Dissimilarity}(a,b) = \frac{\sum_{i=1}^{n}\sum_{k=1}^{p}\left|a_{ki} - b_{ki}\right|}{np},$$

Table 1. Previous Unsupervised Stylometric Analysis Techniques

Technique	Study	Features	Test bed
PCA	Kjell et al. [33]	Letter bigrams (10 total)	Federalist Papers (2 authors)
	Baayen et al. [4]	Function words (50 total)	Literary texts (2 authors)
	Abbasi and Chen [2]	Punctuation, word length distributions, topical words, special characters, letters, and so on (104 total)	Pirated software Web forum (10 authors)
N-gram models	Keselj et al. [30]	Character n-grams (5,000 per author)	Literary texts (8 authors)
	Peng et al. [41]	Character n-grams (5,000 per author)	Literary texts (8 authors)
Markov models	Khmelev [31]	Character bigrams (729 total)	Literary texts (82 authors)
	Khmelev and Tweedie [32]	Character bigrams (729 total)	Project Gutenberg (45 authors), literary texts (2 authors), Federalist Papers (2 authors)
Cross entropy	Juola [26]	Match lengths	Federalist Papers (2 authors)
	Juola and Baayen [28]	Match lengths	Student essays (8 authors)
K–L similarity	Novak et al. [39]	Word unigrams (quantity not available)	Message board postings from www.courttv.com (100 authors)

where a_{ki} and b_{ki} are the coefficients of the kth component of the usage instance i for authors a and b.

N-Gram Models. Proposed by Keselj et al. [30] and Peng et al. [41], this technique requires the construction of a profile for each author, where a profile is the set of the n most frequently used character n-grams. Keselj et al. [30] used between 20 and 5,000 as the value for n, with the best accuracy attained using 5,000 n-grams. They attained the best results using four- to eight-character n-grams. Using this approach, they computed the dissimilarity between two authors as

$$\text{Dissimilarity}\left(\text{profile}_1, \text{profile}_2\right) = \sum_{x \in \text{profile}_1 \cup \text{profile}_2} \left(\frac{2\left(f_1(x) - f_2(x)\right)}{f_1(x) + f_2(x)} \right)^2,$$

where $f_1(x)$ and $f_2(x)$ are frequencies of an n-gram x contained in profile 1 or profile 2. Keselj et al. [30] and Peng et al. [41] were able to attain good performance using this approach on test beds consisting of up to eight authors.

Markov Models. Proposed by Khmelev [31] and later extended by Khmelev and Tweedie [32], this technique requires the creation of a Markov model for each author, using bigrams of letters and the space character. Khmelev [31] removed all other characters and ignored words beginning with capital letters, resulting in a fixed ($27 \times 27 = 729$) feature space for each author. Using this approach, the similarity between two authors can be computed as follows:

$$\text{Similarity}\left(\text{a}, \text{b}\right) = \sum_i \sum_j \left| \ln\left(\frac{f_{ij}(a) f_i(b)}{f_i(a) f_{ij}(b)} \right) \right|,$$

where $f_{ij}(a)$ and $f_{ij}(b)$ are the number of transitions from letter i to j for authors a and b's texts, respectively. The technique has performed well on larger test beds of 45 and 82 authors [31, 32]. However, these data sets consisted of literary texts that tend to be longer and more stylistically consistent due to contextual independence.

Cross Entropy. Proposed by Juola [26, 27] and later applied in Juola and Baayen [28], this technique is based on the concept of match length where

> The match length $L_n(x)$ of a sequence $x_1, x_2, ..., x_k$ is the length of the longest prefix of the sequence $x_{n+1}, x_{n+2}, ..., x_k$ that matches a contiguous substring of $x_1, x_2, ..., x_n$.

The substring $x_1, x_2, ..., x_n$ is referred to as the database. For cross entropy, simply compute the average match length for author b's text $b_1, b_2, ..., b_k$ compared with author a's database $a_1, a_2, ..., a_n$ and author a's $a_1, a_2, ..., a_j$ to author b's database $b_1, b_2, ..., b_n$ as follows:

$$\text{Similarity}(\text{a},\text{b}) = \frac{\sum_{i=1}^{j} L_n(a_i, b)}{j} + \frac{\sum_{i=1}^{k} L_n(b_i, a)}{k},$$

where $L_n(a_i, b)$ is the match length for author a's substring a_i, a_{i+1}, ..., a_j compared with author b's database.

Texts written by the same author should result in higher match lengths. Juola [26] used n = 2,000 characters for each author's database size. The cross entropy method has performed well in prior studies, outperforming PCA on a test bed consisting of eight students' essays [28].

K–L Similarity. Novak et al. [39] used the Kullback–Leibler divergence as follows:

$$\text{Similarity}(\text{a},\text{b}) = \sum_{i=1}^{n} p_i \frac{\log p_i}{\log q_i},$$

where p and q are the feature distributions for the two authors a and b.

Novak et al. [39] performed smoothing to account for nonzero elements in p and applied the approach to message board postings on www.courttv.com. They compared various features, and attained the best performance using word unigrams. Their study is one of the few prior similarity detection studies applied to computer-mediated communication. Kullback–Leibler similarity using word unigrams performed well; however, they acknowledged that their approach was susceptible to topical variation [39], possibly stemming from the use of a feature set comprised only of word unigram features. While topical variation is less of a concern for online feedback comments, the sensitivity of such an approach may make it susceptible to intentional obfuscation.

Most techniques, such as n-gram and Markov models, were designed to be used with character n-grams. Word-based features are too sparse to be used accurately with these techniques [41]. Similarly, Novak et al. [39] attained better performance using the Kullback–Leibler similarity on word unigrams as compared to other features, such as punctuation, function words, misspellings, and a combined feature set. It is unclear if such methods can be effectively applied to online settings, where techniques capable of handling larger feature sets are typically required [1, 49]. Therefore, assessing the efficacy of these approaches (i.e., the combination of features and techniques employed by these prior studies) for online analysis is especially important in order to gauge their applicability for stylometric similarity detection of reputation system feedback comments.

Stylometric Analysis Parameters

Two important stylometric analysis parameters for online authentication are scalability and robustness [49]. Scalability refers to the impact of the number of author classes on classification performance. Typically, there has been a noticeable drop in

performance for prior online message level identification research as the number of authors increased. Zheng at al. [49] noted a 14 percent drop in accuracy when increasing the number of authors from five to 20. Argamon et al. [3] observed as much as a 23 percent drop in accuracy over a similar number of authors. Given the large number of traders in online markets, it is important to assess the impact of the number of traders and identities per trader on stylometric performance.

It is also important to assess robustness of stylometric approaches against intentional stylistic alteration and copycatting/message forging. Fraudulent traders may attempt to avoid detection by altering their style or copying other traders' style (referred to as copycatting or forging). Previous research on intentional authorship obfuscation suggests that such alteration can impact stylometric classification performance [42]. For instance, word substitution, a popular and convenient form of alteration, has been shown to impact identification accuracy [29]. Rao and Rohatgi [42] noted that word substitution via the use of thesaurus tools (altering words with synonyms) could represent a promising stylistic obfuscation mechanism because it would decrease the presence of stylistic elements attributable to an author's vocabulary. Forging/copycatting entails intentionally mimicking other community members' styles or user names [2]. This behavior is fairly common in certain computer-mediated communication (CMC) modes, such as Usenet forums. Mimicking other members' styles by either directly copying their text or attempting to copy their stylistic tendencies is an important and plausible form of deception that must be considered when evaluating stylometric methods in online settings.

Research Gaps, Questions, and Design

BASED ON THE REVIEW OF RELATED RESEARCH on reputation systems and stylometric analysis, we present several research gaps and questions.

Research Gaps

Stylometric Similarity Detection of Feedback Comments

We are not aware of any prior application of stylometric similarity detection techniques to online feedback comments. Most previous stylometric work either focused on the online identification task (known classes) or was applied to literary texts. Successful application to reputation system feedback comments could reduce fraud and deception in reputation systems, and consequently, online markets.

Techniques That Can Handle Richer Feature Sets

There is a need for techniques that can handle richer feature sets for online settings. Existing techniques were designed to use a single category of features (e.g., character n-grams, match length). However, application to online settings necessitates the use of techniques that can incorporate rich feature sets [1, 49].

Analysis of Scalability

There has been limited work done to analyze the scalability of stylometric techniques for application to reputation system feedback comments. Given the large number of identities in online markets, there is a need to apply stylometry effectively in a scalable manner.

Analysis of Robustness Against Intentional Alteration and Forging

We are unaware of any previous research to assess the robustness of stylometric features and techniques against intentional stylistic alteration or forging. Unlike biometrics, writing style may be susceptible to intentional manipulation via stylistic alteration or message forging. It is important to assess the robustness of stylometric similarity detection techniques against such obfuscation of authorship.

Research Questions

Based on the gaps described, we propose the following research questions:

RQ1: Which stylometric technique is most effective for similarity detection of online market feedback comments?

RQ2: How scalable are these techniques in terms of number of traders and identities per trader?

RQ3: How robust are these techniques against intentional stylistic obfuscation?

RQ4: Can techniques using richer feature sets provide improved scalability and robustness?

Research Design

We propose the development of a stylometric similarity detection system capable of differentiating between online traders based on stylistic tendencies inherent in feedback comments left for other buyers and sellers. Our system uses an extended feature set comprised of lexical, syntactic, structural, content-specific, and idiosyncratic style markers. The system also includes a novel Writeprint technique, which compares the style patterns between two identities. Writeprint uses Karhunen–Loeve transforms to assess the similarity for features used by the two identities as well as a pattern disruption mechanism that assesses the degree of dissimilarity for features used by one identity but not the other.

We intend to compare our system, which includes the Writeprint technique and an extended feature set, against existing similarity detection approaches described above, including PCA, n-gram models, Markov models, cross entropy, and Kullback–Leibler

similarity. The evaluation will assess the scalability and robustness of our system and comparison approaches for application to online market feedback comments.

System Design

THE PROPOSED SYSTEM HAS TWO MAJOR COMPONENTS—feature extraction and classifier construction (as shown in Figure 1). The feature extraction phase derives various static and dynamic features (e.g., n-grams) from the trader feedback comments. A subset of the dynamic features is chosen using feature selection in order to create an extended feature set that is passed forward to the classifier construction phase. This stage involves the creation of Writeprints for each trader identity, which can then be compared against each other to assess the degree of stylistic similarity.

Feature Extraction

The extraction phase involves derivation of static and dynamic features resulting in the creation of our extended feature set. For static features, extraction simply involves generating the feature usage statistics (feature vectors) across texts; however, dynamic feature categories such as n-grams require indexing and feature selection. The feature extraction procedure for the extended feature set is described below; Table 2 provides a description of the style markers included. For dynamic feature categories, the number of attributes varies depending on indexing and feature selection. For some such categories, the upper limit of features is already known (e.g., number of character bigrams is less than 676).

Dynamic features incorporated in the extended feature set include several n-gram feature groups and a list of 5,513 common word misspellings taken from various Web sites, including Wikipedia (www.wikipedia.org). N-gram categories utilized include character, word, part-of-speech tag, and digit-level n-grams. These categories require indexing with the number of initially indexed features varying depending on the data set. The indexed features are then sent forward to the feature selection phase. Use of such an indexing and feature selection/filtering procedure for n-grams is necessary and common in stylometric analysis research [34, 41].

Feature selection is applied to all the n-gram and misspelled word categories using the information gain (IG) heuristic. Information gain has been used in many text categorization studies as an efficient method for selecting text features [34]. Specifically, it is computationally efficient compared to search-based techniques and good for multiclass text problems [48]. The information gain for feature j across a set of classes c is derived as $IG(c, j) = H(c) - H(c|j)$, where $H(c)$ is the overall entropy across author classes and $H(c|j)$ is the conditional entropy for feature j. For each identity, information gain is applied using a two-class (one-against-all) setup (size of $c = 2$, c_1 = identity, c_2 = rest). Thus, each trader identity's feature set is intended to be comprised of the set of dynamic features that can best discriminate that specific identity against all others.

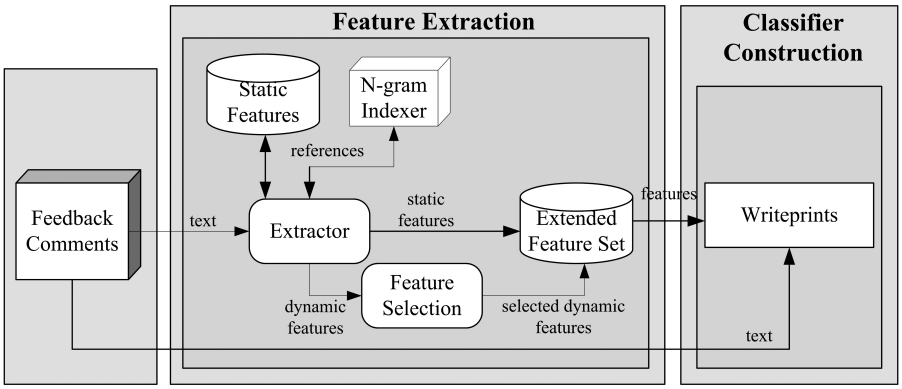

Figure 1. Stylometric Similarity Detection System Design

Classifier Construction: Writeprints

We propose a novel Writeprint technique that has two major components—creation and comparison. The creation steps are concerned with the construction of Writeprint patterns reflective of an identity's writing style variation, based on the occurrence of common identity features as well as lack of occurrence of style markers prevalent in other identities' text. The comparison steps describe how created Writeprints for various trader identities are compared against one another to assess the degree of stylistic similarity.

Writeprint Creation

The Writeprint creation component can be further decomposed into two steps. In the first step, Karhunen-Loeve transforms are applied with a sliding window in order to capture stylistic variation with a finer level of granularity. Writeprints are created for each identity using their key features. Although some unsupervised variants of Karhunen–Loeve transforms are similar to PCA, we used a version that allows the inclusion of class information, in this case for the different identities/aliases [47]. The second step, pattern disruption, uses zero usage features as red flags intended to decrease the level of stylistic similarity between identities when one identity contains important features not occurring in the other. The two major steps, which are repeated for each identity, are shown below.

Writeprint Creation Steps

1. For all identity features with occurrence frequency > 0.
   a. Extract feature vectors for each sliding window instance.
   b. Derive basis matrix (set of eigenvectors) from feature usage covariance matrix using Karhunen–Loeve transforms.

Table 2. Extended Feature Set

Group	Category	Quantity	Description/examples
Lexical	Word level	5	Total words, percent characters per word
	Character level	5	Total characters, percent characters per message
	Character n-grams	< 18,278	Count of letter n-grams (e.g., a, at, ath)
	Digit n-grams	< 1,110	Count of digit n-grams (e.g., 1, 12, 123)
	Word-length distribution	20	Frequency distribution of 1–20 letter words
	Vocabulary richness	8	Richness (e.g., hapax legomena, Yule's K)
	Special characters	21	Occurrences of special characters (e.g., @#$%^&*+=)
Syntactic	Function words	300	Frequency of function words (e.g., of, for, to)
	Punctuation	8	Occurrence of punctuation marks (e.g., !;:,.?)
	Part-of-speech tag n-grams	Varies	Part-of-speech tag n-grams (e.g., NNP, NNP JJ)
Structural	Message level	6	For example, has greeting, has URL, requoted content
	Paragraph level	8	For example, number of paragraphs, sentences per paragraph
	Technical structure	50	For example, file extensions, fonts, use of images
Content specific	Word n-grams	Varies	Bag-of-word n-grams (e.g., "seller", "bad sale")
Idiosyncratic	Misspelled words	< 5,513	Common misspellings (e.g., "beleive", "thougth")

   c. Compute window instance coordinates (principal components) by multiplying window feature vectors with basis. Window instance points in n-dimensional space represent author Writeprint pattern.
2. For all author features with occurrence frequency = 0.
   a. Compute feature disruption value as product of information gain, synonymy usage, and disruption constant K.
   b. Append features' disruption values to basis matrix.
3. Repeat steps 1–2 for each identity.

Step 1: Sliding Window and Karhunen–Loeve Transforms. A lower dimensional usage variation pattern is created based on the occurrence frequency of the identity's features (individual-level feature set). For all features with usage frequency greater than zero, a sliding window of length L with a jump interval of J characters is run over the identity's messages. The feature occurrence vector for each window is projected to an n-dimensional space by applying the Karhunen–Loeve transform. The Kaiser–Guttman stopping rule [24] was used to select the number of eigenvectors in the basis. The formulation for step 1 is presented below:

1. Let $\Omega = \{1, 2, ..., f\}$ denote the set of f features with frequency greater than 0 and $\Phi = \{1, 2, ..., w\}$ represent the set of w text windows. Let X denote the author's feature matrix where x_{ij} is the value of feature $j \in \Omega$ for window $i \in \Phi$.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1f} \\ x_{21} & x_{22} & \cdots & x_{2f} \\ \cdots & \cdots & \cdots & \cdots \\ x_{w1} & x_{w2} & \cdots & x_{wf} \end{bmatrix}.$$

2. Extract the set of eigenvalues $\{\lambda_1, \lambda_2, ..., \lambda_n\}$ for the covariance matrix Σ of the feature matrix X by finding the points where the characteristic polynomial of Σ equals 0:

$$p(\lambda) = \det(\Sigma - \lambda I) = 0.$$

   For each eigenvalue $\lambda_m > 1$, extract its eigenvector $a_m = (a_{m1}, a_{m2}, ..., a_{mf})$ by solving the following system, resulting in a set of n eigenvectors $\{a_1, a_2, ..., a_n\}$:

$$(\Sigma - \lambda_m I)a_m = 0.$$

3. Compute an n-dimensional representation for each window i by extracting principal component scores ε_{ik} for each dimension $k \le n$:

$$\varepsilon_{ik} = a_k^T x_i.$$

Step 2: Pattern Disruption. Because Writeprint uses individual author-level feature sets, an author's key set of features may contain attributes that are significant because

the author never uses them. However, features with no usage by the identity of interest will currently be irrelevant to the process because they have no variance. Nevertheless, these features are still important when comparing a trader identity to other anonymous trader identities. The trader's lack of usage of these features represents an important stylistic tendency. Anonymous identity texts containing these features should be considered less similar (because they contain attributes never used by this author). When comparing two trader identities A and B, we would like A's zero frequency features to act as pattern disruptors, where the presence of these features in identity B's feedback comments decreases the similarity for the particular A–B comparison (and vice versa for the B–A comparison).

The magnitude of a disruptor signifies the extent of the disruption for a particular feature. Larger values of for the disruptor will cause pattern points representing text windows containing the disruptor feature to be shifted further away. However, not all features are equally important discriminators. Koppel et al. [35] developed a machine translation-based technique for measuring the degree of feature "stability." Stability refers to how often a feature changes across authors and documents for a constant topic. They found noun phrases to be more stable than function words and argued that function words are better stylistic discriminators than noun phrases because use of function words involves making choices between a set of synonyms. Based on this intuition, we used the disruptor feature's information gain and synonymy information to assign them a weight (disruptor coefficient), which was appended to the identity's basis matrix (set of eigenvectors).

1. Let $\Psi = \{f + 1, f + 2, ..., f + g\}$ denote the set of g features with zero frequency. For each feature $p \in \Psi$, compute the disruptor coefficient d_p:

$$d_p = IG(c, p)K(\text{syn}_{total} + 1)(\text{syn}_{used} + 1),$$

   where $IG(c, p)$ is the information gain for feature p across the set of classes c; syn_{total} and syn_{used} are the total synonyms and the number used by the author, respectively, for the disruptor feature; and K is a disruptor constant.
2. For each feature $p \in \Psi$, append the value d_{kp} to each eigenvector a_k, where $k \leq n$.

Writeprint Comparisons

When comparing two identities' usage variation patterns, two comparisons must be made because both identities used different feature sets and basis matrices in order to construct their lower-dimensional patterns. The dual comparisons are illustrated in Figure 2. We would need to construct a pattern for identity B using B's text with A's feature set and basis matrix (pattern B) to be compared against identity A's Writeprint (and vice versa). The overall similarity between identity A and B is the sum of the average distance between Writeprint A and pattern B and Writeprint B and pattern A.

As previously mentioned, the pattern disruptors are intended to assess the degree of stylistic dissimilarity based on important features only found in one of the two

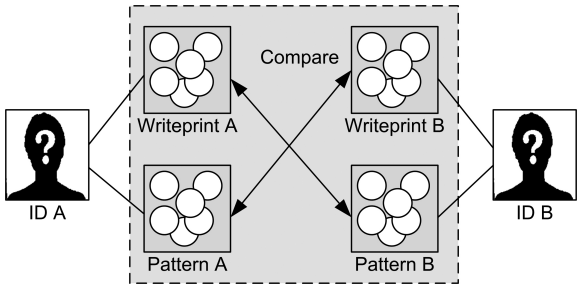

Figure 2. Writeprint Comparisons

identities' feedback comments. Disruptors shift pattern points further away from the Writeprint they are being compared against, thereby increasing the average distance between patterns (and reducing the similarity score). The direction of a pattern window point's shift is intended to reduce the similarity between the Writeprint and comparison pattern. This is done by making d_{kp} positive or negative for a particular dimension k based on the orientation of the Writeprint (WP) and comparison pattern (PT) points along that dimension, as follows:

$$d_{kp} = \begin{cases} -d_{kp}, & \text{if } \sum_{i=1}^{w} \frac{WP_{ik}}{w} > \sum_{i=1}^{w} \frac{PT_{ik}}{w} \\ d_{kp}, & \text{if } \sum_{i=1}^{w} \frac{WP_{ik}}{w} < \sum_{i=1}^{w} \frac{PT_{ik}}{w}. \end{cases}$$

For instance, if identity A's Writeprint is spatially located to the left of identity B's pattern for dimension k, the disruptor d_{kp} will be positive in order to ensure that the disruption moves the comparison pattern away from the Writeprint (toward the right in this case) as opposed to toward it.

Evaluation

In order to evaluate the effectiveness of the proposed system, which includes the Writeprint technique and extended feature set, experiments were conducted that compared the system against previous unsupervised stylometric identification techniques described, including PCA, n-gram and Markov models, cross entropy, and Kullback–Leibler.

Test Bed

The test bed consisted of buyer–seller feedback comments extracted from eBay's online reputation system. We randomly extracted 200 eBay members selling electronic goods. For each trader, 3,000 feedback comments posted by that author were

Table 3. eBay Test Bed Statistics

Number of authors (i.e., traders)	Words (per author)	Comments (per author)	Average comment length (words)	Time duration
200	22,564	3,000	7.94	02/2003–06/2006

included. Table 3 provides summary statistics of the test bed and example feedback comments are listed below:

- "Another quick & easy transaction, thanks for your biz!"
- "Excellent e-bayer!! fast payment, great to deal with, many thanks!!!"
- "PLEASURE doing business with you and thanks for making this business a PLEASURE!"

Experimental Setup

All comparison techniques were run using the best parameter settings determined by tuning these parameters on the actual test bed data. This was done in order to allow the best possible comparison against the proposed Writeprint technique. Most of the parameter values were consistent with prior research. PCA was run using the extended feature set. We extracted feature vectors for 1,500 character text blocks, consistent with prior research [2]. The Kaiser–Guttman stopping rule was used (i.e., extract all eigenvectors with an eigenvalue greater than 1). For the n-gram models, we used character-level n-grams, with profile sizes of 5,000 n-grams per identity. For each identity, we used four- to eight-character n-grams because this configuration garnered the best results, also consistent with Keselj et al. [30] and Peng et al. [41]. Markov models were built using letters and space bigrams. We removed all other characters and ignored words beginning with capital letters, as done by Khemelev [31] and Khemelev and Tweedie [32]. For cross entropy, we used a database size of 5,000 characters for each identity as this size provided the best performance. For the Kullback–Leibler similarity, word unigrams were used and smoothing was performed as outlined by Novak et al. [39].

For the experiments, we created multiple identities for each of the 200 eBay traders by splitting the traders' feedback comment text into multiple parts, as done in prior research [39]. The objective of the experiments was to see how well the proposed Writeprint method and comparison techniques could match up the different trader identities based on their comment texts. Each trader's text was split into 12 parts. If two identities were to be created for a single trader, six parts were randomly assigned to each identity; for example, parts 1, 5, 7, 8, 9, 11 (identity 1), parts 2, 3, 4, 6, 10, 12

(identity 2). In order to test the statistical significance of the techniques' performance, bootstrapping was performed 30 times for each technique, where each iteration the 12 trader text parts were randomly split into the desired number of identities. A trial-and-error method was used to find the optimal similarity threshold for matching for each technique. The same threshold was used throughout the experiments for the Writeprint method. A dynamic threshold yielding optimal results for the particular experimental settings was used for each comparison technique. This was done in order to compensate for differences in performance attributable to thresholds instead of techniques. All identity-identity scores above a technique's threshold were considered a match. The *F*-measure was used to evaluate performance.

$$F\text{-measure} = \frac{2(\text{Precision})(\text{Recall})}{\text{Precision} + \text{Recall}},$$

where Precision = (number of identities assigned correctly)/(total number of identities assigned); Recall = (number of identities assigned correctly)/(total number of identities).

Using these experimental settings, two sets of experiments were conducted. The first assessed the scalability of the proposed stylometric similarity detection system and comparison approaches in terms of number of traders and number of identities' comments. The second attempted to evaluate the effectiveness of these stylometric methods against intentional stylistic alteration and forging/copycatting.

Experiment 1: Scalability

We conducted two experiments to analyze the scalability across traders (experiment 1a) and identities (experiment 1b). In experiment 1a, scalability across traders was evaluated. Each trader's text was split into two anonymous identities. We used 25, 50, 100, and 200 traders (i.e., 50, 100, 200, and 400 identities). In experiment 1b, scalability across identities was the focal point. We used 50 traders, with each trader's text split into *n* anonymous identities. We used 2, 3, 4, and 5 identities per trader (i.e., 100, 150, 200, and 250 identities total). The details of the number of traders and identities used for experiment 1 are presented in Table 4.

Results for Experiment 1a: Scalability Across Traders

Figure 3 shows the *F*-measure percentages for 25, 50, 100, and 200 traders (with two identities per trader), intended to assess the scalability across traders. Overall, all the techniques except PCA performed well. As expected, doubling the number of authors and identities decreased performance, however, the decrease was gradual. Writeprint had the best performance for all four identity levels. The technique only had approximately a 3 percent decrease when going from 100 to 200 identities and from 200 to 400 identities. In contrast, the performance of n-gram models, K–L similarity, and cross entropy fell 6 percent to 7 percent for each such increase.

Table 4. Number of Traders and Identities Used in Experiment 1

Experiment	Number of traders	Number of identities	Words (per identity)	Comments (per identity)
1a (traders)	25	50	11,282	1,500
	50	100	11,282	1,500
	100	200	11,282	1,500
	200	400	11,282	1,500
1b (identities)	50	100	11,282	1,500
	50	150	7,521	1,000
	50	200	5,641	750
	50	250	4,513	600

Techniques	#Identities			
	50	100	200	400
Writeprint	**100.00**	**97.88**	**94.59**	**92.16**
PCA	81.19	77.49	72.25	65.80
N-Gram Models	97.96	97.60	89.39	78.42
Markov Models	98.04	90.91	87.86	76.18
Cross Entropy	**100.00**	96.15	90.34	83.53
K-L Similarity	98.06	93.52	88.90	77.18

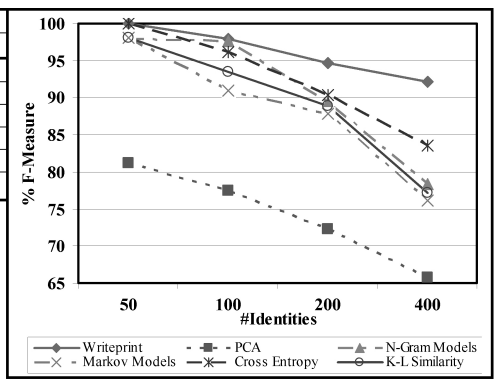

Figure 3. Experiment 1a Results (Scalability Across Traders Using Two Identities per Trader)

Table 5 shows the *p*-values for the pairwise *t*-tests on *F*-measure. For all *t*-tests, a Bonferroni correction was performed to avoid spurious positives stemming from the large number of comparisons. Only *p*-values less than 0.0001 were considered significant. Because this threshold is considerably lower than α/n, we are confident that it ensures the statistical validity of the *t*-tests. Because our primary concern is the effectiveness of Writeprint coupled with the extended feature set, only *p*-values for this technique are depicted in Table 5. However, other significant results of interest are also reported in the text description below.

Writeprint significantly outperformed all comparison techniques. The n-gram and Markov models, cross entropy, and K–L similarity techniques significantly outperformed PCA for all four settings (*p*-values < 0.0001). Furthermore, cross entropy significantly outperformed n-gram and Markov models and K–L similarity when using 400 identities (*p*-values < 0.0001).

Results for Experiment 1b: Scalability Across Identities

Figure 4 shows the *F*-measure percentages for 2, 3, 4, and 5 identities per trader (with 50 traders), intended to assess the scalability across identities. Writeprint again had

Table 5. *p*-Values for Pairwise *t*-Tests on *F*-Measure ($n = 30$)

	Number of traders/number of identities			
Techniques	25/50	50/100	100/200	200/400
Writeprint versus PCA	< 0.0001*	< 0.0001*	< 0.0001*	< 0.0001*
Writeprint versus n-gram models	< 0.0001*	0.1090	< 0.0001*	< 0.0001*
Writeprint versus Markov models	< 0.0001*	< 0.0001*	< 0.0001*	< 0.0001*
Writeprint versus cross entropy	0.8521	< 0.0001*	< 0.0001*	< 0.0001*
Writeprint versus K–L similarity	< 0.0001*	< 0.0001*	< 0.0001*	< 0.0001*

* *p*-values significant at corrected threshold alpha/$n = 0.0001$.

the best performance for all four trader/identity levels. N-gram and Markov models performed worse on this experiment as compared to the trader scalability experiment (1a), with 10 percent to 15 percent lower performance on an equal number of total identities (see values for 200 identities in experiment 1a and four identities per trader in experiment 1b). The results suggest that the number of identities per author has a greater impact on performance than the number of authors for these techniques. Perhaps this is due to the amount of text per identity, which was constant in experiment 1a and decreased in experiment 1b as the number of identities per trader increased. Writeprint, cross entropy, and the K–L similarity method appear more robust against smaller amounts of text. This finding is consistent with Novak et al. [39], who also found the K–L similarity approach to work almost equally well when dealing with two to four aliases.

Table 6 shows the *p*-values for the pairwise *t*-tests on *F*-measure. Writeprint significantly outperformed all comparison techniques. This is likely attributable to the pattern disruptors effectively differentiating between a larger number of identities per author. Cross entropy significantly outperformed n-gram and Markov models, K–L similarity, and PCA for all four settings. This is consistent with prior research, where the technique has been shown to be effective when applied to smaller texts [28].

Results Discussion for Experiment 1

In both experiments, Writeprint had the best performance for all trader/identity levels. The performance gap widened as the number of traders and identities increased, suggesting that the extended feature set and pattern disruption mechanism incorporated by Writeprint allowed improved scalability. The enhanced representational richness of Writeprint allowed it to outperform the word (K–L similarity) and n-gram-based techniques (n-gram and Markov models) while the pattern disruption component allowed improved performance over PCA.

Techniques	#Identities per Trader			
	2	3	4	5
Writeprint	**97.88**	**96.92**	**95.59**	**94.72**
PCA	77.49	71.09	70.43	67.21
N-Gram Models	97.60	84.44	78.66	74.22
Markov Models	90.91	73.23	72.43	69.32
Cross Entropy	96.15	93.27	90.16	88.10
K-L Similarity	98.06	91.21	87.90	85.31

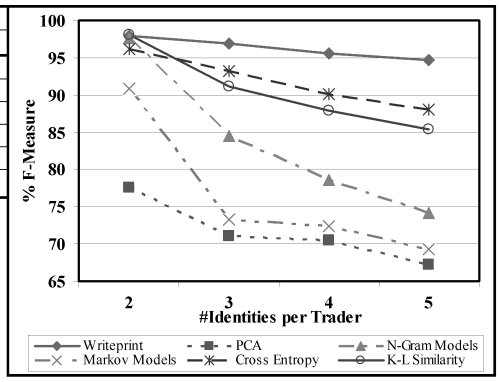

Figure 4. Experiment 1b Results (Scalability Across Identities Using 50 Traders)

Table 6. *p*-Values for Pairwise *t*-Tests on *F*-Measure ($n = 30$)

	Number of traders/number of identities			
Techniques	50/100	50/150	50/200	50/250
Writeprint versus PCA	< 0.0001*	< 0.0001*	< 0.0001*	< 0.0001*
Writeprint versus n-gram models	0.1090	< 0.0001*	< 0.0001*	< 0.0001*
Writeprint versus Markov models	< 0.0001*	< 0.0001*	< 0.0001*	< 0.0001*
Writeprint versus cross entropy	< 0.0001*	< 0.0001*	< 0.0001*	< 0.0001*
Writeprint versus K–L similarity	< 0.0001*	< 0.0001*	< 0.0001*	< 0.0001*

* *p*-values significant at corrected threshold alpha/$n = 0.0001$.

Experiment 2: Robustness

We conducted experiments to analyze the robustness of the proposed system and comparison approaches against intentional stylistic alteration and copycatting/forging. For each experiment, we used 50 traders, with every trader's text split into two identities. For each trader, one identity was kept unchanged while the other was altered using word substitution or forging. In experiment 2a, intentional stylistic alteration was simulated using word substitution and experiment 2b evaluated the impact of message forging.

Results for Experiment 2a: Robustness Against Word Substitution

Word substitution is a popular obfuscation strategy because word-based features are transparent and more easily modifiable [29]. Altering words with semantically equivalent ones using thesauruses is considered a promising technique for stylistic

Table 7. Impact of Different Levels of Word Substitution on an Example Comment

Percent words altered	Number of alterations per comment	Example comment
0	0.000	"Excellent e-bayer!! fast payment, great to deal with, many thanks!!!"
20	1.448	"Superb e-bayer!! swift payment, great to deal with, many thanks!!!"
40	2.883	"Astounding e-bayer!! expedited payment, lovely to deal with, many thanks!!!"
60	4.349	"Awesome e-bayer!! quick payment, wonderful to interact with, lots of thanks!!!"

obfuscation [42]. Based on this rationale, we simulated word synonym substitution using a thesaurus. For each altered identity, WordNet [15] was used to randomly alter n percent of the words with a synonym randomly taken from the synset. We used 20 percent, 40 percent, and 60 percent as values for n. Table 7 shows the average number of alterations per comment for each setting of n and the impact of such alteration on an actual comment.

Figure 5 shows the F-measure percentages for 20 percent, 40 percent, and 60 percent word substitution using 50 traders and two identities per trader. Writeprint had the best performance against alteration with cross entropy also performing very well. These techniques seem more robust against synonymy-based word alteration. N-gram and Markov models and K–L similarity all performed poorly. These techniques' accuracy dropped 50–75 percent with 20 percent synonym alteration. These methods utilize character n-grams and word unigrams, respectively, which may be more susceptible to alteration. In comparison, PCA's performance was more stable. While n-gram and Markov models outperformed PCA by a wide margin when no substitution was performed, PCA considerably outperformed these techniques once different levels of alteration were introduced.

Table 8 shows the p-values for the pairwise t-tests on F-measure for the experiment evaluating robustness against word substitution. Writeprint significantly outperformed all comparison techniques. PCA also outperformed n-gram and Markov models and K–L similarity with p-values less than 0.0001, as previously mentioned. However, cross entropy significantly outperformed PCA for all three alteration levels (all p-values < 0.0001). The t-tests indicate that, once again, Writeprint had the best performance followed by cross entropy.

Results for Experiment 2b: Robustness Against Forging

Message forging (also referred to as copycatting) occurs when an individual attempts to mimic another user by imitating the user's writing style [2]. In order to assess the impact of forging on stylometric similarity detection of online market feedback

Techniques	% Words Substituted			
	0	20	40	60
Writeprint	**97.88**	**97.05**	**96.00**	**85.71**
PCA	77.49	72.44	61.03	50.29
N-Gram Models	97.60	45.78	27.45	20.00
Markov Models	90.91	15.38	9.43	6.00
Cross Entropy	96.15	92.22	88.18	83.33
K-L Similarity	93.52	62.45	51.43	42.09

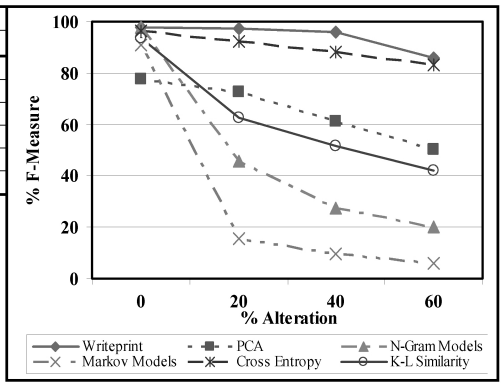

Figure 5. Experiment 2a Results (Robustness Against Word Substitution)

Table 8. *p*-Values for Pairwise *t*-Tests on *F*-Measure ($n = 30$)

	Percent alterations			
Techniques	0	20	40	60
Writeprint versus PCA	< 0.0001*	< 0.0001*	< 0.0001*	< 0.0001*
Writeprint versus n-gram models	0.1090	< 0.0001*	< 0.0001*	< 0.0001*
Writeprint versus Markov models	< 0.0001*	< 0.0001*	< 0.0001*	< 0.0001*
Writeprint versus cross entropy	< 0.0001*	< 0.0001*	< 0.0001*	0.0043
Writeprint versus K–L similarity	< 0.0001*	< 0.0001*	< 0.0001*	< 0.0001*

* *p*-values significant at corrected threshold alpha/$n = 0.0001$.

comments, we simulated identities engaging in different levels of message forging. In similar fashion to the previous experiment, 50 traders and two identities per trader were incorporated, with one of the two trader identities being subjected to different levels of forging. For each altered identity, we randomly substituted *n* percent of the identity's messages with randomly selected messages taken from other author identities. We used 10 percent, 20 percent, and 30 percent values for *n*. Table 9 illustrates the impact of 20 percent forgery on a set of five comments from an author. In this case, one comment out of five (20 percent) is forged with a random comment taken from another identity.

Figure 6 shows the *F*-measure percentages for 10 percent, 20 percent, and 30 percent message forging using 50 traders and two identities per trader. Cross entropy performed the best against forging. Writeprint's performance fell at an increasing rate, especially at 20 percent and 30 percent forging. N-gram and Markov model performance plummeted once again when exposed to message forging. PCA was the only technique

Table 9. Illustration of Impact of 20 Percent Message Forging on Feedback Comments

0 percent messages forged	20 percent messages forged
Another quick & easy transaction, thanks for your biz!	Another quick & easy transaction, thanks for your biz!
Excellent e-bayer!! fast payment, many thanks!!!	Excellent e-bayer!! fast payment, many thanks!!!
A pleasure to do business with, don't be a stranger!!!	A wonderful buyer. Prompt payment, quick response.
Great to deal with, fast payment.	Great to deal with, fast payment.
A superb e-bayer!!! A real pleasure to do business with.	A superb e-bayer!!! A real pleasure to do business with.

Techniques	% Messages Forged			
	0	10	20	30
Writeprint	**97.88**	88.17	74.23	56.10
PCA	77.49	70.42	64.13	57.91
N-Gram Models	97.60	34.43	21.96	10.53
Markov Models	90.91	11.02	5.54	3.48
Cross Entropy	96.15	**92.18**	**87.03**	**76.94**
K-L Similarity	93.52	65.65	58.53	52.63

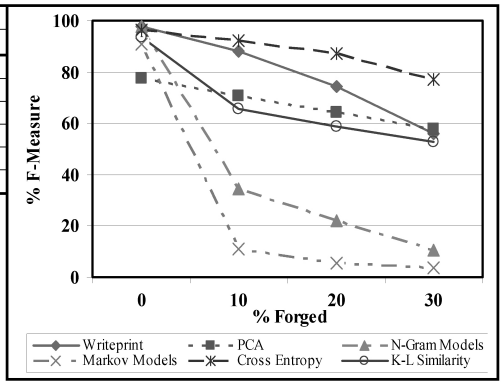

Figure 6. Experiment 2b Results (Robustness Against Message Forging)

that performed marginally better on the forging experiment (2b) as compared to the word substitution experiment (2a).

Table 10 shows the *p*-values for the pairwise *t*-tests on *F*-measure for the experiment evaluating robustness against message forging. Writeprint significantly outperformed n-gram and Markov models, K–L similarity, and PCA for most settings. However, cross entropy significantly outperformed all other techniques including Writeprint and PCA. The improved performance of cross entropy was particularly noticeable at the 20 percent and 30 percent forging levels. The following section provides an analysis of why Writeprint performed poorly on the message forging experiment (2b) as compared to the word substitution experiment (2a) while PCA performed marginally better on message forging (as compared to word substitution, 2a).

Results Discussion for Experiment 2

We analyzed the impact of word substitution–based alteration and forging on the features selected for the altered identities. Because the feature sets are dynamically

Table 10. *p*-Values for Pairwise *t*-Tests on *F*-Measure ($n = 30$)

	Percent forged			
Techniques	0	10	20	30
Writeprint versus PCA	< 0.0001*	< 0.0001*	< 0.0001*	0.0016
Writeprint versus n-gram models	0.1090	< 0.0001*	< 0.0001*	< 0.0001*
Writeprint versus Markov models	< 0.0001*	< 0.0001*	< 0.0001*	< 0.0001*
Writeprint versus cross entropy	< 0.0001*	< 0.0001*	< 0.0001*	< 0.0001*
Writeprint versus K–L similarity	< 0.0001*	< 0.0001*	< 0.0001*	< 0.0001*

* *p*-values significant at corrected threshold alpha/$n = 0.0001$.

generated at the group level (PCA) or individual identity level (Writeprint, n-gram model, cross entropy, K–L similarity) for most of our approaches, word substitution or forging results in a different feature set as compared to no alteration. Thus, the amount of change in the features used by an altered identity as compared to that same identity, devoid of alteration, can shed light on the impact of alteration. We analyzed this by taking the percentage change in altered/forged feature sets from the feature sets used when no alteration/forging was performed. We considered the Writeprint, PCA, n-gram model, cross entropy, and K–L similarity methods. For cross entropy, the features were the match lengths. Markov models were not analyzed because they use a fixed feature set.

Figure 7 shows the impact of word substitution and message forging on the feature sets for the various techniques. Word substitution and forging had a profound impact on character n-gram and word features, resulting in the poor performance of the n-gram models and K–L similarity methods. The cross entropy match lengths also changed considerably; however, the magnitude of the changes was not significant. In other words, although the cross entropy features changed a lot, the manner in which the features are applied is fairly conducive (i.e., insensitive) to word substitution and message forging. For example, a change in the lengths from $\{6, 3, 4\}$ to $\{4, 5, 6\}$ results in 33 percent change in features but only an average match length change of 0.67.

PCA had fewer feature changes for forging as compared to synonym alteration. This was attributable to the fact that PCA used a single feature set. Message forging does not change the overall text across identities, resulting in minimal change in the feature set used by PCA. Consequently, PCA performed better on the forging experiments. In contrast, Writeprint features changed marginally for the word substitution experiment but considerably more for the forging experiments. This resulted in lower accuracy when encountering message forging. For example, the 56.10 percent accuracy for 30 percent forging can be attributed to the fact that 40 percent of the forged identities' features changed. The following paragraph describes why the Writeprint features for the altered identities were generally more effective.

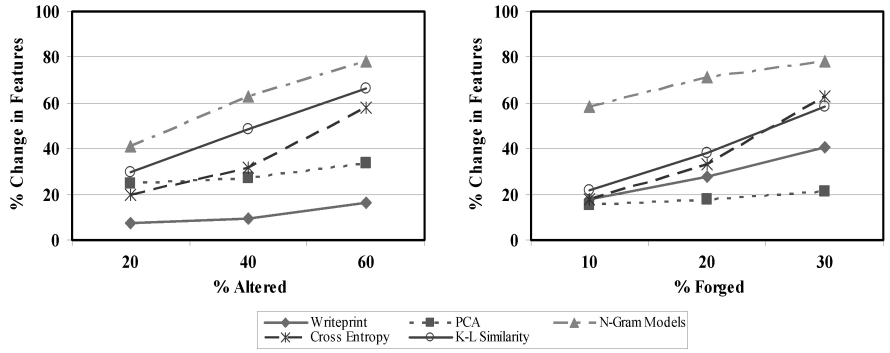

Figure 7. Impact of Word Substitution and Forging on Feature Sets for Various Techniques

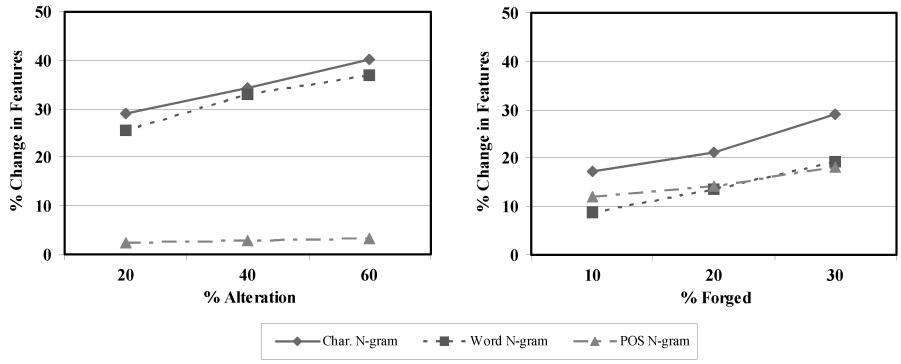

Figure 8. Impact of Word Substitution and Forging on Writeprint N-Gram Features

Figure 8 shows the percentage change in the Writeprint features for character, word, and part-of-speech tag n-grams across the word alteration and forging experiments. For the word alteration experiments, the part-of-speech tag n-grams had minimal change. This led to a reduced impact of word synonym substitution on performance. However, the forging caused considerably higher change in the identities' part-of-speech tags n-gram features, resulting in decreased Writeprint performance for experiment 2b.

Conclusions

IN THIS STUDY, WE DEVELOPED A SYSTEM that can be used for similarity detection of trader feedback comments in online markets. Our research contributions are manyfold. We developed the Writeprint technique that uses Karhunen–Loeve transforms and a novel pattern disruption mechanism to help detect stylistic similarity between traders based on feedback comments. We also incorporated a more comprehensive feature set, allowing improved representation of reputation system feedback comments. Experiments in comparison with existing stylometric techniques demonstrated the scalability and

robustness of the proposed features and technique for differentiating trader identities in online markets. The system proposed in this paper was fairly scalable in terms of number of traders and identities per trader. The approach was also fairly robust against word substitution-based alteration.

The viability of stylometric techniques that can differentiate between hundreds of online traders, coupled with the emergence of large online fraudulent trader databases, has several important research implications. Stylometric analysis techniques can serve as identity authentication systems in online markets, allowing users to compare a potential trading partner against existing fraudulent identities. Such authentication could be especially useful considering that most fraudulent traders engage in such "opportunistic behavior" repeatedly [9], resulting in many documented identities. In the future, we intend to develop such an authentication system that allows individuals to compare traders against hundreds of fraudulent identities collected from various online resources that have emerged in recent years [9]. Moreover, we intend to further enhance the scalability and robustness of the Writeprint-based system using a larger number of online traders. We also plan to investigate the effectiveness of contextual stylometric models segmented temporally or based on genres, emotions, message recipients, or topics.

Acknowledgments: This research was funded in part by the following grant: NSF Information and Data Management Program, "Multilingual Online Stylometric Authorship Identification: An Exploratory Study," 8/2006-8/2007.

References

1. Abbasi, A., and Chen, H. Identification and comparison of extremist-group Web forum messages using authorship analysis. *IEEE Intelligent Systems, 20,* 5 (2005), 67–75.

2. Abbasi, A., and Chen, H. Visualizing authorship for identification. In *Proceedings of the Fourth IEEE Conference on Intelligence and Security Informatics.* Los Alamitos, CA: IEEE Computer Society, 2006, pp. 60–71.

3. Argamon, S.; Saric, M.; and Stein, S. Style mining of electronic messages for multiple authorship discrimination: First results. In L. Getoor, T.E. Senator, P. Domingos, and C. Faloutsos (eds.), *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* New York: ACM Press, 2003, pp. 475–480.

4. Baayen, R.H.; Halteren, H.V.; and Tweedie, F. J. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing, 11,* 3 (1996), 121–132.

5. Bolton, G.E.; Katok, E.; and Ockenfels, A. How effective are electronic reputation mechanisms? An experimental investigation. *Management Science, 50,* 11 (2004), 1587–1602.

6. Brynjolfsson, E., and Smith, M.D. Frictionless commerce? A comparison of Internet and conventional retailers. *Management Science, 46,* 4 (2000), 563–585.

7. Burrows, J.F. Word patterns and story shapes: The statistical analysis of narrative style. *Literary and Linguistic Computing, 2,* 2 (1987), 61–70.

8. Chaski, C.E. Empirical evaluation of language-based author identification techniques. *Forensic Linguistics, 8,* 1 (2001), 1–65.

9. Chua, C.E.H., and Wareham, J. Fighting Internet auction fraud: An assessment and proposal. *IEEE Computer, 37,* 10 (2004), 31–37.

10. Dellarocas, C. The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management Science, 49,* 10 (2003), 1407–1424.

11. De Vel, O.; Anderson, A.; Corney, M.; and Mohay, G. Mining e-mail content for author identification forensics. *ACM SIGMOD Record, 30,* 4 (2001), 55–64.

12. Ding, H., and Samadzadeh, H.M. Extraction of Java program fingerprints for software authorship identification. *Journal of Systems and Software, 72,* 1 (2004), 49–57.

13. Erickson, T., and Kellogg, W.A. Social translucence: An approach to designing systems that support social processes. *ACM Transactions on Computer–Human Interaction, 7,* 1 (2000), 59–83.

14. Fairclough, N. *Analysing Discourse: Textual Analysis for Social Research.* New York: Routledge, 2003.

15. Fellbaum, C. *Wordnet: An Electronic Lexical Database.* Cambridge, MA: MIT Press, 1998.

16. Friedman, E., and Resnick, P. The social cost of cheap pseudonyms. *Journal of Economic Management Strategy, 10,* 1 (2001), 173–199.

17. Garrett, M.F. Production of speech: Observations from normal and pathological language use. In A. Ellis (ed.), *Normality and Pathology in Cognitive Functions*. London: Academic Press, 1982.

18. Grandison, T., and Sloman, M. A survey of trust in Internet applications. *IEEE Communications Surveys and Tutorials, 4,* 4 (2000), 2–16.

19. Gray, A.; Sallis, P.; and MacDonell, S. Software forensics: Extending authorship analysis techniques to computer programs. Paper presented at the Third Biannual Conference of the International Association of Forensic Linguists, Duke University, Durham, NC, September 4–7, 1997.

20. Halliday, M.A.K. *An Introduction to Functional Grammar,* 3d ed. London: Hodder Arnold, 2004.

21. Hayne, C.S., and Rice, E.R. Attribution accuracy when using anonymity in group support systems. *International Journal of Human–Computer Studies, 47,* 3 (1997), 429–452.

22. Hayne, C.S.; Pollard, E.C.; and Rice, E.R. Identification of comment authorship in anonymous group support systems. *Journal of Management Information Systems, 20,* 1 (Summer 2003), 301–329.

23. Holzer, R.; Malin, B.; and Sweeney, L. Email alias detection using social network analysis. In J. Adibi, M. Grobelnik, D. Mladenic, and P. Pantel (eds.), *Proceedings of the Third International Workshop on Link Discovery.* New York: ACM Press, 2005, pp. 52–57.

24. Jackson, D. Stopping rules in principal component analysis: A comparison of heuristical and statistical approaches. *Ecology, 74,* 8 (1993), 2204–2214.

25. Josang, A.; Ismail, R.; and Boyd, C. A survey of trust and reputation systems for online service provision. *Decision Support Systems, 43,* 2 (2007), 618–644.

26. Juola, P. What can we do with small corpora? Document categorization via cross-entropy. In E. Cambouropoulos, U. Hahn, H. Pain, and M. Ramscar (eds.), *Proceedings of the Interdisciplinary Workshop on Similarity and Categorization.* Edinburgh, UK: Department of Artificial Intelligence, Edinburgh University, 1997.

27. Juola, P. The time course of language change. *Computers and the Humanities, 37,* 1 (2003), 77–96.

28. Juola, P., and Baayen, H. A controlled-corpus experiment in authorship identification by cross-entropy. *Literary and Linguistic Computing, 20,* Supplement 1 (2005), 59–67.

29. Kacmarcik, G., and Gamon, M. Obfuscating document stylometry to preserve author anonymity. In N. Calzolari, C. Cardie, and P. Isabelle (eds.), *Proceedings of the Forty-Fourth Annual Meeting of the Association for Computational Linguistics.* East Straudsburg, PA: Association for Computational Linguistics, 2006, pp. 444–451.

30. Keselj, V.; Peng, F.; Cercone, N.; and Thomas, C. N-gram based author profiles for authorship attribution. In *Proceedings of the Sixth Conference of the Pacific Association for Computational Linguistics.* Hoboken, NJ: John Wiley and Sons, 2003, pp. 255–264.

31. Khmelev, D.V. Disputed authorship resolution using relative entropy for Markov chains of letters in human language texts. *Journal of Quantitative Linguistics, 7,* 3 (2000), 115–126.

32. Khmelev, D.V., and Tweedie, F.J. Using Markov chains for identification of writers. *Literary and Linguistic Computing, 16,* 3 (2001), 299–307.

33. Kjell, B.; Woods, W.A.; and Frieder, O. Discrimination of authorship using visualization. *Information Processing and Management, 30,* 1 (1994), 141–150.

34. Koppel, M., and Schler, J. Exploiting stylistic idiosyncrasies for authorship attribution. In A.G. Cohn (ed.), *Proceedings of the IJCAI Workshop on Computational Approaches to Style Analysis and Synthesis.* San Francisco: Morgan Kaufmann, 2003, pp. 69–72.

35. Koppel, M.; Akiva, N.; and Dagan, I. Feature instability as a criterion for selecting potential style markers. *Journal of the American Society for Information Science and Technology, 57,* 11 (2006), 1519–1525.

36. Lee, Z.; Im, L.; and Lee, S. J. The effect of negative buyer feedback on prices in Internet auction markets. In S. Ang, H. Krcmar, W.J. Orlikowski, P. Weill, and J.I. DeGross (eds.), *Proceedings of the Twenty-First International Conference on Information Systems.* Atlanta, GA: Association for Information Systems, 2000, pp. 286–287.

37. Li, J.; Zheng, R.; and Chen, H. From fingerprint to Writeprint. *Communications of the ACM, 49,* 4 (2006), 76–82.

38. Mui, L.; Mohtashemi, M.; and Halberstadt, A. A computational model of trust and reputation. In R.H. Sprague (ed.), *Proceedings of the Thirty-Fifth Annual Hawaii International Conference on System Sciences.* Los Alamitos, CA: IEEE Computer Society Press, 2002 (available at http://csdl2.computer.org/comp/proceedings/hicss/2002/1435/07/14350188.pdf).

39. Novak, J.; Raghavan, P.; and Tomkins, A. Anti-aliasing on the Web. In S.I. Feldman, M. Uretsky, M. Najork, and C.E. Wills (eds.), *Proceedings of the Thirteenth International World Wide Web Conference.* New York: ACM Press, 2004, pp. 30–39.

40. Pavlou, P.A., and Gefen, D. Building effective online marketplaces with institution-based trust. *Information Systems Research, 15,* 1 (2004), 37–59.

41. Peng, F.; Schuurmans, D.; Keselj, V.; and Wang, S. Automated authorship attribution with character level language models. In A. Copestake and J. Hajic (eds.), *Proceedings of the Tenth Conference of the European Chapter of the Association for Computational Linguistics.* East Straudsburg, PA: Association for Computational Linguistics, 2003, pp. 267–274.

42. Rao, J.R., and Rohatgi, P. Can pseudonymity really guarantee privacy? In S. Bellovin and G. Rose (eds.), *Proceedings of the Ninth USENIX Security Symposium.* Berkeley, CA: USENIX Association, 2000, pp. 85–96.

43. Rasmusson L., and Jansson, S. Simulated social control for secure Internet commerce. In H. Hosmer, J. Dobson, C. Meadows, and D. Bailey (eds.), *Proceedings of the ACM SIGSAC Workshop on New Security Paradigm.* New York: ACM Press, 1996, pp. 18–25.

44. Resnick, P.; Zeckhauser, R.; Friedman, E.; and Kuwabara, K. Reputation systems. *Communications of the ACM, 43,* 12 (2000), 45–48.

45. Resnick, P.; Zeckhauser, R.; Swanson, J.; and Lockwood, K. The value of reputation on eBay: A controlled experiment. *Experimental Economics, 9,* 2 (2006), 79–101.

46. Rudman, J. The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities, 31,* 4 (1997), 351–365.

47. Webb, A. *Statistical Pattern Recognition.* New York: John Wiley and Sons, 2000.

48. Yang, Y., and Pedersen, J. O. A comparative study on feature selection in text categorization. In D.H. Fisher (ed.), *Proceedings of the Fourteenth International Conference on Machine Learning.* San Francisco: Morgan Kaufmann, 1997, pp. 412–420.

49. Zheng, R.; Qin, Y.; Huang, Z.; and Chen, H. A framework for authorship analysis of online messages: Writing-style features and techniques. *Journal of the American Society for Information Science and Technology, 57,* 3 (2006), 378–393.