2012 IEEE Symposium on Security and Privacy

Detecting Hoaxes, Frauds, and Deception in Writing Style Online

Sadia Afroz*, Michael Brennan* and Rachel Greenstadt*
*Department of Computer Science
Drexel University, Philadelphia, PA 19104
Emails: sadia.afroz@drexel.edu, mb553@drexel.edu and greenie@cs.drexel.edu

***Abstract*—In digital forensics, questions often arise about the authors of documents: their identity, demographic background, and whether they can be linked to other documents. The field of stylometry uses linguistic features and machine learning techniques to answer these questions. While stylometry techniques can identify authors with high accuracy in non-adversarial scenarios, their accuracy is reduced to random guessing when faced with authors who intentionally obfuscate their writing style or attempt to imitate that of another author. While these results are good for privacy, they raise concerns about fraud. We argue that some linguistic features change when people hide their writing style and by identifying those features, stylistic deception can be recognized. The major contribution of this work is a method for detecting stylistic deception in written documents. We show that using a large feature set, it is possible to distinguish regular documents from deceptive documents with 96.6% accuracy (F-measure). We also present an analysis of linguistic features that can be modified to hide writing style.**

***Keywords*-stylometry; deception; machine learning; privacy;**

I. Introduction

When an American male blogger Thomas MacMaster posed as a Syrian homosexual woman Amina Arraf in the blog "A Gay Girl in Damascus" and wrote about Syrian political and social issues, several news media including The Guardian and CNN thought the blog was "brutally honest," and published email interviews of Amina[1]. Even though no one had ever spoken to or met her and no Syrian activist could identify her, "Amina" quickly became very popular as a blogger. When "Amina's cousin" announced that she had been abducted by the Syrian police, thousands of people supported her on social media and made the US state department investigate her fictional abduction[2]. This scrutiny led to the discovery of the hoax.

The authenticity of a document (or blog post) depends on the authenticity of its source. Stylometry can help answering the question, "who is the writer of a particular document?" Many machine learning based methods are available to recognize authorship of a written document based on linguistic style. Stylometry is important to security researchers as it is a forensics technique that helps detect authorship of unknown documents. If reliable, it can be used to provide attribution for attacks, especially when attackers release manifestos explaining their actions. It is also important to privacy research, as it is necessary to hide the indications of authorship to achieve anonymity. Writing style as a marker of identity is not addressed in current privacy and anonymity tools. Given the high accuracy of even basic stylometry systems this is not a topic that can afford to be overlooked.

In the year 2000, Rao and Rohatgi questioned whether pseudonymity could provide privacy, showing that linguistic analysis could identify anonymous authors on sci.crypt by comparing their writing to attributed documents in the RFC database and on the IPSec mailing list [23]. In the intervening years, linguistic authorship iden-

[1] http://www.telegraph.co.uk/news/worldnews/middleeast/syria/8572884/A-Gay-Girl-in-Damascus-how-the-hoax-unfolded.html

[2] http://www.foxnews.com/world/2011/06/07/gay-girl-in-damascus-blogger-kidnapped-by-syrian-forces/

DOI 10.1109/SP.2012.34

tification techniques have improved in accuracy and scale to handle over fifty potential authors with over 90% accuracy [1], and even 100,000 authors with significant accuracy [17]. At the same time there has been an explosion in user-generated content to express opinions, to coordinate protests against repressive regimes, to whistle blow, to commit fraud, and to disclose everyday information.

These results have not been lost on the law enforcement and intelligence communities. The 2009 Technology Assessment for the State of the Art Biometrics Excellence Roadmap (SABER) commissioned by the FBI stated that, "As non-handwritten communications become more prevalent, such as blogging, text messaging and emails, there is a growing need to identify writers not by their written script, but by analysis of the typed content [29]."

Brennan and Greenstadt [2] showed that current authorship attribution algorithms are highly accurate in the non-adversarial case, but fail to attribute correct authorship when an author deliberately masks his writing style. Their work defined and tested two forms of adversarial attacks: imitation and obfuscation. In the imitation attack, authors hide their writing style by imitating another author. In the obfuscation attack, authors hide their writing style in a way that will not be recognized. Traditional authorship recognition methods perform less than random chance in attributing authorship in both cases. These results were further confirmed by Juola and Vescovi [10]. These results show that effective stylometry techniques need to recognize and adapt to deceptive writing.

We argue that some linguistic features change when people hide their writing style and by identifying those features, deceptive documents can be recognized. According to Undeutsch Hypothesis [26] "Statements that are the product of experience will contain characteristics that are generally absent from statements that are the product of imagination." Deception requires additional cognitive effort to hide information, which often introduces subtle changes in human behavior [6]. These behavioral changes affect verbal and written communication. Several linguistic cues were found to discriminate deceptive communication from truthful communication. For example, deceivers use fewer long sentences, fewer average syllables per word and simpler sentences than truth tellers [3]. Thus, deceptive language appears to be less complex and easier to comprehend. Our analysis shows that though stylistic deception is not lying, similar linguistic features change in this form of deception.

The goal of our work is to create a framework for detecting the indication of masking in written documents. We address the following research questions:

1) Can we detect stylistic deception in documents?
2) Which linguistic features indicate stylistic deception?
3) Which features do people generally change in adversarial attacks and which features remain unchanged?
4) Does stylistic deception share similar characteristics with other deceptions?
5) Are some adversarial attacks more difficult to detect than others?
6) Can we generalize deception detection?

This work shows that using linguistic and contextual features, it is possible to distinguish stylistic deception from regular writing with 96.6% accuracy (F-measure) and identify different types of deception (imitation vs. obfuscation) with 87% accuracy (F-measure). Our contributions include a general method for distinguishing stylistic deception from regular writing, an analysis of long-term versus short-term deception, and the discovery that stylistic deception shares similar features with lying-type deception (and can be identified using the linguistic features used in lying detection).

We perform analysis on the Brennan-Greenstadt adversarial dataset and a similar dataset collected using Amazon Mechanical Turk (AMT)[3]. We show that linguistic cues that can detect stylistic deception in the Extended-Brennan-Greenstadt adversarial dataset can detect indication of masking in the documents collected from the Ernest Hemingway and William Faulker imitation contests. We also

[3]https://mturk.amazon.com

show how long-term deceptions such as the blog posts from "A Gay Girl in Damascus" are different from these short-term deceptions. We found these deceptions to be more robust to our classifier but more vulnerable to traditional stylometry techniques.

The remainder of the paper is organized as follows. In section 2, we discuss related works in adversarial stylometry. Section 3 explains how deception detection is different from regular authorship recognition. Section 4 describes our analytic approach of detecting deception. In section 5, we describe our data collection methods and datasets. We follow with our results of detecting deception on different datasets in Section 6 and discuss their implications in Section 7.

II. Related work

The classic example in the field of stylometry is the Federalist Papers. 85 papers were published anonymously in the late 18th century to persuade the people of New York to ratify the American Constitution. The authorship of 12 of these papers was heavily contested [18]. To discover who wrote the unknown papers, researchers have analyzed the writing style of the known authors and compared it to that of the papers with unknown authorship. The features used to determine writing styles have been quite varied. Original attempts used the length of words, whereas later attempts used pairs of words, vocabulary usage, sentence structure, function words, and so on. Most studies show the author was James Madison.

Several resources give an overview of stylometry methods [14], [28], and describe the state of the field as it relates to computer science and computer linguistics [11] or digital forensics [5]. Artificial Intelligence has been embraced in the field of stylometry, leading to more robust classifiers using machine learning and other AI techniques [9], [27]. There has also been some work on circumventing attempts at authorship attribution [12], [23], using stylometry to deanonymize conference reviews [16], and looking at stylometry as a method of communication security [4], but these works do not deal with malicious attempts to circumvent a specific method. Some research has looked at imitation of authors. Somers [25] compared the work of Gilbert Adair's literary imitation of Lewis Carroll's Alice in Wonderland, and found mixed results. Other work looks into the impact of stylometry on pseudonymity [23] and author segmentation [1].

Most authorship recognition methods are built on the assumption that authors do not make any intentional changes to hide their writing style. These methods fail to detect authorship when this assumption does not hold [2], [10]. The accuracy of detecting authorship decreases to random guessing in the case of adversarial attacks.

Kacmarcik and Gamon explored detecting obfuscation by first determining the most effective function words for discriminating between text written by Hamilton and Madison, then modifying the feature vectors to make the documents appear authored by the same person. The obfuscation was then detected with a technique proposed by Koppel and Scher, "unmasking," that uses a series of SVM classifiers where each iteration of classification removes the most heavily weighted features. The hypothesis they put forward (validated by both Koppel and Scher [13] and Kacmarcik and Gamon [12]) is that as features are removed, the classifier's accuracy will slowly decline when comparing two texts from different authors, but accuracy will quickly drop off when the same is done for two texts by the same author (where one has been modified). It is the quick decline in accuracy that shows there is a deeper similarity between the two authors and indicates the unknown document has most likely been modified.

However, the above work has some significant limitations. The experiments were performed on modified feature vectors, not on modified documents or original documents designed with obfuscation in mind. Further, the experiments were limited to only two authors, Hamilton and Madison, and on the Federalist Papers data set. It is unclear whether the results generalize to actual documents, larger author sets and modern data sets.

We analyzed the differences between the control and deceptive passages on a feature-by-feature

basis and used this analysis to determine which features authors often modify when hiding their style, designing a classifier that works on actual, modern documents with modified text, not just feature vectors.

III. Difference with regular authorship recognition

Deception detection is challenging because though authorship recognition is a well-studied problem, none of the current algorithms are robust enough to detect stylistic deception and perform close to random chance if the author changes his usual writing style. This problem is quite different from distinguishing one author's samples from others. In supervised authorship recognition, a classifier is trained on the sample documents of different authors to build a model that is specific to each author. In deception detection, we trained a classifier on regular and deceptive documents to model the generic characteristic of regular and deceptive documents.

Our classifier is trained on the Extended-Brennan-Greenstadt dataset where participants spent 30 minutes to an hour on average to write documents in a style different from their own. Our test set also consists of imitated documents from the Ernest Hemingway and William Faulkner imitation contests. We investigated the effect of long-term deception using obfuscated documents from a deceptive blog. The skill levels of the participants in these datasets are varied. In the Brennan-Greenstadt dataset, the participants were not professional writers and they had a pre-specified topic to develop a different writing style, but authors in the imitation contests and fictional blog were mostly professional writers and had enough time and chose a topic of their choice to express themselves in a different voice other than their own. The fact that the classifier, trained on the Extended-Brennan-Greenstadt dataset, can detect deception in the test sets—though at lower accuracy—generalizes the underlying similarity among different kinds of deception.

IV. Analytic Approach

Our goal is to determine whether an author has tried to hide his writing style in a written document. In traditional authorship recognition, authorship of a document is determined using linguistic features of an author's writing style. In deceptive writing, when an author is deliberately hiding his regular writing style, authorship attribution fails because the deceptive document lacks stylistic similarity with the author's regular writing style. Though recognizing correct authorship of a deceptive document is hard, our goal is to see if it is possible to discriminate deceptive documents from regular documents.

To detect adversarial writing, we need to identify a set of discriminating features that distinguish deceptive writing from regular writing. After determining these features, supervised learning techniques can be used to train and generate classifiers to classify new writing samples.

A. Feature selection

The performance of stylometry methods depends on the combination of the selected features and analytical techniques. We explored three feature sets to identify stylistic deception.

Writeprints feature set: Zheng et al. proposed the Writeprints features that can represent an author's writing style in relatively short documents, especially in online messages [30]. These "kitchen sink" features are not unique to this work, but rather represent a superset of the features used in the stylometry literature. We used a partial set of the Writeprints features, shown in Table I.

Our adaptation of the Writeprints features consists of three kinds of features: lexical, syntactic, and content specific. The features are described below:

Lexical features: These features include both character-based and word-based features. These features represent an author's lexicon-related writing style: his vocabulary and character choice. The feature set includes total characters, special character usage, and several word-level features such as total words, characters per word, frequency of large words, unique words.

Syntactic features: Each author organizes sentences differently. Syntactic features represent an author's sentence-level style. These features include frequency of function words, punctuation and parts-of-speech (POS) tagging. We use the list of function words from LIWC 2007 [19].

Content Specific features: Content specific features refer to keywords for a specific topic. These have been found to improve performance of authorship recognition in a known context [1]. For example, in the spam context, spammers use words like "online banking" and "paypal;" whereas scientific articles are likely to use words related to "research" and "data."

Our corpus contains articles from a variety of contexts. It includes documents from business and academic contexts, for example school essays and reports for work. As our articles are not from a specific context, instead of using words of any particular context we use the most frequent word n-grams as content-specific features. As we are interested in content-independent analytics, we also performed experiments where these features were removed from the feature set.

Table I: **Writeprints feature set**

Category	Quantity	Description
Character related	90	Total characters, percentage of digits, percentage of letters, percentage of uppercase letters, etc. and frequency of character unigram, most common bigrams and tri-grams
Digits, special characters, punctuations	39	Frequency of digits (0-9), special characters(e.g., %,&, *) and punctuations
Word related	156	Total words, number of characters per word, frequency of large words, etc. Most frequent word uni-/bi-/ tri-grams
Function words and parts-of-speech	422	frequency of function words and parts-of-speech

Lying-detection feature set: Our feature set includes features that were known to be effective in detecting lying type deception in computer mediated communications and typed documents [3], [8]. These features are:

1) Quantity (number of syllables, number of words, number of sentences),
2) Vocabulary Complexity (number of big words, number of syllables per word),
3) Grammatical Complexity (number of short sentences, number of long sentences, Flesh-Kincaid grade level, average number of words per sentence, sentence complexity, number of conjunctions),
4) Uncertainty (Number of words express certainty, number of tentative words, modal verbs)
5) Specificity and Expressiveness (rate of adjectives and adverbs, number of affective terms),
6) Verbal Non-immediacy (self-references, number of first, second and third person pronoun usage).

We use the list of certainty, tentative and affective terms from LIWC 2007 [19].

9-feature set (authorship-attribution features): This minimal feature set consists of the nine features that were used in the neural network experiments in Brennan's 2009 paper [2]. The features are: number of unique words, complexity, Gunning-Fog readability index, character count without whitespace, character count with whitespace, average syllables per word, sentence count, average sentence length, and Flesch-Kincaid readability score.

B. Classification

We represent each document as $(\vec{x}, y)$ where $\vec{x} \in \mathbb{R}^n$ is a vector of n features and $y \in \{Regular, Imitation, Obfuscation\}$ is the type of the document. In our study, $n = 9$ for 9-features, $n = 20$ for lying-detection features and $n = 707$ for the Writeprints features. For classification, we used Support Vector Machine (SVM) with Sequential Minimal Optimization (SMO) [20] implemented in the WEKA tool [7] with a polynomial kernel. We tested our dataset with other classifiers in the WEKA tool such as k-Nearest Neighbor, Naive Bayes, J48 Decision Tree, Logistic Regression and SVM with RBF kernel. We chose to focus on the SMO SVM as it outperformed other

classifiers in most of the test cases. The exception to this is the Lying-detection feature set, in which a J48 Decision Tree[4] outperformed SVMs. J48 is the JAVA implementation of C4.5 algorithm for constructing decision tree [22]. It is notable that the original work using these features also used a decision tree [3].

V. Data Collection

We present results on three datasets. The first one is the Extended-Brennan-Greenstadt corpus which contains the Brennan-Greenstadt corpus, extended with regular, obfuscated and imitated writing samples of the AMT workers. The second dataset, which we call the Hemingway-Faulkner Imitation corpus, contains articles from the International Imitation Hemingway Competition and Faux Faulkner contest. The last dataset, Thomas-Amina Hoax corpus, contains blog posts from "A Gay Girl in Damascus" blog, posts of Thomas MacMaster in the alternate-history Yahoo! group[5] as himself and as Amina Arraf, and writing samples of Britta Froelicher, a graduate student at Center for Syrian Studies at St Andrews, who is also Thomas's wife.

A. Extended-Brennan-Greenstadt corpus

We used the Brennan-Greenstadt adversarial corpus for this study [6]. This dataset consists of two types of writing samples, regular and adversarial, from 12 participants. The regular writing contains approximately 5000 words of pre-existing writing samples per author. The regular writings are formal in nature, written for business or academic purposes. In the adversarial writing samples, participants performed two adversarial attacks: obfuscation and imitation. In the obfuscation attack, each participant tried to hide his identity while writing a 500-word article describing his neighborhood. In the imitation attack, each participant tried to hide his writing style by imitating Cormac McCarthy's writing style in 'The Road' and wrote a 500-word article describing a day of their life in the third person.

We extended this corpus by collecting similar writing samples using AMT. We created a Human Intelligence Task (HIT) where participants were asked to submit the three kinds of writing sample described in the previous paragraph [7]. After collecting the data, we manually verified each submission and only accepted the ones that complied with our instructions. 56 participants' work was accepted.

Participants were also asked to provide their demographic information. According to the provided demographic information, all the participants' native language is English and all of them have some college-level degree.

A total of 68 authors' writing samples are used in this study, 12 of the authors are from the Brennan-Greenstadt corpus and others are the AMT workers.

B. Hemingway-Faulkner Imitation corpus

The Hemingway-Faulkner Imitation corpus consists of the winning articles from the Faux Faulkner Contest and International Imitation Hemingway Competition[8]. The International Imitation Hemingway Competition is an annual writing competition where participants write articles by imitating Ernest Hemingway's writing style. In the Faux Faulkner Contest participants imitate William Faulkner's artistic style of writing, his themes, his plots, or his characters. Each article is at most 500 words long. We collected all publicly available winning entries of the competitions from 2000 to 2005. The corpus contains sixteen 500-word excerpts from different books of Ernest Hemingway, sixteen 500-word excerpts from different books of William Faulkner, 18 winning articles from The International Imitation Hemingway Competition and 15 winning articles from The Faux Faulkner Contest.

In the imitation contests, participants chose different topics and imitated from different novels of the original authors. Table II, III, and IV show imitation samples. Cormac McCarthy imitation sam-

[4]http://weka.sourceforge.net/doc/weka/classifiers/trees/J48.html
[5]http://groups.yahoo.com/group/alternate-history/
[6]This data set is publicly available at https://psal.cs.drexel.edu
[7]https://www.cs.drexel.edu/ sa499/amt/dragonauth_index.php.
[8]Available at http://web.archive.org/web/20051119135221/ http://www.hemispheresmagazine.com/fiction/2005/hemingway.htm

ples are all of same topic but the contest articles are of varied topics and most of winners were professional writers.

Table II: Imitation samples from the Extended-Brennan-Greenstadt dataset.

Cormac McCarthy imitation sample: 1 Laying in the cold and dark of the morning, the man was huddled close. Close to himself in a bed of rest. Still asleep, an alarm went off. The man reached a cold and pallid arm from beneath the pitiful bedspread.
Cormac McCarthy imitation sample: 2 She woke up with a headache. It was hard to tell if what had happened yesterday was real or part of her dream because it was becoming increasingly hard to tell the two apart. The day had already started for most of the world, but she was just stumbling out of bed. Across the hall, toothbrush, shower.

Table III: Imitation samples from the International Imitation Hemingway Competition.

Hemingway imitation sample: 1 At 18 you become a war hero, get drunk, and fall in love with a beautiful Red Cross nurse before breakfast. Over absinthes you decide to go on safari and on your first big hunt you bag four elephants, three lions, nine penguins, and are warned never to visit the Bronx Zoo again. Later, you talk about the war and big rivers and dysentery, and in the morning you have an urge to go behind a tree and get it all down on paper.
Hemingway imitation sample: 2 He no longer dreamed of soaring stock prices and of the thousands of employees who once worked for him. He only dreamed of money now and the lions of industry: John D. Rockefeller, Jay Gould and Cornelius Vanderbilt. They played in the darkened boardrooms, gathering money in large piles, like the young wolves he had hired.

C. Long Term Deception: Thomas-Amina Hoax corpus

In 2010, a 40-year old US citizen Thomas MacMaster opened a blog "A Gay Girl in Damacus" where he presented himself as a Syrian-American homosexual woman Amina Arraf and published blogposts about political and social issues in Syria. Before opening the blog, he started posting as Amina Arraf in the alternate-history Yahoo! group since early 2006. We collected twenty 500-word posts of Amina and Thomas from the alternate-history Yahoo! group, publicly available articles written by Britta Froelicher[9] who was a suspect of this hoax and 142 blog posts from "A Gay Girl in Damacus." The blog posts were divided into 500-word chunks. In total we had 248 articles.

[9]One such article: http://www.joshualandis.com/blog/?p=1831

Table IV: Imitation samples from the Faux Faulkner Contest.

William Faulkner imitation sample: 1 And then Varner Pshaw in the near dark not gainsaying the other but more evoking privilege come from and out of the very eponymity of the store in which they sat and the other again its true I seen it and Varner again out of the near dark more like to see a mule fly and the other himself now resigned (and more than resigned capitulate vanquished by the bovine implacable will of the other) Pshaw in final salivary resignation transfixed each and both together on a glowing box atop the counter.
William Faulkner imitation sample: 2 From a little after breakfast until almost lunch on that long tumid convectionless afternoon in a time that was unencumbered by measure (and before you knew to call it time: when it was just the great roiling expressionless moment known only elliptically and without reference to actual clocks or watches as When We Were Very Young) Piglet, his eyes neither seeing nor not-seeing, stood motionless as if riveted to the iron landscape from which he had equally motionlessly emerged until he became the apotheosis of all tiny pigs wearing scarves standing on two legs and doing absolutely nothing but blinking at what lay before them in the dust.

VI. Evaluation and Results

A. Evaluation Methodology

To evaluate our approach, we perform a three class classification, where the three classes are Regular, Imitation, and Obfuscation, with three feature sets. We use 10-fold cross-validation with Support Vector Maching (SVM) and J48 Decision Tree classifier. In k-fold cross-validation the original dataset is randomly partitioned into k equal folds or subsets. Then the classifier is trained on k-1 subsets and tested on the remaining one subset. The cross-validation process is then repeated k times (the folds), with each of the k folds used exactly once as the validation data. The final estimation of the classifier is the average of the k results from the folds.

We also split the Writeprints feature set into three sets, lexical, syntactic, and content specific and run the three class classification using each set separately. In this experiment, we show that non-specific features are as effective as content specific

features in detecting adversarial attacks.

B. Results

1) **Can we detect stylistic deception in documents?:** The results show that a classifier trained on sets of adversarial and non-adversarial documents can detect deceptive documents with 96.6% accuracy on our best feature set, as is shown in Table V.

The classification was performed on the Brennan-Greenstadt dataset, Amazon Mechanical Turk dataset, and Extended-Brennan-Greenstadt dataset which combines both of the sets. On all of the datasets, the Writeprints features showed the best performance in detecting adversarial attacks. With this feature set, an SVM classifier can detect imitation attacks with 85% accuracy and obfuscation attacks with 89.5% accuracy on the Extended-Brennan-Greenstadt dataset.

Though deception in writing style is significantly different from lying, both deceptions have similar characteristics. With the Lying-detection features we can detect imitation attacks with 75.3% accuracy and obfuscation attacks with 59.9% accuracy.

The 9-feature set, which can detect authorship of regular documents with over 90% accuracy, performed poorly (less than 50%) in detecting adversarial attacks.

The type of machine learning method used in classification is another important factor in detecting deception. The SVM classifier worked best with the Writeprints features whereas the J48 decision tree performed well with the Lying-detection features.

2) **Which linguistic features indicate stylistic deception?:** To understand the effect of different features, we rank the Writeprints features based on their Information Gain Ratio (IGR) [21]. IGR of a feature f_j in class C_i is calculated using the following formula,

$$IGR(C_i, f_j) = (H(C_i) - H(C_i|f_j))/H(f_j),$$

where H is entropy. The top features are mostly function words, as shown in Table VI. Other than function words, some syntactic features such as personal pronoun, adverbs, adjectives, and average word length were some of the most discriminating features.

In our dataset, the non-content-specific features performed similar to the content-specific features in detecting deception, as shown in Figure 1, which suggests the possibility of generalizing these features to detect multiple forms of adversarial attacks.

Table VI: This table shows the features that discriminate deceptive documents from regular documents. The top discriminating features are mostly function words.

Top 20 features	
Imitated documents	**Obfuscated documents**
whats	alot
atop	near
lately	up
wanna	theres
underneath	thousand
anymore	ours
beside	shall
she	thats
herself	cuz
beneath	whats
like	havent
he	Frequency of comma
till	lots
her	tons
onto	anyway
soon	plus
Frequency of dot	other
Personal pronoun	maybe

3) **Which features do people generally change in adversarial attacks and which features remain unchanged?:** We analysed the Extended-Brennan-Greenstadt dataset to understand which features people change in stylistic deception. We computed change in a feature f (C_f) in regular and adversarial documents using the following formula:

$$C_f = 100 * (f_{adv} - f_{reg})/(f_{reg} + 1) \quad (1)$$

where, f_{adv} and f_{reg} are the average values of feature f in the adversarial documents and regular documents respectively. We added 1 with f_{reg} in

Table V: The table shows performance of different feature sets in detecting regular and adversarial writing samples. The Writeprints feature set with SVM classifier provides the best performance in detecting deception.

Dataset	Feature set, Classifier	Type	Precision	Recall	F-measure	Overall F-measure
Extended-Brennan-Greenstadt	**Writeprints, SVM**	**Regular**	**97.5%**	**98.5%**	**98%**	**96.6%**
		Imitation	**87.2%**	**82.9%**	**85%**	
		Obfuscation	**93.2%**	**86.1%**	**89.5%**	
	Lying-detection, J48	Regular	95.2%	96.2%	95.7%	92%
		Imitation	80.6%	70.7%	75.3%	
		Obfuscation	60.3%	59.5%	59.9%	
	9-feature set, J48	Regular	92.3%	96.8%	94.5%	89%
		Imitation	52.9%	43.9%	48%	
		Obfuscation	61.9%	32.9%	43%	
Amazon Mechanical Turk	Writeprints, SVM	Regular	96.5%	98.6%	97.5%	95.6%
		Imitation	82.3%	72.9%	77.3%	
		Obfuscation	96.4%	79.1%	86.9%	
	Lying-detection, J48	Regular	94.2%	96.2%	95.2%	90.9%
		Imitation	71.7%	54.3%	61.8%	
		Obfuscation	58.5%	56.7%	57.6%	
	9-feature set, J48	Regular	92.5%	96.3%	94.3%	88%
		Imitation	45.5%	35.7%	40%	
		Obfuscation	45.5%	29.9%	36%	
Brennan-Greenstadt	Writeprints, SVM	Regular	94%	100%	96.9%	94.7%
		Imitation	100%	83.3%	90.9%	
		Obfuscation	100%	50%	66.7%	
	Lying-detection, J48	Regular	90%	92.9%	91.4%	85.3%
		Imitation	90.9%	83.3%	87%	
		Obfuscation	11.1%	8.3%	9.5%	
	9-feature set, J48	Regular	89.4%	93.7%	91.5%	84%
		Imitation	25%	25%	25%	
		Obfuscation	83.3%	41.7%	55.6%	

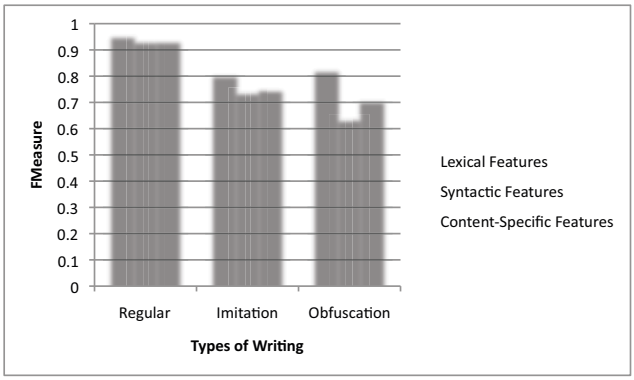

Figure 1: Effect of different feature types in detecting adversarial attacks. In each case, SVM is used for classification.

the denominator to avoid divide-by-zero error, as f_{reg} can be zero for some features.

Figure 2 shows the mostly changed features among parts-of-speech and 9-features. In Figure 3, we grouped similar Writeprint features together and added their corresponding changes to show the overall change. In each graph, the y-axis contains a list of features that have been adjusted in the passages and x-axis of these graphs denotes the change in each feature.

In our experiments, participants used more existential *there*[10] and adverbs in obfuscated passages. We also noticed an increase in the usage of particles and personal pronouns and decrease in the usage of noun, and Wh-pronoun. We do see hints that authors are "dumbing down" their writing style with shorter sentences, simpler words with less syllables, lower readability scores and higher readability ease (Figure 2). Another interesting result is that participants used a large number of function

[10]Existential *there* (EX) is the unstressed *there* that triggers inversion of the inflected verb and the logical subject of a sentence [24]. For example: There/EX were lots of people in town.

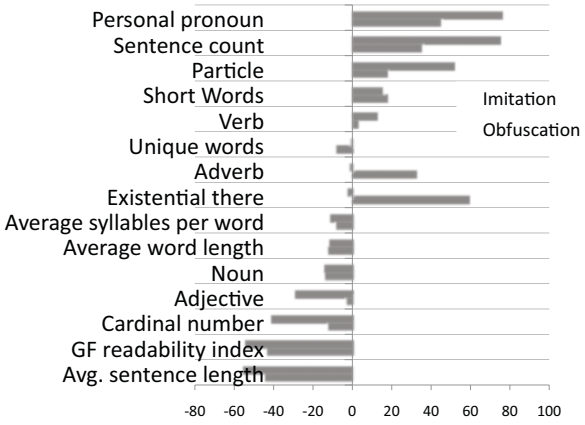

Figure 2: Mostly changed features among parts-of-speech and 9-features.

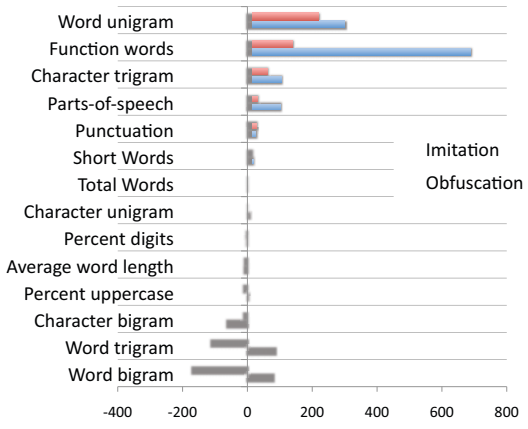

Figure 3: Overall changes in the Writeprint features.

words to obfuscated their writings (Figure 3). This suggests that changing function words might be an important way to anonymize a document. In our dataset, the function words that were used more in obfuscated documents than regular documents are: I, my, there, are, you. The function words that were used less are: as, his, by, her, be.

In imitated passages, more usage of personal pronouns, particles and less usage of adjectives were noticed (shown in Figure 2). Participants also used shorter sentences and words with less syllables which is why the imitated passages had less readability score (Figure 2). The personal pronouns can likely be attributed to the direction of the writing survey which asked participants to describe their day in the third person. The shorter sentences are likely a result of imitating Cormac McCarthy who often uses short sentences in his prose. Participants were found to use more function words than regular in this case also, although not as much as in the obfuscated documents.

Information Gain Ratio (IGR) and usage change of a feature are positively correlated (shown in Figure 5 and Figure 4). All of the highly changed features have high IGR and unchanged features have zero IGR. But some function words were found to have high information gain ratio in spite of being changed very little.

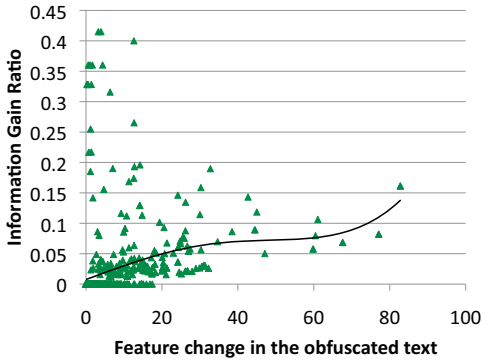

Figure 4: Relation between Information Gain Ratio and absolute usage change in features in the obfuscated documents.

4) **Does stylistic deception share similar characteristics with other deceptions?:** Stylistic deception shares similar characteristics with lying. As shown in Table V, features used to detect lying can detect imitation attacks with 75.3% accuracy and obfuscation attacks with 59.9% accuracy on the Extended-Brennan-Greenstadt dataset. This result indicates the similarity of linguistic changes associated with any kind of information hiding.

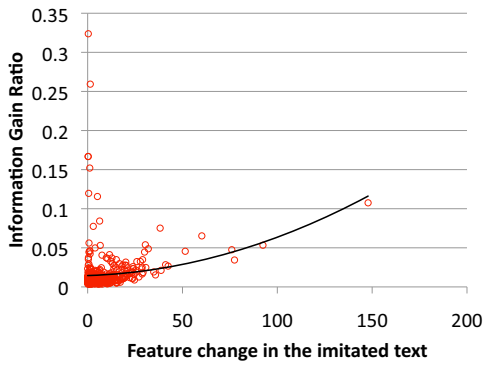

Figure 5: Relation between Information Gain Ratio and absolute usage change in features in the imitated documents.

5) **Are some adversarial attacks more difficult to detect than others?:** In our experiment, obfuscation was harder to detect than imitation. In imitation, participants followed one specific writing style, the writing style of Cormac McCarthy. Different people followed different linguistic aspects in imitating him, for example, some participants used short sentences, some used descriptive adjectives and some used a conversational format with dialogs. But the overall writing style was limited to the style of Cormac McCarthy. Obfuscation is different than imitation as in obfuscation an author can choose to imitate more than one authors' writing style or develop a new style different from his own. However, when we include multiple imitated authors it becomes correspondingly more difficult to detect imitation attacks.

6) **Can we generalize deception detection?:** We check whether our deception detection approach that can detect imitation and obfuscation on the Extended-Brennan-Greenstadt can detect imitation samples from the Ernest Hemingway and William Faulkner imitation contests. Our performed a 10-fold cross-validation on the Hemingway-Faulkner imitation corpus. We used Writeprints and Lying-detection features with SVM and J48 classifiers respectively from the WEKA tool. Our classifier can distinguish imitated articles from the original writings of Ernest Hemingway and William Faulkner with 88.6% accuracy (Table VII).

Table VII: Imitated document prediction result: Hemingway-Faulkner imitation corpus. (P = Precision, R= Recall and F= F-measure)

Type	Lying-detection, J48			Writeprints, SVM		
	P	R	F	P	R	F
Imitation	69.7%	69.7%	69.7%	83.8%	93.9%	88.6%
Regular	61.5%	61.5%	61.5%	92.6%	80.6%	86.2%
Weighted Avg.	66.1%	66.1%	66.1%	88.1%	87.5%	87.4%

We also performed an experiment where a classifier trained on the Extended-Brennan-Greenstadt dataset was tested on Hemingway-Faulkner Imitation corpus. Only 57.1% of the imitated documents were considered as imitation in that case. The Hemingway-Faulkner Imitation corpus is different from our dataset in several ways. The participants in the training set imitated Cormac McCarthy using one pre-specified excerpt from 'The Road' in writing about their day. But in the imitation contests, participants imitated two different authors without any topic constraint. Also the contest winners were found to be more successful than the mechanical turkers in imitating, as shown in Table VIII. To see how often a classifier can be fooled into predicting imitated document as written by the original authors, we trained an SMO SVM classifier with the Writeprints features using the original writing excerpts of Cormac McCarthy, Ernest Hemingway, William Faulkner and tested the classifier with the imitated documents. In this test, we classified imitated documents into three classes: Cormac McCarthy, Ernest Hemingway, William Faulkner. The result shows that the contest winners can imitate Ernest Hemingway in 84.27% cases, and William Faulkner in 66.67%, whereas the turkers were successful in 47.05% cases in imitating Cormac McCarthy.

C. Detecting long term deception

Detecting long term deception is similar to detecting fiction as deception. Fiction and elaborate

Table VIII: This table shows the success rate (Precision) of participants in imitating different authors. Imitation contest winners were more successful in imitating than the AMT participants.

Author name	Imitation success rate	Writer's skill
Cormac McCarthy	47.05%	Not professional writers
Ernest Hemingway	84.21%	Writing contest winners
William Faulkner	66.67%	Writing contest winners

deception have different linguistic characteristics than short-term on-the-spur deception, as in the long-term deception the author has sufficient time and topic to write descriptively and edit sufficiently to make it appear as a truthful document. This is why a different approach is required to detect long-term hoaxes and deception. Regular authorship recognition can be helpful to find inconsistencies in writing and to discover real authorship of the deceptive documents.

To test our method on long-term deception, we used the Thomas-Amina Hoax corpus. We performed an authorship attribution test on the posts he created as himself and as Amina in the alternate-history Yahoo! group. We found that he consistently maintained a separate writing style as Amina in the Yahoo! group as none of the Amina's posts were attributed to Thomas in the authorship attribution test. Thomas's writing style as Amina was different than his regular writing style. The use of upper-case letters and modal verbs [11] were noticeable in Amina's posts, whereas Thomas used longer sentences and more adjective and adverbs. Table IX lists the top Writeprints features that discriminate Thomas and Amina's posts in the Yahoo! group.

Moreover, all of the posts Thomas wrote as Amina and as himself and posts of Britta were considered as regular when tested on an SVM classifier which was trained with the Extended-Brennan-Greenstadt corpus. Deception classification of the posts from "A Gay Girl in Damascus" also did not show any indication of masking. In our test, only 14% of the blog posts were considered as deceptive which is less than the error rate, suggesting a random effect. 13 blog posts were classified as obfuscated documents, 22 were classified as imitated document. Table X shows that 57.14% of the deceptive documents were attributed to Amina during authorship attribution.

[11]Modal verbs are verbs that do not take an -s ending in the third person singular present, i.e. can, could, may, might, ought.

Table IX: Features that discriminate Thomas and Amina.

Feature	Information Gain Ratio
Percent upper case	0.758
t	0.620
than	0.620
all	0.516
only	0.516
tha	0.501
though	0.501
Frequency of *	0.432
less	0.433
can	0.433
not	0.433
a	0.423
that	0.423
Modal verb	0.423
that	0.423
them	0.373
even	0.372
doing	0.361

Table X: Blog post classification.

	Thomas	Amina
Imitation	11	11
Obfuscation	4	9

But maintaining an alternate writing style consistently for a long time is hard, which was evident in the Thomas-Amina case. When people started questioning Amina's existence, Thomas and his wife Britta were suspected as possible writers of the blog based on various pieces of evidence, for example, Thomas's address was used in Amina's account, and photos from Britta's picasa album were used in Amina's blog. In the end, Thomas admitted that he was "Amina." Regular authorship recognition also supports this fact. More than half of the blog posts (54.03%) were attributed to

Thomas during authorship attribution with an SVM classifier and the Writeprints feature set. Only 10 posts were attributed to Britta and the rest were attributed to "Amina." Native language detection, age and other demographics analysis are other possible ways to detect this form of deception, which are not explored in this paper.

VII. Discussion

In this research, we showed that two kinds of adversarial attacks—imitation and obfuscation—can be detected with high accuracy using a large feature set. However, the results of cross-validation on the Extended-Brennan-Greenstadt corpus are considerably better than the results on the imitation contest participants (96.6% vs. 87.4%).

There are several potential explanations for this discrepancy. The first is that the original dataset only handles a single imitated author, Cormac McCarthy, and the imitation classifier may be overfit to detect imitations of his style. This suggests that training on a combination of different author imitations will produce a stronger general classifier and that an imitation classifier should be trained on the target of the attack, if known.

The second potential explanation is the skill and incentive differential between the AMT participants and the contest winners. We included all reasonable attempts by the mechanical turk participants, but only the top three contestants per year were included in the contest data set. We offered our participants $5. A Faulker contest winner receives airfare for two to Memphis on United Airlines, five days free car rental, two free passes to the Faulkner Conference, and a complimentary room for five days and nights at an Oxford motel. First prize in the Heminway contest was round-trip tickets and dinner for two at Harry's in Florence, Italy.

A third potential explanation is that part of the success on the original data set is due to topic detection, not deception detection. In the Extended-Brennan-Greenstadt corpus, regular documents are from a variety of documents, obfuscation documents are a description of a neighborhood, and imitation documents provide a third-person narrative of a day.

Our analysis of the feature set shows that the non-content specific features have the same accuracy as of the content-specific features (Figure 1). Also, as most top ranked discriminating features are function words, even by ignoring contextual similarity of the documents, it is possible to detect adversarial documents with sufficient accuracy.

While it is true that content features may indicate authorship or deception, we do not believe this is the case here. Our non-deceptive writing samples consist of multiple documents per author, yet our authorship recognition techniques identify them properly with high levels of accuracy. The different content features there did not dissuade the standard authorship recognition techniques and we do not believe they greatly alter the outcome of the deception analysis. Furthermore, previous linguistic research has shown that the frequencies of common function words are content neutral and indicative of personal writing style [15].

What the "Gay Girl in Damascus" results show is that obfuscation is difficult to maintain in the long term. While Tom's posts as Amina were not found to be deceptive by our classifier, we show that traditional authorship attribution techniques work in this case.

Implications for Future Analyses: The current state of the art seems to provide a perfect balance between privacy and security. Authors who are deceptive in their writing style are difficult to identify, however their deception itself is often detectable. Further, the detection approach works best in cases where the author is trying fraudulently present themselves as another author.

However, while we are currently unable to unmask the original author of short term deceptions, further analyses might be able to do so, especially once a deception classifier is able to partition the sets. On the other hand, the Extended-Brennan-Greenstadt data set used contains basic attacks by individuals relying solely on intuition (they have no formal training or background in authorship attribution) and the results on the more skilled contest winners are less extensive.

We are currently working on a software appli-

cation to facilitate stylometry experiments and aid users in hiding their writing style. The software will point out features that are identifying to users and thus provide a mechanism for performing adaptive countermeasures against stylometry. This tool may be useful for those who need longer term anonymity or authors who need to maintain a consistent narrative voice. Even though Thomas MacMaster proved extremely skilled in hiding his writing style, half his posts were still identifiable as him rather than the fictional Amina.

In addition, these adaptive attacks may be able to hide the features that indicate deception, especially those in our top 20. It is also possible that attempts to change these features will result in changes that are still indicative of deception, particularly in the case of imitations. The fact is that most people do not have enough command of language to convincingly imitate the great masters of literature and the differences in style should be detectable using an appropriate feature set. A broader set of imitated authors is needed to determine which authors are easier or harder to imitate. Despite our ability to communicate, language is learned on an individual basis resulting in an individual writing style [11].

Implications for Adversarial Learning: Machine learning is often used in security problems from spam detection, to intrusion detection, to malware analysis. In these situations, the adversarial nature of the problem means that the adversary can often manipulate the classifier to produce lower quality or sometimes entirely ineffective results. In the case of adversarial writing, we show that using a broader feature set causes the manipulation itself to be detectable. This approach may be useful in other areas of adversarial learning to increase accuracy by screening out adversarial inputs.

VIII. Conclusion

Stylometry is necessary to determine authenticity of a document to prevent deception, hoaxes and frauds. In this work, we show that manual countermeasures against stylometry can be detected using second-order effects. That is, while it may be impossible to detect the author of a document whose authorship has been obfuscated, the obfuscation itself is detectable using a large feature set that is content-independent. Using Information Gain Ratio, we show that the most effective features for detecting deceptive writing are function words. We analyze a long-term deception and show that regular authorship recognition is more effective than deception detection to find indication of stylistic deception in this case.

IX. Acknowledgement

We are thankful to the Intel Science and Technology Center (ISTC) for Secure Computing and DARPA (grant N10AP20014) for supporting this work. We also thank Ahmed Abbasi for clarifying the Writeprints feature set and our colleagues Aylin Caliskan, Andrew McDonald, Ariel Stolerman and anonymous reviewers for their helpful comments on the paper.

References

[1] Ahmed Abbasi and Hsinchun Chen. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Trans. Inf. Syst.*, 26(2):1–29, 2008.

[2] M. Brennan and R. Greenstadt. Practical attacks against authorship recognition techniques. In *Proceedings of the Twenty-First Conference on Innovative Applications of Artificial Intelligence (IAAI), Pasadena, CA*, 2009.

[3] J. Burgoon, J. Blair, T. Qin, and J. Nunamaker. Detecting deception through linguistic analysis. *Intelligence and Security Informatics*, pages 958–958, 2010.

[4] K Calix, M Connors, D Levy, H Manzar, G MCabe, and S Westcott. Stylometry for e-mail author identification and authentication. *Proceedings of CSIS Research Day, Pace University*, 2008.

[5] Carole E. Chaski. Who's at the keyboard: Authorship attribution in digital evidence investigations. In *8th Biennial Conference on Forensic Linguistics/Language and Law*, 2005.

[6] M.G. Frank, M.A. Menasco, and M. O'Sullivan. Human Behavior and Deception Detection.

[7] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[8] J.T. Hancock, L.E. Curry, S. Goorha, and M. Woodworth. On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, 45(1):1–23, 2008.

[9] D.I. Holmes and R.S. Forsyth. The federalist revisited: New directions in authorship attribution. *Literary and Linguistic Computing*, 10:111–127, 1995.

[10] P. Juola and D. Vescovi. Empirical evaluation of authorship obfuscation using JGAAP. In *Proceedings of the 3rd ACM workshop on Artificial Intelligence and Security*, pages 14–18. ACM, 2010.

[11] Patrick Juola. Authorship attribution. *Foundations and Trends in information Retrieval*, 1(3):233–334, 2008.

[12] Gary Kacmarcik and Michael Gamon. Obfuscating document stylometry to preserve author anonymity. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 444–451, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

[13] Moshe Koppel, Jonathan Schler, Shlomo Argamon, and Eran Messeri. Authorship attribution with thousands of candidate authors. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 659–660, New York, NY, USA, 2006. ACM.

[14] M.B. Malyutov. Information transfer and combinatories. *Lecture Notes in Computer Science*, 4123(3), 2006.

[15] F. Mosteller and D. Wallace. Inference and disputed authorship: The federalist. 1964.

[16] Mihir Nanavati, Nathan Taylor, William Aiello, and Andrew Warfield. Herbert westdeanonymizer. In *6th Usenix Workshop on Hot Topics in Security (HotSec)*, 2011.

[17] A. Narayanan, H. Paskov, N. Gong, J. Bethencourt, E. Stefanov, R. Shin, and D. Song. On the feasibility of internet-scale author identification. In *Proceedings of the 33rd conference on IEEE Symposium on Security and Privacy*. IEEE, 2012.

[18] Michael P. Oakes. Ant colony optimisation for stylometry: The federalist papers. *Proceedings of the 5th International Conference on Recent Advances in Soft Computing*, pages 86–91, 2004.

[19] J.W. Pennebaker, R.J. Booth, and M.E. Francis. Linguistic inquiry and word count (LIWC2007). *Austin, TX: LIWC (www. liwc. net)*, 2007.

[20] J.C. Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. *Advances in Kernel MethodsSupport Vector Learning*, 208(MSR-TR-98-14):1–21, 1998.

[21] J.R. Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[22] J.R. Quinlan. *C4. 5: programs for machine learning*. Morgan kaufmann, 1993.

[23] Josyula R. Rao and Pankaj Rohatgi. Can pseudonymity really guarantee privacy? In *SSYM'00: Proceedings of the 9th conference on USENIX Security Symposium*, Berkeley, CA, USA, 2000. USENIX Association.

[24] B. Santorini. Part-of-speech tagging guidelines for the Penn Treebank Project (3rd revision). 1990.

[25] Harold Somers and Fiona Tweedie. Authorship attribution and pastiche. *Computers and the Humanities*, 37:407–429, 2003.

[26] M. Steller and JC Yuille. Recent developments in statement analysis. *Credibility assessment*, pages 135–154, 1989.

[27] Fiona J. Tweedie, S. Singh, and D.I. Holmes. Neural network applications in stylometry: The federalist papers. *Computers and the Humanities*, 30(1):1–10, 1996.

[28] Üzlem Uzuner and Boris Katz. A comparative study of language models for book and author recognition. In *IJCNLP*, page 969, 2005.

[29] James Wayman, Nicholas Orlans, Qian Hu, Fred Goodman, Azar Ulrich, and Valorie Valencia. Technology assessment for the state of the art biometrics excellence roadmap. http://www.biometriccoe.gov/SABER/index.htm, March 2009.

[30] R. Zheng, J. Li, H. Chen, and Z. Huang. A framework of authorship identification for online messages: Writing style features and classification techniques. *Journal American Society for Information Science and Technology*, 57(3):378–393, 2006.