Tsikerdekis, M., & Zeadally, S. (2014). Multiple Account Identity Deception Detection in Social Media Using Non-Verbal Behavior. IEEE Transactions on Information Forensics and Security, 9(8), 1311–1321.
http://doi.org/10.1109/TIFS.2014.2332820
http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6843931

Multiple Account Identity Deception Detection in Social Media Using Non-Verbal Behavior

Michail Tsikerdekis and Sherali Zeadally, *SMIEEE*

***Abstract*—Identity deception has become an increasingly important issue in the social media environment. The case of blocked users initiating new accounts, often called sockpuppetry, is widely known and past efforts, which have attempted to detect such users, have been primarily based on verbal behavior (e.g., using profile data or lexical features in text). Although these methods yield a high detection accuracy rate, they are computationally inefficient for the social media environment which often involves databases with large volumes of data. To date, little attention has been paid to detecting online deception using non-verbal behavior. We present a detection method based on non-verbal behavior for identity deception which can be applied to many types of social media. Using Wikipedia as an experimental case, we demonstrate that our proposed method results in high detection accuracy over previous methods proposed while being computationally efficient for the social media environment. We also demonstrate the potential of non-verbal behavior data that exists in social media and how designers and developers can leverage such non-verbal information in detecting deception to safeguard their online communities.**

***Index Terms*— Algorithm, deception, identity, performance, social media**

I. Introduction

IN the past decade we have experienced an increasing level of interest in online social media which enable users to not only create content but also exchange it using Web 2.0 technologies [1]. The number of users registering with social networking sites such as Facebook and Twitter keeps increasing at a rapid pace amounting to 82 percent of the world's online population [2]. Social network usage has increased by 64% since 2005 [3]. The ease with which we can generate online profiles at a low cost has also led to ample opportunities for identity deception which at times can have fatal consequences. A recent well-known example is the case of a mother pretending to be a teenage boy on the social networking site MySpace in order to obtain information from a teenage girl eventually leading to the girl committing suicide [4]. Other social media services such as collaborative projects have to engage in "cat-mouse" games by constantly having to block user accounts for individuals joining in with different account names not long after a block has been applied.

Solutions have been proposed that can assist in detecting multiple accounts owned by the same individual but their effectiveness vary in terms of computational efficiency and complexity of practical implementation depending on the availability of the appropriate data [5], [6]. Moreover, these past methods have mainly focused on detecting deception through verbal communication (e.g., speech or text) and have ignored the potential of non-verbal (e.g., user activity or movement) deception detection, which has shown high success rates in the offline world [7], considering that non-verbal cues are 4.3 times more powerful than verbal cues in face-to-face communication [8]. This is a promising detection method that we have identified in our previous work and for which we presented experimental results in [9].

In this paper we propose a novel approach in using user non-verbal behavior data in social media in order to detect multiple account identity deception. In section II, we provide

M. Tsikerdekis and S. Zeadally are with the College of Communication and Information, University of Kentucky, KY 40506 USA. (e-mails: tsikerdekis@uky.edu, szeadally@uky.edu).

an overview on deception and identity deception, discuss about the problems revolving around current identity deception detection methods and highlight the contributions of this paper. In section III, we present our method along with the variables and our experimental case. In section IV, we provide results on the performance of our proposed method. Finally, section V presents the implications of this method in expanding the field of identity deception detection for the ever-growing social media domain.

II. Related Works

A. Deception and Identity Deception

Deception has been defined as the deliberate transfer of false information to a recipient that is not aware that the information received has been falsified [6], [10]. In nature it can be seen as a mechanism for gaining a strategic advantage [11]. Similarly, human deception is motivated by instrumental (goal-driven), relational (relationship-driven) and identity-driven goals [12]. The intent behind these goals may be benign (e.g., white lies) or hostile [13]. Online, the success of an attempt to deceive others is dependent upon multiple factors associated with the components involved: deceiver, social medium, potential victim and deceptive action [9]. Factors that affect a deceiver's behavior and effectiveness in achieving deception include a deceiver's expectations, goals, motivations, his/her relation to target and a target's degree of suspicion [12]. The last element in particular has been found to indirectly affect human deception detection rates [14]. A deceiver's goal is to use everything at his/her disposal to keep a low suspicion from his/her target and this applies to both verbal and non-verbal behaviors. There is also a moral cost for a deceiver that will affect the likelihood of using deception [15]. The software design of the social medium also affects deception through factors such as the perceived level of security provided by the system along with mechanisms that enhance trust and make assurances [16]. The deceptive action transmitted through cyberspace also has attributes such as the number of targets and the expiry date associated with it that influence its success [9]. Finally, a victim's ability to detect deception is an important factor that influences deception success. Humans have been consistently shown to be bad deception detectors [17]. Another important factor is a victim's Information Communication Technology (ICT) literacy [9]. For example, in a study involving Internet fraud through page-jacking techniques (developing fake pages of legitimate websites) only a handful of individuals detected inconsistencies with the fake websites [16].

Deception is achieved by manipulating content, the communication channel, the sender information, or any combinations of these three components [9]. Manipulating content involves tampering with images [18] or even text as can be seen in collaborative projects such as Wikipedia where special user task forces are focused on monitoring for text manipulation with the intention to spread inaccurate information [5]. Communication channels can be tampered with to disrupt communications of a user in an attempt to access his/her account or cause confusion between two parties (e.g., a case that can be observed online in video gaming consoles) [19].

Identity deception (a subcategory of deception) focuses on manipulating the sender's information [20] and can be divided into three categories: identity concealment (e.g., concealing or altering part of an individual's identity), identity theft (e.g., mimicking another person's real identity) and identity forgery (e.g., forging a fictional identity) [6]. Of particular interest for social media is identity forgery. Social media services tend to allow individuals to easily register new accounts without a thorough verification of an identity. In fact, an individual can have an unlimited number of accounts appearing as seemingly different users to unsuspected individuals.

B. Deception Detection

Deception detection theories are divided into those that are based on *leakage cues* (cues sent by the deceiver unwillingly due to factors such as cognitive overload) and *strategic decisions* (cues indicative of deception that are willingly transmitted by a deceiver in order to ensure deception success) [21]. To detect deception, both categories pick up cues from verbal and non-verbal communications.

Human deception detection is arguably the most widely used method. Individuals can pick up cues from the environment in which an interaction takes place (e.g., a photograph that looks edited) with a deceiver and interpret these cues by understanding a deceiver's goals [16]. The most critical factor in detecting deception is the time, which can vary from days to months, until a truth is uncovered by a previously deceived individual [22]. People, however, are bad at detecting deception with detection success bounded between 55 to 60 percent [23] at best while others have measured an even lower success of 34 percent [24]. Even more troublesome is that a study has found that upon training people in detecting verbal and non-verbal cues detection accuracy actually decreased [25]. A more standardized perspective of examining deception detection is necessary to achieve and engineer deception detection solutions with high success rates.

Three of the most popular theories used in the deception field are Interpersonal Deception Theory (IDT), Leakage Theory (LT), and Expectancy Violations Theory (EVT) [21], [26]. In Interpersonal Deception Theory, deception is seen as a series of exchanges between the deceiver and the victim. IDT sees deception as a goal-driven event. After each exchange, the deceiver adapts his/her behavior depending on the responses that he/she receives from his/her potential victim [12]. The adjustments made by the deceiver give away verbal and non-verbal indications for deception. IDT has been used as the theoretical premise for developing a framework for intent detection in deception (detecting whether intent is hostile or benign) [13].

Similarly, Leakage Theory also involves detecting indicators for deception but these are delivered unwillingly by the deceiver due to an inability to reproduce the equivalent of a truthful behavior in terms of verbal and non-verbal behavior

[27]. One possible explanation for this behavior is the cognitive overload created by a deceiver who attempts to control multiple facets of his/her behavior. In addition, a deceiver's awareness may also play a role in the types of cues that are leaked. People are naturally adapted to have a good control over their facial expression while body language (especially that of the lower body) is often seen as less useful from a deceiver. Similarly, in social media non-verbal behavior such as the time taken to type a text (although this can vary in different contexts) is likely to be seen as a deception cue by a deceiver much like pauses in speech help distinguish deceptive from non-deceptive speech [28].

Finally, Expectancy Violations Theory states that a person's normal behavior (i.e., baseline behavior) and the context in which this behavior takes place should also be considered [26], [29]. Instead of looking for indicators of deception, one can focus on comparing an expected interaction (based on one's baseline and context) with a received interaction. Any discrepancies between a baseline and an actual behavior can signal some probability for deception. For example, a profile page from an experienced social networking user is expected to vary compared to that of a freshly registered user. EVT has been used as the conceptual background for detecting deception through digital analysis of head and hand movements [26].

C. Identity Deception Detection

A particular issue with identity deception in social media is the presence of multiple identities by one user. Both online and offline studies have been conducted in an attempt to solve the problem of detecting duplicate account records. Wang et al. [6] in their study attempted to identify duplicate records in a criminal database using a variety of similarity-based detection algorithms. Attributes such as name, address, social security number and date of birth from a criminal database were compared as strings using a string comparator and the level of disagreement for these items was obtained between different user records. Furthermore, they obtained the overall disagreement between records based on these attributes, and those matches that had a disagreement below a certain threshold were considered as the same account. The most direct solution to identify duplicates in a database with the highest accuracy is a cross-comparison for the full length of accounts in a database. If one simply compares each account to all other accounts in the database this results in high computational overheads of $O(N^2)$. The solution adopted by Wang et al. was to use an adaptation of the Sorted Neighborhood Method (SNM). The original SNM develops a sorting key, sorts a database and then merges the duplicate records using a window of fixed size *w* that moves through the sorted records. The adapted SNM version has a shorter window *w'*, where *w'* is smaller than *w*. The window in the adapted version is smaller since once a duplicate record is found the rest of the comparisons for a window are ignored. The method produced high detection accuracy (80.4% for a dataset containing missing values for the previously mentioned attributes and 98.6% for a dataset without missing values) with a computational complexity of $O(w'N)$. The adapted SNM version took 6.5 minutes to complete with 1.3 million records while a record comparison (first approach) would have taken 87 days on the same machine. However, the time complexity for the adapted SNM does not include the sorting of the database. Furthermore, the method is focused on identity concealment and probably has limited application for cases of identity forgery where verbal information (e.g., profile text) in social media can be freely manipulated to an extraordinary degree compared to criminal databases. Finally, social media include a variety of types varying from blogs to social networking sites and even virtual social world that differ drastically in terms of what they offer to their users and their databases tend to be even larger than criminal databases.

A more recent study by Solorio et al. attempted to detect sockpuppets (new accounts of previously blocked users) on Wikipedia [5]. They used natural language processing techniques to detect users who maintain multiple accounts based on their verbal output. Textual features were used such as punctuation count, quotation count or the variation between using capital or lowercase "I". These features were tested against all revisions made by the users on pages throughout Wikipedia. Due to the volume of users on Wikipedia in conjunction with the number of revisions that each account may have (which can reach thousands), the similarity-based method used to identify a positive match between two accounts needs to receive manual input (an individual needs to set which two accounts need to be compared). As such, the method can be considered as a human-augmenting deception detection technique since it requires individuals to provide input for two potential accounts that match. A Support Vector Machine (SVM) model has shown 68.83% overall accuracy against an experimental dataset of 77 cases of legitimate users and sockpuppets. The limitation of this method is its computational cost involved if one would like to test all accounts against all accounts in a database; a time complexity of $O((N*R)^2)$ where R is the number of revisions made by a user. Testing every new account against all accounts currently in the database would result in a time complexity of $O(N*R)$.

The two aforementioned methods described earlier demonstrate the limited capabilities of using verbal communication to detect identity deception using account comparison techniques. These methods yield relatively high levels of detection accuracy. However, the cost for such accuracy may be too high for detecting duplicate accounts on social media. Similarity analyses, when used to evaluate a newly registered user with the rest of the database, also incur high computational overheads. Moreover, as we mentioned in a previous work [9], verbal deception detection as a detection methodology completely ignores non-verbal aspects which have shown to be highly effective in exposing deceivers [26], [30]. The most common argument based on the literature [11], [21] is the fact that in the case of the online environment, a deceiver will maximize his/her effort in ensuring that his/her verbal behavior does not expose the deception being carried out.

D. Contributions of this work

The main contributions of this work can be summarized as follows:

- We propose a computationally efficient method (applicable to all social media classifications [1]) for detecting identity deception through the use of non-verbal user activity in the social media environment. This contribution ensures that a relatively high level of overall detection accuracy is obtained that is comparable to similar methods that make use of verbal communication [5], [6] but with lower computational overheads.
- To demonstrate the computational efficiency (to withstand the immense traffic experienced by social media services) of our proposed non-verbal method to deception detection we use publicly available data from Wikipedia and machine learning algorithms.
- Finally, we present design guidelines for designers and developers interested in implementing this method as an added level of security for their social media communities and additional considerations based on various social media classifications in existence today.

III. Proposed Method for Detecting Online Identity Deception

A. Research Objectives

Our research goal in this work is to develop a method that can automatically detect online identity deception which can be very useful in many online social media scenarios. For instance, one scenario where such detection would be useful is in the case of an open source software development collaborative project website where, for security reasons, allows just one account per individual. Since new account registration is available to anyone, a user can therefore register an unlimited number of times every time his/her account gets blocked. Succeeding in identity deception is important for a deceiver who wants to inject malicious code into a project. Once an account is discovered, all changes made to the code by the owner of that account will be investigated and closely examined. We argue that an early detection system can help identify those individuals who experience a disproportionate familiarity with the collaborative software (according to their non-verbal behavior) which may indicate that they are not in fact newcomers or novices. Post-examination and close monitoring of suspect cases will help ensure the security of an open source project.

In this work, we investigate answers to the following research questions: *Can a user's non-verbal behavior in social media be effectively used in detecting identity deception in terms of multiple account ownership and is it more effective than previously proposed methods in the literature which used verbal similarity searching?*

Can the method be implemented with high computational efficiency and low overheads in large social media environment where we often have a large number of users?

To demonstrate our proposed method's effectiveness we use Wikipedia, which falls under the collaborative projects classification of social media [1] (shown in Table I), as our experimental case. We used publicly available data for Wikipedia in order to evaluate our approach. It is worth pointing out that although we have used Wikipedia as an example of a social medium, our method can be applied to virtually any other social medium environment. We briefly describe below some of the non-verbal user activities that can be observed on Wikipedia before describing our proposed method.

B. The Wikipedia Environment

Wikipedia is a free online encyclopedia in which everyone can contribute without an account (anonymously when only IP address is visible) and with an account using a pseudonym or even real name. Wikipedia operates on the concept of namespaces where each namespace is meant to include a specific type of content (or pages). For example, all encyclopedic articles belong to the "(Main/Article)" namespace (with numeric identifier 0) while all article discussions (involving discussions on improving articles and resolving issues) belongs to its own namespace ("Talk" namespace with numeric identifier 1). Wikipedia's policy pages and discussion on Wikipedia proposals or projects belong to different namespaces. We have a total of 28 namespaces in Wikipedia.

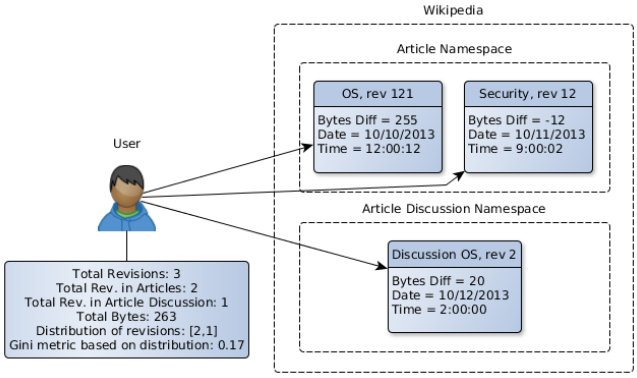


Fig. 1. An example of user activity on Wikipedia along with associated non-verbal activity.

Users leave a revision footprint on pages when they make a change to them. A page revision log is maintained for each page where everyone can find who did a specific revision, the revision itself and other associated matters relating to the revision (when it was made, how many bytes were added or removed from a page). A single user interaction with the Wikipedia's environment and two of its namespaces are illustrated in Fig. 1. The logged data on page revisions provide us with non-verbal user behavior on Wikipedia. For example, the time taken between each revision is a measurable non-verbal behavior.

C. Non-verbal Behavior Variables

We used simple and more complex variables to represent user behavior. Variables of online non-verbal behavior fall under two major categories: time-independent and time-dependent (henceforth these variables are denoted with index *t*).

We started with the number of total revisions (R_t) made by a user for a specific time window since their initial registration with the website. In addition, we obtained the number of revisions as they were distributed in the various namespaces such as article (Ra_t), article discussion (Rd_t), user page (Ru_t), user discussion page (Rt_t), Wikipedia-related pages and Wikipedia-related discussion pages combined under one variable (Rw_t). A final category was added for all the rest of the namespaces such as file uploads, images etc. (Ro_t). Based on these namespaces we also used a variable called the Gini coefficient that represents differences in activity distribution across these namespaces (bounded between 0 and 1) and is formally defined as [31]:

$$GR_t := 100\left[\frac{2\sum_{x=1}^{6}\left(w_x Rx_t \sum_{j=1}^{x} w_j\right)-\sum_{x=1}^{6} w_x{}^2 Rx_t}{\left(\sum_{x=1}^{6} w_x\right)\sum_{x=1}^{6}(w_x Rx_t)} - 1\right] \quad (1)$$

where *x* represents the set of items (revisions on each namespace for our case) and *w* is the relevant weight that may be assigned to each item. Equal weights were applied to our data because, conceptually, namespaces on Wikipedia do not hold any weight and any attempt to assign weights would introduce a bias. In addition, we measured the mean number of bytes of bytes added or removed by all revisions:

$$\overline{RB}_t = \frac{\sum_{i=1}^{R_t} RBi}{R_t} \quad (2)$$

The total number of bytes added (Ba_t) and total number of bytes removed (Br_t) from all the revisions during the observation window were also calculated. Furthermore, the time difference in seconds between the time (TR) a user registered their account until the time of the first revision was measured along with the namespace (FE) where their first revision was made. Finally, the average duration (AD_t) between revisions was used and is defined as follows:

$$AD_t = \frac{\sum_{i=2}^{Rt} T_i - T_{i-1}}{Rt} \quad (3)$$

where *n* is the total number of revisions and *T* is the set of all Unix times for each revision made.

TABLE II
EXAMPLES OF USER BLOCKS FOUND IN WIKIPEDIA BLOCK LOGS

Example of user block due to vandalism (e.g., adding false or inaccurate information to pages with malicious intent)	Example of user block due to sockpuppetry
id="4933953" **user**="198.202.26.110" **by**="Ronhjones" **timestamp**="2013-12-30T00:23:57Z" **expiry**="2014-01-13T00:23:57Z" **reason**="[[WP:Vandalism|Vandalism]]"	**id**="4933944" **user**="Niroshvthanaw" **by**="Anna Frodesiak" **timestamp**="2013-12-29T23:56:29Z" **expiry**="infinity" **reason**="Abusing [[WP:Sock puppetry|multiple accounts]]: See [[Wikipedia:Sockpuppet investigations/Masu7]]"

D. Data Retrieval and Model Testing

We collected a list of all publicly available logs of blocked users on Wikipedia during the period since February 2004 until October 2013. The logs include various reasons for blocking user accounts including account blocks for verified sockpuppet cases (examples of block logs shown in Table II). Using regular expressions we kept only sockpuppet cases with an infinite time of block issued for these accounts.

These are users who make a great effort in using deception to masquerade as legitimate users while still trying to achieve their end goals (e.g., altering a text in a particular article page). For example, in the page that holds the discussion[1] over the block of user "Niroshvthanaw" the following is written about the account puppeteer: "Masu 7 has created another sock, this time changing references to Horana Royal College to Royal College, Horana against consensus. Similar behavior has been shown by User:Xe2oner, User:Wo2gana, User:Samudrab all of whom have been blocked as socks of Masu 7". Individuals like this user attempt to deceive without getting caught and this is revealed by the time taken to block these accounts since the initial first revision on Wikipedia. On average it takes approximately 75 days for a sockpuppet account to get blocked (median is 3.19 days) as evident in our block log dataset. Fig. 2 depicts all the sockpuppet cases showing their first revision and time when the account was blocked. About 38.96 percent of sockpuppets have their accounts blocked

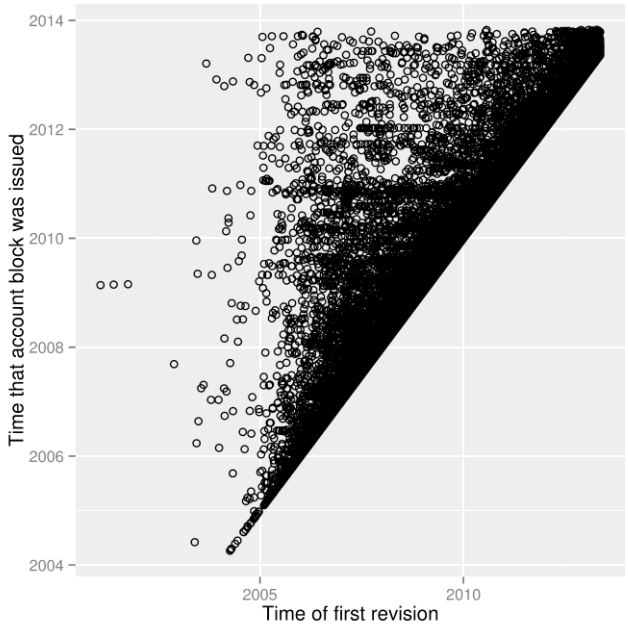


Fig. 2. Scatterplot showing the time when a sockpuppet made the first revision and the time when the account was blocked.

[1] http://en.wikipedia.org/wiki/Wikipedia:Sockpuppet_investigations/Masu7/Archive

during the first day after their first revision on Wikipedia. Ten days after their first revision, the percentage of sockpuppets being caught rises to 62.24 percent. By 30 days, the percentage rises to 74.43. It is quite clear that, while many users are caught early on, others evade detection for a considerable amount of time.

For testing our proposed method we sampled 7500 cases of sockpuppets. In addition, we retrieved a list of all users who made at least one revision through the revision records on all Wikipedia namespaces (these are provided as dump xml files and were parsed). Verified sockpuppet cases were removed from this list and an additional sample of non-blocked users was obtained so that our final user list contained 7500 verified sockpuppet cases and 7500 legitimate user cases. As such, a fair coin toss for our sample would produce approximately 50% accuracy in detecting sockpuppets. Human deception detection is usually placed at much lower rates (as low as 30-50%) [24].

For each one of these users in our total sample (sockpuppets and legitimate users) we obtained all activity on Wikipedia. This activity can be translated into variables which can help us test models (e.g., a model can consist of one or more variables described previously) for our proposed method. The time window set for the user activity will affect all time-dependent non-verbal behavior variables and in turn the efficiency of a model in terms of its predictive accuracy. It will also force some of the cases in our sample to be omitted due to inactivity (e.g., a user who made his/her first revision two hours after registration would be omitted from the sample if the time window is set for an hour after registration). We obtain all activity for users in the first 30 days. The users in the sample were not banned before these 30 days. Those who have not been active for that time window were excluded because without the presence of any behavior we cannot build a classification. The final sample (for 30 days of user activity) consisted of a total of 12,723 users of which roughly 48.23 percent were sockpuppets. We calculated all variables of non-verbal behavior for all users in our sample. Just like in the real world, these variables are similar to measuring non-verbal behavior accompanying verbal interactions such as measuring the speed of delivery of a speech of a person and looking for deviations from a context specific baseline.

Time-dependent variables are likely to affect our models depending on the time *t* we would set when testing their performance (and subsequently the overall performance of our proposed method). We hypothesized that since Wikipedia does not encourage the use of multiple accounts, the expectation is that a newly registered user will also behave as a newcomer. Newcomers are generally unfamiliar with the environment (or then tend to exhibit limited familiarity with the system). In contrast, a deceiver is not only expected to be familiar with the Wikipedia space but to be also very familiar with many of the norms and behaviors of legitimate users. Since deception can be easily detected through verbal communication (e.g., the textual contents of a revision), a deceiver is likely to be extra cautious when delivering text. In contrast, control over a deceiver's non-verbal behavior is less likely because he/she is not aware that this may be monitored and it is less obvious to him/her while interacting with the social medium. In addition, long-term behaviors among deceivers and real users are expected to vary. To be able to identify the best time window we have calculated all non-verbal, time-dependent variables for each hour during the first 30 days of activity for all users in the sample. Then the standardized difference between sockpuppets and real users was calculated for each variable and the trend was represented using locally weighted scatterplot smoothing (Fig. 3). For the standardized difference we used the correlation coefficient

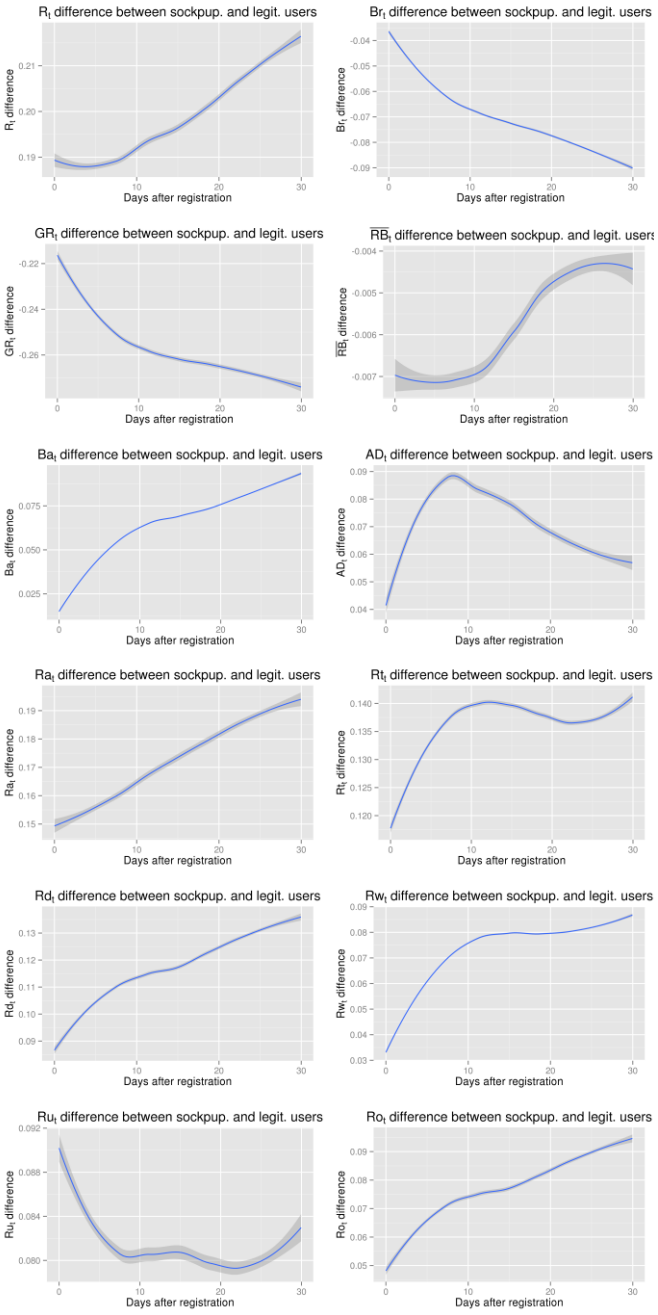


Fig. 3. Variation of differences in non-verbal user activity variables between sockpuppets and legitimate users over a period of a month. Positive scores along the y axis indicate increased activity for sockpuppets whereas negative scores indicate decreased activity for sockpuppets compared to legitimate users.

produced by point-biserial correlation:

$$r_{pb} = \frac{M_1 - M_0}{s_n} \sqrt{\frac{n_1 n_0}{n^2}} \quad (4)$$

where s_n is the standard deviation, M_1 is the mean of a variable (e.g., R_t) for the sockpuppet group A and M_0 is the mean of the same variable for the legitimate user group, n_1 is the number of data points for group A, n_2 is the number of data points for group B, and, n is the total number of data points for both groups.

There are substantial differences for some variables early on but behaviors of deceivers seem to deviate more as time progresses. This is evident particularly for variables R_t, GR_t, Rd_t and Rt_t where changes in the correlation coefficient are particularly large. Additionally, in some cases (e.g., GR_t) deceptive accounts seem to deviate from legitimate accounts in a negative trend. In this particular case, it is evident that sockpuppets tend to distribute their efforts in more namespaces as time progresses contrary to legitimate users. Others tend to reach a maximum and then stabilize or reach a lower value and then become stable. In most cases, it seems that, as time progresses, a deceiver's behavior tends to deviate more than that of a legitimate user. This is similar to real life human deception detection where time is an important factor for uncovering a lie [22] due to the likelihood of exposing such deviations.

To identify the differences in accuracy as time progresses we have calculated several models (Mx_t) for t = 1 day (11207 cases, baseline for non-sockpuppet cases at 52.86%) and t = 30 days after registration (12,723 cases, baseline for non-sockpuppet cases at 51.77%). We developed several binary outcome models aimed at using these non-verbal behavior variables to detect identity deception. The following models were developed:

- $M1 \sim FE + TR$
- $M2_1 \sim FE + TR + R_1 + GR_1$
- $M3_1 \sim FE + TR + R_1 + GR_1 + \overline{RB}_1 + Ba_1 + Br_1 + AD_1$
- $M4_1 \sim FE + TR + R_1 + GR_1 + \overline{RB}_1 + Ba_1 + Br_1 + AD_1 + Ra_1 + Rd_1 + Ru_1 + Rt_1 + Rw_1 + Ro_1$

TABLE III
CLASSIFICATION MATRIX USED TO EVALUATE THE EFFICIENCY OF A MODEL-ALGORITHM PAIR

	Verified Identity Deception (Sockpuppetry)	Verified Legitimate User
Predicted Identity Deception (Sockpuppetry)	True Positive (TP)	False Positive (FP)
Predicted Legitimate User	False Negative (FN)	True Negative (TN)

- $M2_{30} \sim FE + TR + R_{30} + GR_{30}$
- $M3_{30} \sim FE + TR + R_{30} + GR_{30} + \overline{RB}_{30} + Ba_{30} + Br_{30} + AD_{30}$
- $M4_{30} \sim FE + TR + R_{30} + GR_{30} + \overline{RB}_{30} + Ba_{30} + Br_{30} + AD_{30} + Ra_{30} + Rd_{30} + Ru_{30} + Rt_{30} + Rw_{30} + Ro_{30}$

IV. PERFORMANCE EVALUATION

We used a popular set of machine learning algorithms, which includes Support Vector Machine (SVM), Random Forest (RF) and Adaptive Boosting (ADA), to implement our proposed models. A description of how these algorithms work is beyond the scope of this paper but more details can be found in books describing them [32], [33]. However, it should be noted that all the selected algorithms used are considered ideal for models that involve binary outcomes as it was the case in our study [32].

A. Performance Metrics Used

To evaluate the efficiency of our models for our proposed method we used the following classification matrix shown in Table III.

Using this matrix, we derive results to measure the following performance metrics in order to evaluate the performance of our models for our proposed method: ***recall*** (the fraction of valid sockpuppet cases that are returned), ***precision*** (the fraction of returned cases that are valid sockpuppet cases), ***F-measur***e (the test of a model's accuracy bounded between 0 and 1 that combines recall and precision), ***accuracy*** (the fraction of true positives and true negatives returned over the total number of cases), ***false positive rate*** (indicating the rate of falsely identified sockpuppets), and ***Matthews correlation coefficient*** (a performance metric used in machine learning that provides a balanced result even if cases in the sample vary substantially in size). These performance metrics are formally defined as follows [34]:

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

procedure TenTimesTenFoldCrossValidation()
1. *// Algorithm builds a single model (e.g., RF) and produces final results*
2. Set a predefined number *w*
3. LOOP: *n* for *T=[1,2,…,9,10]*
4. Set random seed $S = w * n * 10$
5. Create fold sample list *FLi* by randomly assigning fold numbers to the full length of dataset
6. LOOP: *f* in *TT=[1,2,…,9,10]*
7. Build Random Forest model *RF* based on training data (*FLi* not equal to *f)* and *S*
8. Calculate predictions *Pi* using *RF* for testing data (*FLi* equal to *f*)
9. Set *Oi* as observed values (is or is not a sockpuppet) based on testing data
10. Build classification matrix using observed *Oi* and predicted *Pi* values
11. Calculate Recall RE_f, Precision PR_f, and F-measure FM_f
12. END LOOP
13. Calculate $RE_n = \sum_{i=1}^{f} RE_f / f$, $PR_n = \sum_{i=1}^{f} PR_f / f$, and $FM_n = \sum_{i=1}^{f} FM_f / f$
14. END LOOP
15. Calculate $RE = \sum_{i=1}^{n} RE_n / n$, $PR = \sum_{i=1}^{n} PR_n / n$, and $FM = \sum_{i=1}^{n} FM_n / n$

End TenTimesTenFoldCrossValidation

Algorithm 1. A repeated ten times ten-fold cross-validation algorithm for testing a model using random forest.

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

$$F-measure = \frac{2*Precision*Recall}{Precision+Recall} \quad (7)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

$$False\ Positive\ Rate\ (FPR) = \frac{FP}{FP+TN} \quad (9)$$

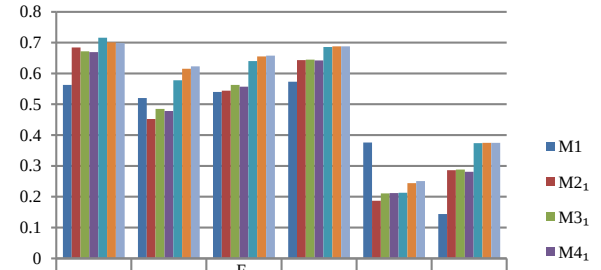
$$MCC = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (10)$$

B. *Experimental Procedure*

To evaluate the performance efficiency of our models for our proposed method we repeated ten times a ten-fold cross-validation procedure to obtain the mean values for all of our performance metrics. Algorithm 1 is used to evaluate our models. The algorithm involves splitting the data in ten parts and using nine of them to build a model while one part is used for testing the model.

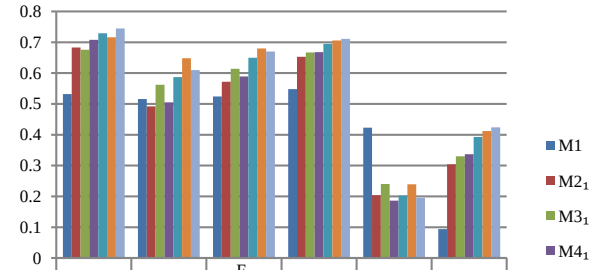
The algorithm is sequentially executed until all possible ten combinations have been used. This process is repeated ten times and each time we used a different seed for splitting the dataset. We used this algorithm because it has been previously proven to produce highly accurate estimates in terms of how models (and as a result our overall method) would perform in previously unseen data [32].

Test results obtained are presented in Figs. 5a, 5b, 5c for all models and algorithms used. These are rounded to three decimal digits.



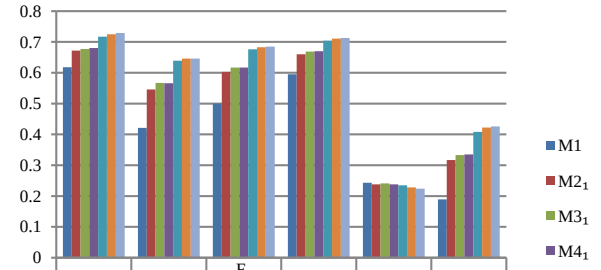
	Precision	Recall	F-measure	Accuracy	FPR	MCC
M1	0.563	0.52	0.54	0.573	0.376	0.144
M2$_1$	0.684	0.452	0.544	0.643	0.187	0.286
M3$_1$	0.672	0.485	0.563	0.645	0.211	0.288
M4$_1$	0.669	0.478	0.557	0.642	0.212	0.281
M2$_{30}$	0.716	0.578	0.64	0.686	0.213	0.374
M3$_{30}$	0.701	0.615	0.655	0.688	0.244	0.375
M4$_{30}$	0.698	0.623	0.658	0.688	0.251	0.375

Fig. 5a. SVM results for all models.



	Precision	Recall	F-measure	Accuracy	FPR	MCC
M1	0.532	0.516	0.524	0.548	0.423	0.094
M2$_1$	0.683	0.492	0.572	0.653	0.204	0.304
M3$_1$	0.676	0.562	0.614	0.667	0.24	0.33
M4$_1$	0.708	0.505	0.589	0.668	0.186	0.337
M2$_{30}$	0.729	0.587	0.65	0.695	0.203	0.393
M3$_{30}$	0.716	0.648	0.68	0.706	0.239	0.412
M4$_{30}$	0.745	0.61	0.67	0.711	0.195	0.424

Fig. 5b. RF results for all models.



	Precision	Recall	F-measure	Accuracy	FPR	MCC
M1	0.618	0.421	0.5	0.595	0.243	0.189
M2$_1$	0.672	0.546	0.603	0.66	0.238	0.317
M3$_1$	0.677	0.567	0.617	0.669	0.241	0.333
M4$_1$	0.68	0.566	0.617	0.67	0.238	0.335
M2$_{30}$	0.717	0.639	0.676	0.704	0.235	0.408
M3$_{30}$	0.725	0.646	0.683	0.711	0.228	0.422
M4$_{30}$	0.729	0.646	0.685	0.713	0.224	0.426

Fig. 5c. ADA results for all models.

Fig. 5. Performance results for all models and algorithms used in this study. Y-axis represents results for all of our performance metrics (bounded between 0 and 1).

We summarize the results obtained with our proposed detection method compared to two other previously proposed approaches and the results are summarized in the Table IV.

V. Discussion of Results

Based on the results obtained, we found that Adaptive Boosting appears to provide the best balance between recall and precision while maintaining the highest achieved accuracy. Recall levels are relatively high (64.8 percent as best case) which means that most cases are picked up by the our proposed method. In terms of precision, we found that a relatively large amount of false positives is obtained (best case RF $M4_{30}$ still leaves 25.5 percent of false positives). This is not necessarily a bad result unless the detection method were to be implemented so that it automatically would block suspect cases. If the detection method is used to report suspect cases so that administrators can keep a close eye on or restrict certain features for suspect accounts for a time period, then recall is the most important feature and low precision can be tolerated.

TABLE IV
Comparison of identity deception detection techniques for multiple accounts owned by the same user.

	Adaptive SVM Text Attribute Disagreement Algorithm [6]	Natural Language Processing Similarity Searching [5]	Non-Verbal Expectancy Violations Detection **(Our Proposed Detection Method)**
Overall accuracy	80.4% - 98.6%	68.8%	71.3%
Indicators used	Verbal	Verbal	Non-verbal (Verbal can be added however)
Limitations	Limited to cases where profile attributes are provided	Limited to cases where text is communicated through	Limited to cases where data on user activity is available
Efficiency for analyzing a newly registered user	*O(w'N')* (N' is smaller than the total number of users in database focusing on records close to the new account)	*O(N*R)* (N number of users in database and R all revisions made by each user)	*O(1*R')*, R' is a limited amount of revisions (*R'<R*) made by the user in question in the window of observation
Time of application	As soon as data is added on a profile (missing values are allowed)	As soon as a user posts a text somewhere (preferably enough cases for the algorithm to pick up on cues)	After a set time window (e.g., 12 hours) that distinguishes newcomer from old user

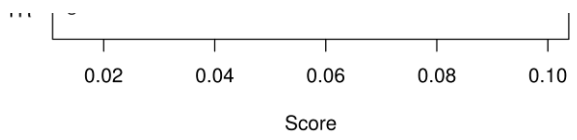


Fig. 6. Variable importance plot for $M4_{30}$ based on the adaptive boosting algorithm.

The machine learning algorithms that we used in this study usually provide higher accuracy results than traditional models (e.g., binary logistic regression [35]) used in statistical research). However, they are "black-boxes" in the sense that they produce results but it is less evident how variables affect a prediction. The adaptive boosting package in R (ada) does offer a measure of the importance of variables that are included in a model [36]. The measure calculates how each variable improves the predictive accuracy. We present results for variable importance for $M4_{30}$ based on the adaptive boosting algorithm shown in Fig. 6. In conjunction with the results shown in Fig. 3, the distribution of edits is a powerful predictor for deception just as we hypothesized that is likely to be. Deceivers have a higher probability of delivering content to multiple namespaces as opposed to newcomers who are less likely to do so. The total number of revisions made by a deceiver also demonstrate that deceivers tend to be more active and in namespaces other than the article's namespace. Moreover, the average duration between revisions shows that deceivers take longer times between posting their revisions. A plausible explanation for this result is that deceivers need to take longer to make strategic decisions to ensure success for their deception.

The results obtained show that the use of non-verbal user activity is a viable and efficient method for detecting identity deception (specifically sockpuppetry). Our method achieved an overall accuracy of 71.3% in identifying deceivers. The method also incurs a much lower computational overhead over previous methods while achieving an overall accuracy that renders it a valid choice for an early filtering system. Moreover, although we have used Wikipedia as an example of a social medium, this deception detection method can be applied to other social media domains. In fact, the detection method can be used with any social media service that contains user footprints that are not only verbal (e.g., text, audio, video) but also non-verbal (e.g., frequency of posting, time between updates, length or duration of messages). Moreover, the method also demonstrated the value of using non-verbal communication to detect identity deception in real time with limited resources (given that it is computationally efficient).

One possible explanation as to why our proposed detection method is effective can be found from IDT and Leakage Cues Theory. A deceiver is constantly evaluating the receiver and is continuously adjusting his or her behavior accordingly. Such adjustments are likely to be applied to what a deceiver can perceive as something communicated to a receiver and other third parties present within the observable vicinity of a deceiver. Non-verbal activity for a deceiver is less likely to be perceived as monitored especially in a digital environment. In addition, even if such activity is controlled, certain cues will still leak and leave a footprint which others can make use of later on. For example, the impatience of sending messages to multiple namespaces right after an account registration is less likely to be controlled. Based on results obtained in this work, we argue that a deceiver is less likely to attribute importance

to non-verbal activity on social media. The deceiver is less aware that there is a footprint for that online activity, and also less aware that others can detect this footprint and he or she is also likely to have less control over controlling such non-verbal activity.

Our results have also contributed to a new perspective on sockpuppetry where it is a challenging to detect identity deception. Our results show that in online communities where one account per user is enforced by a social media service's policy, sockpuppets will deviate from the baseline behavior of newcomers. This deviation is in line with EVT and we have demonstrated that sockpuppets tend to be more active than newcomers possibly due to their prior knowledge and skills with various other systems.

A. *Limitations of our Proposed Detection Method*

The efficiency and effectiveness of our proposed detection method is influenced by several context specific factors. The time window set for observing early new user behavior has a significant impact on the method's effectiveness. It can also affect the efficiency if the window is too large given that more data will be needed to be examined by the detection method. Another issue is the identification of measurable non-verbal behavior in social media. We have demonstrated a couple of examples in our paper. More work is needed in the future to implement such a method to identify the most optimum set of variables that can assist in detecting identity deception and are also computationally efficient. However, our method based on the expectancy violations theory is still superior in deception detection compared to methods of similarity searching and text comparative methods used by other previously proposed detection techniques [5], [6] . Finally, the social medium under examination will also determine the data that can be used. It is worth pointing out that although our method is portable to any social media classification, adaptations may be needed to ensure its proper implementation. Research is a necessary step to identify what non-verbal behaviors can be consistently and quantitatively be translated into variables that can be included in a predictive model. These behaviors will need to be good indicators (at least conceptually) of a substantial difference between how legitimate and deceitful users operate. After these variables are identified, one will need to develop models to find the most optimum model with the highest predictive accuracy. It requires a lot of work up front but the computational and practical efficiency of the method may prove beneficial to other social media services.

B. *Future work*

Future work will need to examine other non-verbal behavior variables in different social media services that can be used as good indicators of deception. Moreover, combining research on verbal detection deception with the non-verbal behavior deception detection method presented in this study may help improve prediction accuracy.

VI. Conclusion

Despite the explosive growth of social media applications and networks, deception in social media environment is an area that has not received commensurate attention from researchers, designers, and developers. Identity deception in particular is something that has haunted the Internet with a number of incidents receiving attention because of the ease of creating new accounts. Given the increasing number of Internet users and social media users, identity deception is likely to increase and the discussion on deception detection will become even more important. There are automated solutions that guarantee higher detection rates than human detection but the computational challenges of monitoring verbal communications are many. Non-verbal behavior monitoring for deception detection is an alternative path that can be used as a leading or complimentary detection solution. A coordinated effort is required to test these solutions on different platforms and advance the field of social media identity deception detection.

References

[1] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," *Bus. Horiz.*, vol. 53, no. 1, pp. 59–68, 2010.

[2] X. (Sherman) Shen, "Security and privacy in mobile social network [Editor's Note]," *IEEE Netw.*, vol. 27, no. 5, pp. 2–3, Sep. 2013.

[3] J. Brenner and A. Smith, "72% of Online Adults are Social Networking Site Users," 2013.

[4] C. S. Bhat, "Cyber Bullying: Overview and Strategies for School Counsellors, Guidance Officers, and All School Personnel," *Aust. J. Guid. Couns.*, vol. 18, no. 01, pp. 53–66, 2008.

[5] T. Solorio, R. Hasan, and M. Mizan, "A Case Study of Sockpuppet Detection in Wikipedia," in *Proceedings of the Workshop on Language Analysis in Social Media*, 2013, pp. 59–68.

[6] G. A. Wang, H. Chen, J. J. Xu, and H. Atabakhsh, "Automatically detecting criminal identity deception: an adaptive detection algorithm," *Syst. Man Cybern. Part A Syst. Humans, IEEE Trans.*, vol. 36, no. 5, pp. 988–999, 2006.

[7] T. O. Meservy, M. L. Jensen, J. Kruse, J. K. Burgoon, J. F. Nunamaker Jr., D. P. Twitchell, G. Tsechpenakis, and D. N. Metaxas, "Deception detection through automatic, unobtrusive analysis of nonverbal behavior," *Intelligent Systems, IEEE*, vol. 20, no. 5. pp. 36–43, 2005.

[8] M. Argyle, V. Salter, H. Nicholson, M. Williams, and P. Burgess, "The Communication of Inferior and Superior Attitudes by Verbal and Non-verbal Signals *," *Br. J. Soc. Clin. Psychol.*, vol. 9, no. 3, pp. 222–231, 1970.

[9] M. Tsikerdekis and S. Zeadally, "Online Deception in Social Media," *To Appear Commun. ACM*, 2014.

[10] P. Ekman, "Deception, Lying, and Demeanor," in in *States of Mind : American and Post-Soviet Perspectives on Contemporary Issues in Psychology: American and Post-Soviet Perspectives on Contemporary Issues in Psychology*, D. F. Halpern and A. E. Voiskounsky, Eds. Oxford University Press, 1997, pp. 93–105.

[11] J. S. Donath, "Identity and deception in the virtual community," in in *Communities in Cyberspace*, M. A. Smith and P. Kollock, Eds. Routledge, 1999.

[12] D. B. Buller and J. K. Burgoon, "Interpersonal Deception Theory," *Commun. Theory*, vol. 6, no. 3, pp. 203–242, Aug. 1996.

[13] J. Burgoon, M. Adkins, J. K. M. L. Jensen, T. Meservy, D. P. Twitchell, A. Deokar, J. F. Nunamaker, S. Lu, G. Tsechpenakis, D. N. Metaxas, and R. E. Younger, "An Approach for Intent Identification by Building on Deception Detection," *System Sciences, 2005. HICSS '05. Proceedings of the 38th Annual Hawaii International Conference on*. p. 21a–21a, 2005.

[14] R. J. Boyle and C. P. Ruppel, "The Impact of Media Richness, Suspicion, and Perceived Truth Bias on Deception Detection," in *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, 2005, p. 20a–20a.

[15] A. C. Squicciarini and C. Griffin, “An Informed Model of Personal Information Release in Social Networking Sites,” in *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, 2012, pp. 636–645.

[16] S. Grazioli and S. L. Jarvenpaa, “Perils of Internet fraud: an empirical investigation of deception and trust with experienced Internet consumers,” *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 30, no. 4. pp. 395–410, 2000.

[17] C. E. Lamb and D. B. Skillicorn, “Detecting deception in interrogation settings,” in *Intelligence and Security Informatics (ISI), 2013 IEEE International Conference on*, 2013, pp. 160–162.

[18] A. C. Popescu and H. Farid, “Exposing digital forgeries by detecting traces of resampling,” *Signal Processing, IEEE Transactions on*, vol. 53, no. 2. pp. 758–767, 2005.

[19] A. Podhradsky, R. D’Ovidio, P. Engebretson, and C. Casey, “Xbox 360 Hoaxes, Social Engineering, and Gamertag Exploits,” in *System Sciences (HICSS), 2013 46th Hawaii International Conference on*, 2013, pp. 3239–3250.

[20] T. Madhusudan, “On a text-processing approach to facilitating autonomous deception detection,” *System Sciences, 2003. Proceedings of the 36th Annual Hawaii International Conference on*. p. 10 pp., 2003.

[21] S. L. Humpherys, K. C. Moffitt, M. B. Burns, J. K. Burgoon, and W. F. Felix, “Identification of fraudulent financial statements using linguistic credibility analysis,” *Decis. Support Syst.*, vol. 50, no. 3, pp. 585–594, Feb. 2011.

[22] H. S. Park, T. Levine, S. McCornack, K. Morrison, and M. Ferrara, “How people really detect lies,” *Commun. Monogr.*, vol. 69, no. 2, pp. 144–157, Jun. 2002.

[23] T. H. Feeley and M. J. Young, “Humans as lie detectors: Some more second thoughts,” *Commun. Q.*, vol. 46, no. 2, pp. 109–126, Mar. 1998.

[24] T. R. Levine, R. K. Kim, H. Sun Park, and M. Hughes, “Deception Detection Accuracy is a Predictable Linear Function of Message Veracity Base-Rate: A Formal Test of Park and Levine’s Probability Model,” *Commun. Monogr.*, vol. 73, no. 3, pp. 243–260, Sep. 2006.

[25] S. Kassin and C. Fong, “‘I’m Innocent!’: Effects of Training on Judgments of Truth and Deception in the Interrogation Room,” *Law Hum. Behav.*, vol. 23, no. 5, pp. 499–516, 1999.

[26] S. Lu, G. Tsechpenakis, D. N. Metaxas, M. L. Jensen, and J. Kruse, “Blob Analysis of the Head and Hands: A Method for Deception Detection,” *System Sciences, 2005. HICSS ’05. Proceedings of the 38th Annual Hawaii International Conference on*. p. 20c–20c, 2005.

[27] P. Ekman and W. V. Friesen, “Nonverbal leakage and clues to deception.,” *Psychiatry Interpers. Biol. Process.*, vol. 32, no. 1, pp. 88–106, 1969.

[28] S. Benus, F. Enos, J. Hirschberg, E. Shriberg, S. R. I. International, and M. Park, “Pauses in Deceptive Speech,” *Speech Prosody*, vol. 18, pp. 2–5, 2006.

[29] J. K. Burgoon, “A communication model of personal space violations: Explication and an initial test,” *Hum. Commun. Res.*, vol. 4, no. 2, pp. 129–142, Dec. 1978.

[30] J. Hirschberg, S. Benus, J. M. Brenier, F. Enos, S. Friedman, S. Gilman, C. Girand, M. Graciarena, A. Kathol, and L. Michaelis, “Distinguishing deceptive from non-deceptive speech,” in *Interspeech 2005*, 2005, pp. 1833–1836.

[31] A. Alfons and M. Templ, “Estimation of Social Exclusion Indicators from Complex Surveys: The R Package laeken,” *J. Stat. Softw.*, vol. 54, no. 15, 2013.

[32] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques*. Elsevier Science, 2011.

[33] G. Williams, *Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery*. Springer, 2011.

[34] D. M. W. Powers, “Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation,” *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.

[35] A. Field, *Discovering Statistics Using SPSS*, vol. 58, no. 3. London: SAGE, 2009, p. 821.

[36] M. Culp, K. Johnson, and G. Michailides, “ada: An r package for stochastic boosting,” *J. Stat. Softw.*, vol. 17, no. 2, 2006.