Towards a Human-AI Hybrid for Adversarial Authorship

Jordan Allred
Department of Computer Science and Software Engineering
Auburn University
Auburn, USA
jordan.allred@auburn.edu

Sadaira Packer
Department of Computer Science and Software Engineering
Auburn University
Auburn, USA
smp0043@auburn.edu

Gerry Dozier
Department of Computer Science and Software Engineering
Auburn University
Auburn, USA
gvdozier@auburn.edu

Sarp Aykent
Department of Computer Science and Software Engineering
Auburn University
Auburn, USA
sza0112@auburn.edu

Alexicia Richardson
Department of Computer Science and Software Engineering
Auburn University
Auburn, USA
adr0021@auburn.edu

Michael C. King
School of Computing
Florida Institute of Technology
Melbourne, FL, USA
michaelking@fit.edu

Abstract—**In this paper, we compare two types of masking methods for Adversarial Authorship. One method is a human-based interactive form of masking (referred to as AuthorCAAT-V) while the second method is hybrid of three state-of-the-art author masking techniques (referred to as AIM-IT). Our results show that the performances of AuthorCAAT-V and AIM-IT are equal to or better than the performances of the three state-of-the-art author masking techniques in reducing the identification rate of four well-known authorship attribution systems (AASs). Furthermore, our results show that the hybridization of AuthorCAAT-V and AIM-IT provides a greater reduction in the identification rate.**

I. Introduction

According to Shou [1], one of the imminent threats to the anonymity of our cyber identities is that it is difficult for current AI systems to 'forget' our digital exhaust – the data that has been collected from us. Even if one could develop methods that would allow AI systems to forget, there may still be some who would refuse to eradicate the digital exhaust of others. This is an ever-growing problem with respect to Internet users and their privacy. AI systems for Authorship Attribution [2][3] are now becoming ever more sophisticated and efficient in identifying individuals based on their writing style.

One method that can be used to preserve the privacy of internet user, with respect to their writing style, is known as of Adversarial Authorship [4][5]. Some forms of Adversarial Authorship include Adversarial Stylometry [6][7][8] and Author Obfuscation / Masking [9][10][11][12][13][14][15]. For each of these methods, the objective is to mask the true identity of an author. In this paper, we compare a number of methods for Adversarial Authorship in an effort to determine their effectiveness in developing adversarial texts in an effort to conceal the identity of an author against a number of well-known authorship attribution systems (AASs) [6][16][17][18]. In this paper, we also present a human-AI hybrid for Adversarial Authorship that outperforms a number of state-of-the-art author masking techniques (AMTs). Hybridizing human intelligence with artificial intelligence has the potential to provide better performance than just relying on human intelligence or artificial intelligence alone [19]. Such collaborative systems allow for the strengths of both humans and AI to be utilized in a complementary way. AI can provide a quick analysis of large amounts of data, while a human's intuition is invaluable when it comes to decision making.

The remainder of this paper is as follows. In Section II, we present work related to the research that is presented in this paper. In Section III, we introduce two Adversarial Authorship methods: one is automated and the other is interactive. In Section IV, we present our experimental setup and introduce the datasets that will be used. In Section V, we present our results, and in Section VI, we provide our conclusions and future work.

II. Related Work

A. Author Masking Techniques (AMTs)

The first approaches to author masking as a part of the PAN shared tasks by Webis were established by Keswani et al. [9], Mansoorizadeh et al. [10], and Mihaylova et al. [11] in 2016. The technique developed by Keswani et al. [9] proposes a round-trip translation method to attempt to mask the text. Since there are competition restraints restricting the use of online translators, the group used the Moses SMT toolkit [20]. The language round-trip used for author masking by Keswani et al. [9] is as follows: English-to-German, German-to-French, and French-to-English. Mansoorizadeh et al. [10] implemented a technique to modify word frequency in the text. The way they did so was to replace specific instances of the two hundred most frequently used words with their synonyms. These synonyms were obtained from Princeton's Wordnet [21] with the constraint that this method would replace a maximum of one word per sentence.

Mihaylova et al. [11] developed a technique that targets many different style indicators typically used in author

978-1-7281-6861-6/20/$31.00 ©2020 IEEE

identification. The authors present three main categories: text transformations, noise, and general transformations. Text transformation consists of methods such as adding or removing punctuation, splitting or merging sentences, etc. Noise consists of replacing American English words with their British English counterparts and vice versa as well as adding or removing function words at beginnings of sentences. The general transformation consists of techniques like contraction replacement, replacing possessive phrases using regular expressions, etc. All of these methods are used to push the features of a given text toward the average of a specified training corpus.

The work of Bakhteev and Khazov [12] continued the shared task in 2017, and Castro et al. [13] furthered the state of the art. Bakhteev and Khazov [12] develop a technique that uses two methods to alter the text: an encoder-decoder and a set of sentence transformation rules. The encoder-decoder either uses nearest-neighbor synonym replacement or a pre-trained long short-term memory recurrent neural network to modify a sentence. The sentence transformation rules used were: contraction replacement, either sentence splitting or sentence merging, and either adding or removing introductory phrases.

The technique described by Castro et al. [13] uses a simple method for masking the original text in an attempt to shorten it. This is primarily done through contraction replacement, synonym substitution, and sentence simplification.

The most recently developed methods for author masking belong to Rahgouy et al. [14] and Kocher and Savoy [15]. The technique crafted by Rahgouy et al. [14] is a method similar to the one proposed by Mihaylova et al. [11]. Rahgouy et al. [14] used word replacement, phrase replacement, contraction replacement, and either sentence splitting or merging, in order to transform their text samples.

Kocher and Savoy [15] devised a rule-based approach for masking original texts using a set of twenty rules (e.g., changing contractions as seen in other methods). This rule-based approach seeks to change the original text enough to mask the true author while keeping the changes inconspicuous.

Anonymouth [6] is an authorship anonymization tool that aids users in making changes to a prewritten document to hide its true author via Adversarial Stylometry [22]. The most recent version of Anonymouth works by extracting features from documents and using them to determine what the feature targets should be. It utilizes the authorship attribution program, JStylo [6][23], for extracting the features and classifying the document. Anonymouth generates suggestions for potential changes to the document in order to move away from the writing style of the author. These suggestions come in the form of two lists of words for the user to add and/or delete from their document. Anonymouth also provides the user with a list of sentences that were created by two-way translation (otherwise known as iterative language translation) [7][24]. This list is sorted based on how helpful the sentence would be in reaching the feature targets. Users can apply changes to the document and then reclassify it until they reach the feature target. This version of Anonymouth has not been tested.

In the previous version of Anonymouth, the user would be presented with the extracted features (unlike the current version). Along with the extracted features, Anonymouth provides the user with suggestions as to which features should be changed [6]. Users found it difficult to alter the frequencies of the extracted features. The users were ultimately able to implement the suggestions offered using the Basic-9 AAS and were successful in deceiving the Basic-9 AAS. However, they were unsuccessful in deceiving the Writeprints (Limited) AAS due to it having a more extensive set of stylometric features. Using Writeprints (Limited), all of the resultant adversarial texts were correctly identified.

B. *AuthorCAAT*

AuthorCAAT (Author Cyber Analysis & Advisement Tool) assists users in anonymizing a document via Adversarial Authorship [25][4][5][26][27]. The previous published version of AuthorCAAT, AuthorCAAT-III [26], included a character unigram feature extractor, an author identification system based on the nearest neighbor matching method, AuthorWebs, and an interactive evolutionary hill-climbing process [25]. An AuthorWeb [27] consists of a set of author clusters. Each author cluster has arcs directed to and from it. The arcs represent writing samples and are directed toward the author cluster to which they belong.

The process for using AuthorCAAT-III starts with the user entering a sample, referred to as the parent text, and then selecting an author cluster to either mimic or avoid. Next, the user does either iterative language translation (ILT), iterative paraphrasing, or steepest ascent ILT hill-climbing to create an offspring text. The user then decides whether the offspring text is better than the parent text. If so, the offspring text becomes the parent text. The user can edit the parent and/or continue the process of creating offspring using the three operations mentioned previously until an adversarial text is generated that is satisfactory to the user. Gaston et al. developed AuthorCAAT-IV and AuthorCAAT-V by incorporating LIWC, Sentiment Analysis, Topic Models, GEFeS, and LSVM [4][5][27]. Figure 1 provides a view of the AuthorCAAT-IV user interface. AuthorCAAT-V will be discussed in more detail in the upcoming section.

C. *Authorship Attribution Systems (AASs)*

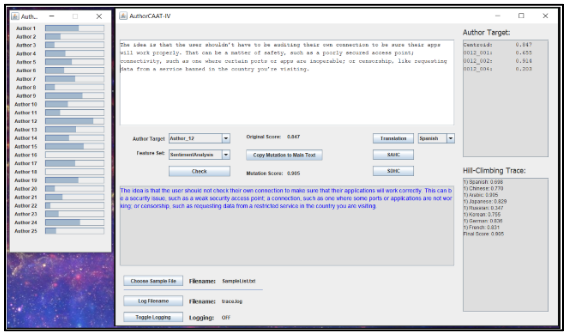
The work in AASs can also mainly be found in works of the PAN shared tasks. The AASs used to evaluate the texts were taken from the best systems as found by Neal et al. [2]. Koppel11 is an AAS that only uses 4-grams to create an author profile. The author profile is created for each candidate author using a limit of the 20,000 most common features. Next, a random subset of the particular unknown text is compared against the author profiles using cosine similarity. The process is repeated 100 times, and the candidate author with the most votes is chosen as the predicted author of the unknown text. Due to the random nature of the pieces of text chosen in Koppel11, the results are not deterministic. To counter this, we run the system five times and use majority voting to determine an author. This value was determined to be the lowest value to give a deterministic author prediction through our evaluation.

Teahan03 also uses n-grams, but this method is using unigrams to create an author profile for each candidate author. Every unknown text is then compared against the author profiles, this time using cross-entropy. Keselj03 uses the training set to create author profiles for candidate authors and a validation set to tune two parameters – n-grams and vocabulary size. The n-grams are chosen from 3, 4, 5, 6 and the vocabulary size is chosen from 500, 1000, 2000, 3000, 5000. The test set is then evaluated using a dissimilarity function to compare the given text features to the average features of each candidate author.

III. AuthorCAAT-V and AIM-IT

A. AuthorCAAT-V

The process for creating adversarial text with AuthorCAAT-V begins with the user entering their parent text that they wish to anonymize. The user picks an author target from a list of 25 that they would like to move towards or away from (write more or less like) and they pick a feature set that they would like to focus on. The feature sets available are character unigrams, sentiment analysis, linguistic inquiry and word count (LIWC), topic model, bag of words, and stylometry. AuthorCAAT-V uses multiple feature sets because changing a document based on one feature set may not provide anonymity when an authorship attribution system classifies using a different feature set than the one that was used to make changes to the document [5].



```
The idea is that the user shouldn't have to be auditing their own connection to be sure their apps
will work properly. That can be a matter of safety, such as a poorly secured access point;
connectivity, such as one where certain ports or apps are inoperable; or censorship, like requesting
data from a service banned in the country you're visiting.
```

Author Target: Author_12

Feature Set: SentimentAnalysis

Original Score: 0.847

Mutation Score: 0.905

Spanish

The idea is that the user should not check their own connection to make sure that their applications will work correctly. This can be a security issue, such as a weak security access point; a connection, such as one where some ports or applications are not working; or censorship, such as requesting data from a restricted service in the country you are visiting.

Filename: SampleList.txt

Filename: trace.log

Logging: OFF

Author Target:

```
Centroid:     0.847
0012_001:     0.655
0012_002:     0.914
0012_004:     0.203
```

Hill-Climbing Trace:

```
1) Spanish: 0.698
1) Chinese: 0.770
1) Arabic: 0.905
1) Japanese: 0.829
1) Russian: 0.347
1) Korean: 0.755
1) German: 0.836
1) French: 0.831
Final Score: 0.905
```

Fig. 1. The AuthorCAAT-IV User Interface

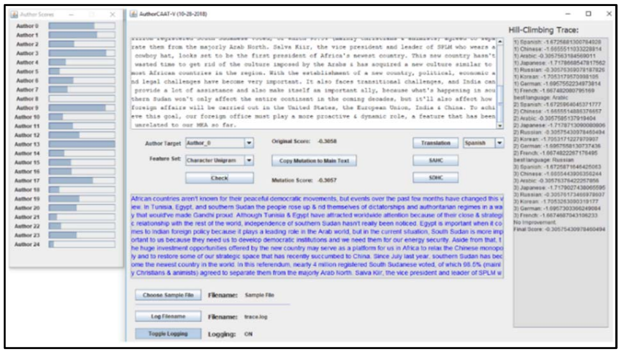
For the topic model feature set we use the MALLET program to extract a set number of topics that are specific to the group of documents being analyzed [28]. Topics are groupings of words that have been determined to be related because they commonly appear together in a set of documents. MALLET gathers the topics by analyzing the document using the topic modeling algorithm latent Dirichlet allocation (LDA) and Gibbs sampling. The program returns the words with their associated topics and the distribution of the topics in each document.

LIWC is a program that extracts information from text that provides an understanding of the psychological state of an author [29]. The words from the text are taken and compared to LIWC's dictionary of 6400 words. These 6400 words belong to different hierarchical categories. Each word from the sample text that is found in the dictionary counts towards the category associated with the word from the dictionary.

OpinionFinder is the program used for the sentiment analysis feature set [30][31][32][33]. It allows users to observe the polarity and subjectivity of the text. OpinionFinder examines the words of the documents and compares them to its dictionary. The words in the dictionary have an assigned polarity and subjectivity.

AuthorCAAT-V uses GEFeS (Genetic & Evolutionary Feature Selection) for off-line feature selection. GEFeS allows for the evaluation of various subsets of a feature set in order to determine the best features for a specific dataset [34][35][36]. AuthorCAAT-V also utilizes a linear support vector machine because it has shown to perform well in [4] and [5] when used with multiple feature sets.

Once an author target or feature set has been selected, a visual representation of the fitness scores for all 25 authors is displayed. The numerical value for the selected author target is displayed under the parent text. The higher the score is the closer the text is to the author target. These values change depending on the selected feature set. The user then mutates their text using two-way translations or iterative language translation hill-climbing (steepest ascent or descent). The languages used for translating are Spanish, Chinese, Arabic, Japanese, Russian, Korean, German, and French. After running one of these methods a mutated text is produced along with a fitness score for the mutation. Like the score for the parent text, this score is relative to the selected author target. The user can then decide if the mutated text is good enough to become the parent text or if they would like to mutate the parent text again. Once a satisfactory mutated text has been created, the user can copy this text to the main text box where they can make modifications and repeat the process. Figure 2 shows the AuthorCAAT-V user interface.



```
rate them from the majorly Arab North. Salva Kiir, the vice president and leader of SPLM who wears a
 cowboy hat, looks set to be the first president of Africa's newest country. This new country hasn't
 wasted time to get rid of the culture imposed by the Arabs & has acquired a new culture similar to
most African countries in the region. With the establishment of a new country, political, economic a
nd legal challenges have become very important. It also faces transitional challenges, and India can
 provide a lot of assistance and also make itself an important ally, because what's happening in sou
thern Sudan won't only affect the entire continent in the coming decades, but it'll also affect how
foreign affairs will be carried out in the United States, the European Union, India & China. To achi
eve this goal, our foreign office must play a more proactive & dynamic role, a feature that has been
 unrelated to our MEA so far.
```

Author Target: Author_0

Feature Set: Character Unigram

Original Score: -0.3058

Mutation Score: -0.3057

Spanish

African countries aren't known for their peaceful democratic movements, but events over the past few months have changed this view. In Tunisia, Egypt, and southern Sudan the people rose up & rid themselves of dictatorships and authoritarian regimes in a way that would've made Gandhi proud. Although Tunisia & Egypt have attracted worldwide attention because of their close & strategic relationship with the rest of the world, independence of southern Sudan hasn't really been noticed. Egypt is important when it comes to Indian foreign policy because it plays a leading role in the Arab world, but in the current situation, South Sudan is more important to us because they need us to develop democratic institutions and we need them for our energy security. Aside from that, the huge investment opportunities offered by the new country may serve as a platform for us in Africa to relax the Chinese monopoly and to restore some of our strategic space that has recently succumbed to China. Since July last year, southern Sudan has become the newest country in the world. In this referendum, nearly 4 million registered South Sudanese voted, of which 98.5% (mainly Christians & animists) agreed to separate them from the majorly Arab North. Salva Kiir, the vice president and leader of SPLM w

Filename: Sample File

Filename: trace.log

Logging: ON

Hill-Climbing Trace:

```
1) Spanish: -1.6725881300784928
1) Chinese: -1.6655511033228814
1) Arabic: -0.3057563184569011
1) Japanese: -1.7178668547817562
1) Russian: -0.3057639078187826
1) Korean: -1.7053179570998105
1) German: -1.6957552234973814
1) French: -1.667482080795169
best language: Arabic
2) Spanish: -1.6725964045371777
2) Chinese: -1.6655514886376657
2) Arabic: -0.3057585137919404
2) Japanese: -1.7178713090080806
2) Russian: -0.30575430978460494
2) Korean: -1.7053171227970907
2) German: -1.6957558130737436
2) French: -1.6674822267176495
best language: Russian
3) Spanish: -1.6725871646425063
3) Chinese: -1.6655443906356244
3) Arabic: -0.30576376422257856
3) Japanese: -1.7179027438065595
3) Russian: -0.30576173466978607
3) Korean: -1.7053263090319177
3) German: -1.6957303306249084
3) French: -1.6674687043106233
No Improvement
Final Score: -0.30575430978460494
```

Fig. 2. The AuthorCAAT-V User Interface

B. AIM-IT

The Automated Intelligent Masking & Information Tool (AIM-IT) is a novel AMT that is an instance of a (1+3) evolution strategy [37]. AIM-IT consist of five steps as shown in Figure 3. In Step 0, the initial parent text is evaluated using the fitness function in Figure 3. In Step 1, the text with the highest fitness is evaluated, and if it is incorrectly classified, the procedure finishes. In Step 2, three offspring texts (children) of the parent are created using the AMTs by Castro et al. [13], Mihaylova et al. [11] and Rahgouy et al. [14]. In Step 3, the offspring are evaluated and the best performing child is selected. In Step 4, if the selected child is better than the parent then it replaces the parent; otherwise, the process has stalled and the procedure finishes. In Step 5, the procedure is terminated and the text with the highest fitness is returned.

The fitness function shown in Figure 3 is used to evaluate the fitness of a text given a dataset (minus the parent writing sample, *D'*), parent text, and a target vector. The fitness function is implemented as follows. Initially, the parent text is removed from the dataset resulting in *D'*. Next, features are extracted from the candidate adversarial text in the form of character unigrams and a bag of words in an effort to develop a feature vector. The feature vector is then provided as input to a linear support vector machine (LSVM). The LSVM then returns a decision function and this decision function is then scaled from zero to one inclusive and subtracted from the target vector, *t*. The target vector used consists of values of 1.0, t_k, for all authors that are not the true author of the text and the value of 0.0 for, t_j, associated with the true author. The only constraint for these two values is all values of t_k are greater than t_j. Finally, the sum of the absolute value of each element in the resulting vector is taken, and this sum is returned. Figure 4 provides an example how the AIM-IT target vector is used.

```
Procedure AIM-IT(D', parent, t)
{
        Step 0: $fitness_{parent}\ (D', parent, t)$
        Step 1: create mutants of parent
                $child_0 = mutate(parent, Castro)$
                $child_1 = mutate(parent, Mihaylova)$
                $child_2 = mutate(parent, Rahgouy)$
        Step 2: evaluate offspring and select best child
                $\underset{i}{\operatorname{argmax}}\ (fitness_i\ (D', child_i, t))$
        Step 3: if $fitness_i > fitness_{parent}$ then $parent = child_i$
        Step 4: if $\forall_k\ (t_j > t_k)$ then go to Step 1   // where j is the index of the current author
}

Function $fitness_x$ (D', x, t)
{
        return $|LSVM(x, D') - t|$
        where $\forall_k\ (t_j > t_k)$                           // where j is the index of the current author
}
```

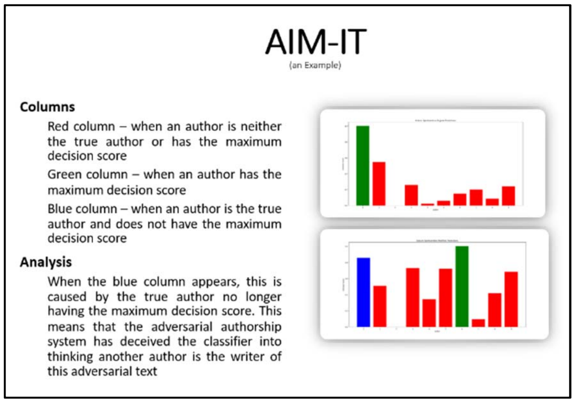
Fig. 3. Pseudocode of the AIM-IT method

IV. Experiment

In our experiment, we compare the masking performances of the five AMTs presented earlier on their ability to mask the first 25 instances of the CASIS-1000 dataset [38][2]. The masking performance of a method is simply the change in accuracy, Δ_{Acc} [37], given by the original text when classified by the following authorship attribution systems:

1. Keselj-2003 [16],
2. Teahan-2003 [17],
3. Koppel-2011 [39],
4. CNN [40].

The CASIS-1000 dataset is a collection of 1000 blogs from 1000 authors. Each of the blogs is divided into 4 writing samples for each authors for a total of 4000 writing samples. Each writing sample has on average 13 sentences [2]. Table I shows the baseline performance of the four AASs on the CASIS-25 dataset.



AIM-IT

(an Example)

Columns

Red column – when an author is neither the true author or has the maximum decision score

Green column – when an author has the maximum decision score

Blue column – when an author is the true author and does not have the maximum decision score

Analysis

When the blue column appears, this is caused by the true author no longer having the maximum decision score. This means that the adversarial authorship system has deceived the classifier into thinking another author is the writer of this adversarial text

Fig. 4. An Example of How the AIM-IT Target Vector is Used

TABLE I.
A Comparison of the AASs on the CASIS-25 Dataset

	Keselj2003	Teahan2003	Koppel2011	CNN
Original	0.60	0.92	0.76	0.84

V. Results

The results presented in this section were generated by applying the adversarial authorship techniques to the fourth writing instances of the first 25 authors of the CASIS-1000 dataset. The adversarial authorship techniques were 'blind' in that they had no prior interaction with author attribution or verification systems. After the AMTs 'evaded' classification of their internal author identification systems, the resulting adversarial texts were submitted to the set of author AASs systems.

Table II provides the results of our experiment. The first column contains the names associated with the authorship attribution systems and the change in accuracy, Δ_{ACC}, using only the writing samples of the first 25 authors of the CASIS-1000 dataset (referred to as CASIS-25). The entries within the first column of Table I corresponding to the authorship attribution systems are: Keselj2003 [16], Teahan2003 [17], and Koppel2011 [39].

The next five columns of Table II correspond to the five following AMTs presented earlier: Castro [13], Mihaylova [11], Rahgouy [14], AIM-IT, and AuthorCAAT-V [25]. Castro, Mihaylova, and Rahgouy were selected because of the top ten AMTs presented in [37], they were ranked 1st, 2nd, and 3rd respectively. The last column corresponds to the AuthorCAAT-V + AIM-IT hybrid. The hybrid adversarial text were formed by taking the adversarial texts developed by AuthorCAAT-V and running them through AIM-IT.

In Table II, with respect to the first five AMTs, one can see that AuthorCAAT-V had the best performance against three AAS, (in green), while Castro, Mihaylova, Rahgouy, and AIM-IT had the best performance against 0, 0, 1, and 2 AASs respectively. In Table II, one can see that the hybrid has the best performance against three of four the AASs. Notice also, that on the three for which the hybrid has the best performance, that the hybrid dramatically reduces the identification accuracy. In terms of the average reduction against the four well-known

AASs, the hybrid has the greatest average reduction (-32%), followed by AuthorCAAT-V (-14%), AIM-IT (-11%), Rahgouy (-10%), Mihaylova (-8%), and Castro (4%).

Tables III and IV show the detailed results of the classification of the adversarial texts when submitted to the three AASs (Table III) and the CNN (Table IV). The leftmost column of each table shows the name of each of the 25 adversarial texts followed by the change in accuracy for each AAS, the average change in accuracy of the three systems for each author masking technique, the accuracy for each attribution system with respect to the AMT, and the average accuracy for each attribution system. In the tables, 'x' means neither the original nor the adversarial texts were correctly classified, '✓' means both the original and the adversarial texts were correctly classified, 'TF' means the original text was correctly classified and the adversarial text was incorrectly classified, and 'FT' means the original text was incorrectly classified and the adversarial text was correctly classified.

In Tables III and IV, the last four rows represent the change in accuracy given an AAS or the CNN (denoted as Δ Acc.), the average change in accuracy (denoted as Δ Acc. Avg.), the reduced adversarial accuracy (denoted as RAA), and the average RAA (denoted as RAA Avg.). In Table III, one can see that had a total of 10 TFs, 47 ✓s, and 1 FT. This can be represented using the following notation, *<10, 47, 1>*. Similarly, the performances of AIM-IT, Mihaylova, Rahgouy, Castro, and the hybrid can be represented as *<8, 47, 1>*, *<6, 51, 0>*, *<7, 49, 1>*, *<4, 52, 1>*, and *<30, 27, 5>*. Given these results one can see that the hybrid dramatically outperforms the other AMTs in terms of evading detection by the AASs (with 30 TFs) while having only 27 checkmarks. However, this improvement in performance of the hybrid comes at the cost of an increased number of FTs.

In Table IV, one can see the performances of the five AMTs are as follows *<5, 16, 0>* for AuthorCAAT-V, *<3, 19, 0>* for AIM-IT, *<2, 19, 0>* for Mihaylova, *<2, 19, 0>* for Rahgouy, *<1, 20, 1>* for Castro, and *<7, 14, 0>* for the hybrid. Given these results one can see that the hybrid has the overall best performance; however, it narrowly outperforms AuthorCAAT-V.

VI. Conclusion & Future Work

In this paper, we compared five AMTs for adversarial authorship. All of the performances with respect to Δ_{ACC} were fairly close. Overall, the performance of AIM-IT and AuthorCAAT-V was equal to or better than the performances of three state-of-the-art AMTs. Furthermore, our results show that the hybridization of AuthorCAAT-V and AIM-IT provides a greater reduction in the identification rate against three of the four well-known AASs. Our future work will be devoted towards the development and analysis of other human-AI hybrids for adversarial authorship.

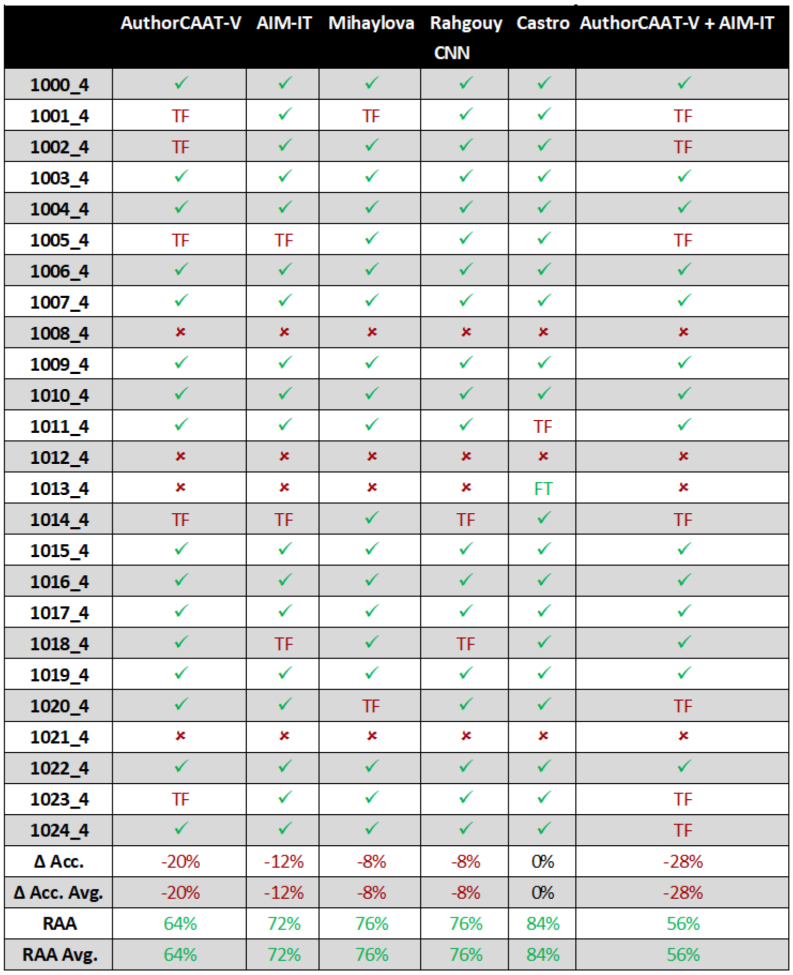
Table II
Results of the Reduction in Accuracy based on the Adversarial Texts

	Castro	Mihaylova	Rahgouy	AIM-IT	AuthorCAAT-V	AuthorCAAT-V + AIM-IT
Keselj2003	-0.12	-0.16	-0.16	**-0.20**	**-0.20**	-0.16
Teahan2003	0.00	-0.04	**-0.12**	**-0.12**	-0.08	-0.32
Koppel2011	0.00	-0.04	0.04	0.00	**-0.08**	-0.52
CNN	0.00	-0.08	-0.08	-0.12	**-0.20**	-0.28

TABLE III.
DETAILED AUTHORSHIP AT TRIBUTON RESULTS

	AuthorCAAT-V			AIM-IT			Mihaylova			Rahgouy			Castro			AuthorCAAT-V + AIM-IT		
	keselj03	teahan03	koppel11	keselj03	teahan03	koppel11	keselj03	teahan03	koppel11	keselj03	teahan03	koppel11	keselj03	teahan03	koppel11	keselj03	teahan03	koppel11
1000_4	✗	✓	✓	✗	✓	✓	✗	✓	✓	✗	✓	✓	✗	✓	✓	FT	✓	✓
1001_4	✗	TF	✗	✗	✓	✗	✗	✓	✗	✗	✓	✗	✗	✓	✗	✗	TF	✗
1002_4	TF	✓	TF	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	TF
1003_4	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	TF	TF	TF
1004_4	✓	✓	✓	TF	✓	✓	✓	✓	✓	TF	TF	✓	✓	✓	TF	TF	TF	TF
1005_4	TF	TF	TF	TF	✓	TF	TF	✓	✓	✓	✓	✓	TF	✓	✓	✓	✓	✓
1006_4	TF	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	TF
1007_4	TF	✓	✓	✓	✓	✓	TF	✓	✓	✓	✓	✓	✓	✓	✓	TF	✓	TF
1008_4	FT	✗	✓	✗	✗	✓	✗	✗	✓	✗	✗	✓	✗	✗	✓	✗	FT	TF
1009_4	✗	✓	✓	✗	✓	✓	✗	✓	✓	✗	✓	✓	✗	✓	✓	FT	✓	✓
1010_4	✗	✓	✗	✗	TF	FT	✗	✓	✗	✗	TF	✗	✗	✓	FT	✗	TF	✗
1011_4	✗	✓	✓	✗	✓	✓	✗	✓	✓	✗	✓	✓	✗	✓	✓	✗	✓	✓
1012_4	✓	✓	✗	✓	✓	✗	✓	✓	✗	✓	✓	FT	✓	✓	✗	✓	✓	✗
1013_4	TF	✓	✓	TF	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	TF	✓	TF
1014_4	✗	✓	✗	✗	TF	✗	✗	✓	✗	✗	TF	✗	✗	✓	✗	FT	✓	✗
1015_4	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	TF
1016_4	✓	✓	✓	✓	✓	✓	TF	✓	TF	TF	✓	✓	✓	✓	✓	TF	TF	TF
1017_4	✗	✓	✓	✗	TF	✓	✗	✓	✓	✗	✓	✓	✗	✓	✓	✗	TF	TF
1018_4	✓	✓	✓	✓	✓	✓	TF	✓	✓	TF	✓	✓	TF	✓	✓	TF	TF	TF
1019_4	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
1020_4	✗	✓	✓	✗	✓	✓	✗	✓	✓	✗	✓	✓	✗	✓	✓	FT	✓	✓
1021_4	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
1022_4	✓	✓	✓	✓	✓	✓	✓	TF	✓	✓	✓	✓	TF	✓	✓	TF	TF	TF
1023_4	✓	✓	✓	TF	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	TF	TF	TF
1024_4	TF	✓	✗	TF	✓	✗	✓	✓	✗	TF	✓	✗	✓	✓	✗	✓	✓	✗
Δ Acc.	-20%	-8%	-8%	-20%	-12%	0%	-16%	-4%	-4%	-16%	-12%	4%	-12%	0%	0%	-16%	-32%	-52%
Δ Acc. Avg.	-12%			-11%			-8%			-8%			-4%			-33%		
RAA	40%	84%	68%	40%	80%	76%	44%	88%	72%	44%	80%	80%	48%	92%	76%	44%	60%	24%
RAA Avg.	64%			65%			68%			68%			72%			43%		

TABLE IV
DETAILED CONVOLUTIONAL NEURAL NETWORK RESULTS



	AuthorCAAT-V	AIM-IT	Mihaylova	Rahgouy CNN	Castro	AuthorCAAT-V + AIM-IT
1000_4	✓	✓	✓	✓	✓	✓
1001_4	TF	✓	TF	✓	✓	TF
1002_4	TF	✓	✓	✓	✓	TF
1003_4	✓	✓	✓	✓	✓	✓
1004_4	✓	✓	✓	✓	✓	✓
1005_4	TF	TF	✓	✓	✓	TF
1006_4	✓	✓	✓	✓	✓	✓
1007_4	✓	✓	✓	✓	✓	✓
1008_4	×	×	×	×	×	×
1009_4	✓	✓	✓	✓	✓	✓
1010_4	✓	✓	✓	✓	✓	✓
1011_4	✓	✓	✓	✓	TF	✓
1012_4	×	×	×	×	×	×
1013_4	×	×	×	×	FT	×
1014_4	TF	TF	✓	TF	✓	TF
1015_4	✓	✓	✓	✓	✓	✓
1016_4	✓	✓	✓	✓	✓	✓
1017_4	✓	✓	✓	✓	✓	✓
1018_4	✓	TF	✓	TF	✓	✓
1019_4	✓	✓	✓	✓	✓	✓
1020_4	✓	✓	TF	✓	✓	TF
1021_4	×	×	×	×	×	×
1022_4	✓	✓	✓	✓	✓	✓
1023_4	TF	✓	✓	✓	✓	TF
1024_4	✓	✓	✓	✓	✓	TF
Δ Acc.	-20%	-12%	-8%	-8%	0%	-28%
Δ Acc. Avg.	-20%	-12%	-8%	-8%	0%	-28%
RAA	64%	72%	76%	76%	84%	56%
RAA Avg.	64%	72%	76%	76%	84%	56%

REFERENCES

[1] Darren Shou, “The Next Big Privacy Hurdle? Teaching AI to Forget,” *Wired*, 2019. [Online]. Available: https://www.wired.com/story/the-next-big-privacy-hurdle-teaching-ai-to-forget/. [Accessed: 05-Dec-2019].

[2] T. Neal, K. Sundararajan, A. Fatima, Y. Yan, Y. Xiang, and D. Woodard, “Surveying stylometry techniques and applications,” *ACM Computing Surveys*. 2017.

[3] S. Elmanarelbouanani and I. Kassou, “Authorship Analysis Studies: A Survey,” *Int. J. Comput. Appl.*, 2014.

[4] J. Gaston *et al.*, “Authorship Attribution via Evolutionary Hybridization of Sentiment Analysis, LIWC, and Topic Modeling Features,” in *Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence, SSCI 2018*, 2019.

[5] J. Gaston *et al.*, “Authorship Attribution vs. Adversarial Authorship from a LIWC and Sentiment Analysis Perspective,” in *Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence, SSCI 2018*, 2019.

[6] A. W. E. McDonald, S. Afroz, A. Caliskan, A. Stolerman, and R. Greenstadt, “Use fewer instances of the letter ‘i’: Toward writing style anonymization,” 2012.

[7] A. W. E. Mcdonald, J. Ulman, M. Barrowclift, and R. Greenstadt, “Anonymouth Revamped : Getting Closer to Stylometric Anonymity,” 2012.

[8] M. Brennan, S. Afroz, and R. Greenstadt, “Adversarial stylometry,” *ACM Trans. Inf. Syst. Secur.*, 2012.

[9] Y. Keswani, H. Trivedi, P. Mehta, and P. Majumder, “Author Masking through Translation Notebook for PAN at CLEF 2016,” pp. 1–5, 2016.

[10] M. Mansoorizadeh, T. Rahgooy, M. Aminiyan, and M. Eskandari, “Author obfuscation using WordNet and language models Notebook for PAN at CLEF 2016,” pp. 1–8, 2016.

[11] T. Mihaylova, G. Karadzhov, P. Nakov, Y. Kiprov, G. Georgiev, and I. Koychev, “SU@ PAN’2016: Author Obfuscation—Notebook for PAN at CLEF 2016,” *Conf. Labs Eval. Forum*, 2016.

[12] O. Bakhteev and A. Khazov, “Author masking using sequence-to-sequence models: Notebook for PAN at CLEF 2017,” *CEUR Workshop Proc.*, vol. 1866, 2017.

[13] D. Castro, R. Ortega, and R. Muñoz, “Author Masking by Sentence Transformation,” 2017.

[14] M. Rahgouy, H. B. Giglou, T. Rahgooy, H. Zeynali, S. Khayat, and M. Rasouli, “Author Masking Directed by Author ’ s Style,” pp. 1–6, 2018.

[15] M. Kocher and J. Savoy, “UniNE at CLEF 2018 : Author Masking,” 2018.

[16] V. Keselj, F. Peng, N. Cercone, and C. Thomas, “N-gram-

based Author Profiles for Authorship Attribution," 2003.

[17] D. V. Khmelev and W. J. Teahan, "A repetition based measure for verification of text collections and for text categorization," 2003.

[18] G. Kacmarcik and M. Gamon, "Obfuscating document stylometry to preserve author anonymity," 2010.

[19] M. H. Jarrahi, "Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making," *Bus. Horiz.*, vol. 61, no. 4, pp. 577–586, Jul. 2018.

[20] "Moses SMT Toolkit." [Online]. Available: http://www.statmt.org/moses/. [Accessed: 05-Dec-2019].

[21] G. A. Miller, "WordNet: a lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995.

[22] M. Brennan, S. Afroz, and R. Greenstadt, "Adversarial stylometry," *ACM Trans. Inf. Syst. Secur.*, vol. 15, no. 3, pp. 1–22, Nov. 2012.

[23] M. Brennan and R. Greenstadt, "Practical attacks against authorship recognition techniques," in *Proceedings of the 21st Innovative Applications of Artificial Intelligence Conference, IAAI-09*, 2009.

[24] N. Mack, J. Bowers, H. Williams, G. Dozier, and J. Shelton, "The Best Way to a Strong Defense is a Strong Offense: Mitigating Deanonymization Attacks via Iterative Language Translation," *Int. J. Mach. Learn. Comput.*, 2015.

[25] S. Day, J. Brown, Z. Thomas, I. Gregory, L. Bass, and G. Dozier, "Adversarial Authorship, AuthorWebs, and Entropy-Based Evolutionary Clustering," in *2016 25th International Conference on Computer Communication and Networks (ICCCN)*, 2016, pp. 1–6.

[26] C. Faust, G. Dozier, J. Xu, and M. C. King, "Adversarial authorship, interactive evolutionary hill-climbing, and author CAAT-III," in *2017 IEEE Symposium Series on Computational Intelligence, SSCI 2017 - Proceedings*, 2018.

[27] M. Narayanan *et al.*, "Adversarial Authorship, Sentiment Analysis, and the AuthorWeb Zoo," in *Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence, SSCI 2018*, 2019.

[28] A. McCallum, "MALLET: A Machine Learning for Language Toolkit," *Http://Mallet.Cs.Umass.Edu*, 2002.

[29] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods," *Journal of Language and Social Psychology*. 2010.

[30] J. Wiebe, T. Wilson, and C. Cardie, "Annotating expressions of opinions and emotions in language," *Language Resources and Evaluation*. 2005.

[31] E. Riloff, J. Wiebe, and T. Wilson, "Learning subjective nouns using extraction pattern bootstrapping," 2007.

[32] E. Riloff and J. Wiebe, "Learning extraction patterns for subjective expressions," 2007.

[33] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-Level sentiment analysis," in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP-2005)*, 2005.

[34] L. D. Davis and M. Mitchell, *Handbook of Genetic Algorithms*. 1991.

[35] G. Dozier *et al.*, "GEFeS: Genetic & evolutionary feature selection for periocular biometric recognition," in *IEEE SSCI 2011 - Symposium Series on Computational Intelligence - CIBIM 2011: 2011 IEEE Workshop on Computational Intelligence in Biometrics and Identity Management*, 2011.

[36] H. C. Williams, J. N. Carter, W. L. Campbell, K. Roy, and G. V. Dozier, "Genetic & Evolutionary Feature Selection for Author Identification of HTML Associated with Malware," *Int. J. Mach. Learn. Comput.*, 2014.

[37] M. Potthast, F. Schremmer, M. Hagen, and B. Stein, "Overview of the author obfuscation task at PAN 2018: A new approach to measuring safety," *CEUR Workshop Proc.*, vol. 2125, 2018.

[38] N. Mack, J. Bowers, H. Williams, G. Dozier, and J. Shelton, "The Best Way to a Strong Defense is a Strong Offense: Mitigating Deanonymization Attacks via Iterative Language Translation," *Int. J. Mach. Learn. Comput.*, vol. 5, no. 5, pp. 409–413, Sep. 2015.

[39] M. Koppel, J. Schler, and S. Argamon, "Authorship attribution in the wild," *Lang. Resour. Eval.*, vol. 45, no. 1, pp. 83–94, 2011.

[40] P. Shrestha, S. Sierra, F. A. González, P. Rosso, M. Montes-Y-Gómez, and T. Solorio, "Convolutional neural networks for authorship attribution of short texts," in *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference*, 2017.