Detecting Deception in Online Social Networks

Jalal S. Alowibdi[1,2], Ugo A. Buy[1], Philip S. Yu[1], Leon Stenneth[3]

[1] Department of Computer Science, University of Illinois at Chicago, Illinois, USA
[2] Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia
[3] Nokia Connected Driving, Chicago, Illinois, USA

{jalowibd,buy,psyu}@cs.uic.edu and leon.stenneth@nokia.com

***Abstract*—Over the past decade Online Social Networks (OSNs) have been helping hundreds of millions of people develop reliable computer-mediated relations. However, many user profiles in OSNs contain misleading, inconsistent or false information. Existing studies have shown that lying in OSNs is quite widespread, often for protecting a user's privacy. In order for OSNs to continue expanding their role as a communication medium in our society, it is crucial for information posted on OSNs to be trusted. Here we define a set of analysis methods for detecting deceptive information about user genders in Twitter. In addition, we report empirical results with our stratified data set consisting of 174,600 Twitter profiles with a 50-50 breakdown between male and female users.**

Our automated approach compares gender indicators obtained from different profile characteristics including first name, user name, and layout colors. We establish the overall accuracy of each indicator and the strength of all possible values for each indicator through extensive experimentations with our data set. We define *male trending* users and *female trending* users based on two factors, namely the overall accuracy of each characteristic and the relative strength of the value of each characteristic for a given user. We apply a Bayesian classifier to the weighted average of characteristics for each user. We flag for possible deception profiles that we classify as male or female in contrast with a self-declared gender that we obtain independently of Twitter profiles. Finally, we use manual inspections on a subset of profiles that we identify as potentially deceptive in order to verify the correctness of our predictions.

***Index Terms*—Deception detection, profile characteristics, gender classification, Twitter.**

I. Introduction

Online Social Networks (OSNs) have spread at stunning speed over the past decade. They are now a part of the lives of hundreds of millions of people world wide. The growth in the user base has led to a dramatic increase in the volume of generated data. The onset of OSNs has stretched the traditional notion of "community" to include groups of people who have never met in person but communicate with each other through OSNs to share knowledge, opinions, interests and activities.

Unfortunately, a lot of information posted on OSNs is not trusted. A recent study by McAfee showed that around 90% of young people believe that it is dangerous to post such information as personal data in OSNs [21]. Ensuring integrity, accessibility and confidentiality of information in OSNs is a major concern to people, organizations and companies that use OSNs.

The long-term objective of this project is to flag automatically deceptive information in user profiles and posts, based on detected inconsistencies in a user's profile and posts. In the past we investigated gender classification on Twitter using profile colors, first names, and user names [3], [4]. Here we focus on detection of false information involving user gender. We specifically studied the effectiveness of each characteristic in predicting the gender of Twitter users contained in a data set that we harvested between January and February 2014. The data set consists of about 174,600 Twitter user profiles, stratified for a 50-50 breakdown between male and female users. Each user profile in our data set has a link to a Facebook page in which users declare explicitly their gender. We have used this information from linked Facebook profiles as the ground truth throughout our studies. The outcome of those studies is that such characteristics as the first name, user name and background color chosen by a user for her profile can provide reasonably accurate predictions of the user's gender.

Here we report our most recent empirical results on a broader range of profile characteristics for gender classification than we considered previously and we define a new method for detecting information likely to be false in Twitter user profiles. Our method is centered on a Bayesian classifier that takes into account seven different profile characteristics and returns so-called *male trending* and *female trending* factors, which correlate to the probability that a Twitter user is male or female. For each Twitter user, the profile characteristics include the following seven items: First name, user name, and five colors that users can choose in their profiles (e.g., background color and text color of Twitter postings). Our classifier works in such a way that the computed male and female trending factors will take non-negative values complementing each other with respect to one. Thus, if the male trending factor for a given user u is m, with $0 \leq m \leq 1$, the female trending factor for u will be $f = 1 - p$.

An additional advantage of our approach is language independence. This is an important factor for an OSN like Twitter, which contains more than 70 languages [24]. Of the seven characteristics contained in our feature set, the five color-based features (i.e., background color, text color, link color, sidebar fill color, and sidebar border color) are clearly independent of a user's language. We achieve language

independence for first names and user names by translating those features into a phonetic representation independent of the name's original language and alphabet. Consequently, we can apply our predictions on deceptive profiles to any Twitter users, regardless of their language.

Our preliminary results with our dataset are quite encouraging. On the one hand, when used in combination our profile characteristics predict the gender of Twitter users with a high degree of accuracy. We can also flag gender deception with reasonable accuracy. On the other hand, our approach uses a relatively modest number of features, resulting in a low-dimensional feature space. We have deliberately excluded text-based characteristics, such as words and phrases appearing in posts, because this strategy would result in millions of features and introduce dependencies on the language of the posts. Consequently, our approach combines a good accuracy and language independence with low computational complexity.

Our Bayesian gender classifier uses two factors for each characteristic in order to predict the gender of a Twitter user. The first factor is the overall *accuracy* of the characteristic as a gender predictor. For instance, when predicting gender with first names alone, we obtain an accuracy of 82% over our entire data set; user names yield an accuracy of 70%, and the 5 color features combined yield an accuracy of 75%. These results are in line with results we published previously [3]. The second factor is the *gender sensitivity* of a user's value for a given characteristic. For instance, the English-language first name "Mary" has high sensitivity for the female gender whereas the name "Pat" has a low sensitivity because it is common among male users as well as female users. We have ignored text-based characteristics such as posted texts because of their high complexity. Other authors have shown that gender can be predicted from text with accuracy ranging from 68% to 78% [8], [28].

We identify deceptive profiles in our data set according to the following paradigm.

1) First, our Bayesian classifier returns the male trending factor m for each user profile in our data set. The female trending factor f is obtained by subtracting $f = 1 - m$.
2) Next, we divide the profiles into 5 disjoint subsets depending on the values of m and f of each profile. Profiles exhibiting extreme female features are classified as *strongly trending female*. Likewise, profiles exhibiting extreme male features are classified as *strongly trending male*. Profiles otherwise exhibiting above average female (male) features are classified as *weakly trending female (male)*. Profiles exhibiting neither female or male features are classified as *neutral*.
3) Finally, we compare the profiles (strongly or weakly) trending female or male with the self-declared gender in the corresponding Facebook profiles. We consider strongly trending profiles that conflict with the gender in the corresponding Faceboook profile to be *likely deceptive*. Weakly trending profiles similarly conflicting with the corresponding self-declared gender are defined to be *potentially deceptive*.

Through our analyses we have identified several thousands potentially deceptive and likely deceptive profiles. We have manually inspected likely deceptive profiles, as we report below, and found that a large proportion of those profiles (about 42.85%) were indeed deceptive. Manual inspection was inconclusive in an additional 7.8% of profiles, as those profiles were either deleted before we could inspected them or associated with multiple Twitter users (e.g., members of a club or an interest group) rather than individual users. We also manually inspected a statistically-significant randomized sample (about 5%) of the potentially deceptive profiles that we identified. We found that about 8.7% of these potentially deceptive profiles were indeed deceptive. We also found that many potentially deceptive profiles, about 19.6% of the total, had been deleted before we could examine them or belonged to groups of people. We conclude that our approach can provide reasonably accurate indications of gender deception.

Our main contributions are outlined below.

1) We defined a novel framework for detecting deception in user profiles using different profile's characteristics with inconsistent information (i.e., conflict indications). Our framework supports multiple approaches to deception detection. Here we report results with one specific approach.
2) We created a large data set of Twitter users, and we applied our approach to the data set in an effort to assess the performance of the approach.
3) We found that considering a combination of multiple profile's characteristics from each Twitter profile leads to a reasonable degree of accuracy for detecting deception.

The remainder of this paper is organized as follows. Section II gives some background information. In Section III, we briefly summarize related work on deception and gender classification. In Section IV, we describe our data-set collection including gender information. In Section V, we detail our method for gender classification. In Section VI, we report our empirical results. Finally, in Section VII, we give some conclusions and outline future work directions.

II. Background and Rationale

The key factor underlying the success of OSN-mediated communities, similar to traditional communities, is the trust that exists among community members [10], [33]. However, in OSNs, it is easy to provide false information in someone's profile in order to deceive others. In fact, lying in profiles and posts is apparently quite widespread. For instance, a survey by the eMedia Group showed that as many as 31% of OSN users had actually entered false information about themselves in OSNs to protect their privacy [14]. Another study, posted on the Pew Internet & American Life Project, showed that 56% of teenagers surveyed in the United States entered false information in their online profiles primarily "to keep themselves safe from unwanted online attention" [18]. More recently, a study by the Advertising Standards Authority shows that 42% of Internet users under the age of 13 reported

that they lie (i.e., provide false information) about their age in order to see products or content with age restrictions [1].

The above surveys on deception in OSNs make it more important for users and administrators of OSNs to be empowered with tools for automatically detecting false or misleading personal information posted in OSNs; however, tools of this kind are currently lacking. One reason for this state of affairs is that there are no reliable indicators for detecting deception; it is unclear which indicators will help and which will not help. In fact, deceiving people will sometimes use great efforts to disguise their deceit. For instance, a man posing as a woman may choose a false identity (e.g., profile's name and description) in order to make his false statements believable. We address these difficulties by considering a wide variety of features indicators for gender, such as several profile colors, in an effort to detect automatically situations of this kind.

This research seeks to detect automatically deception in the gender of a user's OSN profile. This line or work can benefit law enforcement who may use it as part of legal investigations. It will also help millions of OSN users develop authentic and healthy relationships with others. Our approach to detecting deception compares gender indicators obtained from different profile characteristics. Gender indicators are linearly weighted. We then compare the indicators obtained from each feature and we flag for potential deception users' profiles with conflicting indications.

There are several challenges to be considered in detecting deception in OSNs:

- There is no shared universal culture within the OSN community. Each country has its own culture that makes the detection of deception difficult. Age-related issues further complicate the detection of deception. For instance, older people communicate differently compared to teenagers. Therefore, different communication styles make it harder to evaluate profile characteristics correctly.
- Deceivers sometimes go to great lengths to disguise their deceit, for instance, by using fake identities.
- Deception has not been investigated to date. Thus, we do not know about reliable indicators that may exist for detecting deception.
- A wealth of information available in OSNs is text-based (e.g., OSN postings). However, analyzing text leads to high dimensional spaces and computational complexity. For example, there are around 15.4 millions features extracted from the texts for gender classification by Burger et al. [8]. A feature space of that size can cause scalability and efficiency issues.
- Most OSNs support multiple languages. Twitter, for instance, contains more than 70 languages [24]. Our challenge is to have a language independent algorithm that deals with massive data sets.
- Various companies offer services for providing fake followers. Thus, few users might buy services that create fake followers for them in order to increase their total follower count. Thereby, it is harder to distinguish between genuine and fake user profiles.
- To date the effects of cultural factors on the behavior of social-network users are not fully understood. Evidently, the subtle interactions between gender preferences and cultural factors complicate the task of gender identification. Other authors have began investigating the relationships between gender identity and cultural factors on social networks [5]; however, the effects of these factors on color preferences by gender are currently unclear.

We address the above challenges by considering a broad variety of indicators and by carefully defining the relative strength of those indicators through extensive experimentations with our data set.

III. Related Work

To our knowledge, ours is the first method for detecting deception with respect to gender in OSN profiles. Related work to ours falls into two categories, namely detection of fake OSN accounts and gender classification based on text. We discuss developments in these two fields in the next two subsections.

A. Detection of Fake Accounts

In recent years, the field of the deception has attracted many researchers such as Castelfranchi et al. [9]. Thomas et al. [30] address some issues behind the black market of Twitter accounts. A fake Twitter account is considered as one form of deception (i.e., deception in both the content and the personal information of the profiles as well as deception in having the profile follow others not because of personal interest but because they get paid to do so). Thomas et al. investigated Twitter to study, monitor and explore around 120,000 fraudulent accounts. Their work was unique in the area of the spamming in OSNs because they are working at the Twitter corporation where they have internal access privileges. They applied their approach to Twitter contents to distinguish spamming messages from authentic ones while our approach is to identify deceptive profiles (e.g., the person responsible for the information). Prior to this work, many researchers studied spam on different platforms (e.g., emails, OSNs and forums) [13], [15], [26], [29].

Due to the popularity of OSNs, including Twitter, many researchers have analyzed the behavior of profiles in OSNs. Castillo et al. [11] proposed an automatic method for assessing the credibility of Twitter contents. In addition, Yardi et al. [34] examined the differences between fake and legitimate Twitter users while Chu et al. [12] proposed a model to classify legitimate users, fake users and a combination of both in Twitter. Furthermore, Wang [32], Benevenuto et al. [6], McCord and Chuah [22], and Wang [31] investigated spam detection in Twitter using a content-based approach. However, due to limitations and the scope of the research, none of the previous researchers investigated and analyzed gender deception in OSNs. In fact, no research to date could answer the following question: Is a user's profile deceptive or trusted based on inconsistent information originating from the profile? In this research, we define a model for automatically detecting deception and flag it for further investigation.

B. Gender Classification

To our knowledge, the first work on gender classification using a data set extracted from OSNs (e.g., Twitter) is by Rao et al. [28]. They proposed a novel classification algorithm called stacked-SVM-based classification. Their approach depends on simple features such as n-grams, stylistic features, and some statistics on a user's profile. Another work on Twitter, by Pennacchiotti and Popescu [25], used a different set of features extracted from profile contents. These features are derived from an in-depth analysis of profile contents, such as content structure features, text content sentiment features, lexical features and explicit links pointing to outside sources. Other authors' works have investigated gender classification (i.e., inference). See, for instance, [2], [8], [19], [20], [23], [27]. Those works achieved different accuracy results depending on the method used. A general disadvantage is that those works use text-based characteristics for gender classification, resulting in an explosion in the resulting number of features (sometimes in the order of tens of millions of features.) In contrast with those methods, our approach uses only a few hundred features, resulting in low computational complexity and a high degree of scalability [3], [4].

IV. Data Set and Gender Information

Typically, in OSNs users create profiles describing their interests, activities and additional personal information. Then, users often start looking for friendships with other users who could be friends, family members, co-workers, classmates and even perfect strangers, who might happen to share common interests. We chose Twitter profiles as the starting point of our data collection for several reasons that were mentioned in our previous work [3], [4]. Twitter profiles consist of about 30 fields containing biographical and other personal information, such personal interests and hobbies. However, many fields in the form are optional, meaning that they are often left blank. In general, users choose their own preferences for many fields (e.g., name, username, description, colors) while editing their profiles.

Here we are specifically interested in the following seven fields from the profile of each Twitter user.

- Name.
- Username.
- Background color.
- Text color.
- Link color.
- Sidebar fill color.
- Sidebar border color.

We collected information about user profiles on Twitter by running our crawler between January and February 2014, subject to Twitter's limitation of less than 150 requests per hour. We started our crawler using the same technique as in our previous works, that is, by selecting a set of random profiles and adding any profile that the crawler encountered (e.g., profiles of users whose names were mentioned in tweets we harvested). Subsequently, we filtered out all the profiles missing a valid URL. The URL is a profile field that allows a Twitter user to create a link to a profile hosted by another OSN, such as Facebook. This field is important because profiles hosted by other OSNs often contain an explicit gender field, which Twitter profiles do not include [3], [4].

In total, we collected 194,292 profiles, of which 104,535 were classified as male and 89,757 were classified as female according to the self-declared gender field in the Facebook profile. We considered only profiles for which we obtained gender information independently of Twitter content (i.e., by following links to other profiles in Facebook). For each profile in the dataset, we collected the seven profile fields listed above. We also stratified the data by randomly sampling 174,600 profiles, of which 87,300 are classified as male and 87,300 are classified as female. In this manner, we obtain an even baseline containing 50% male and female profiles.

The main threat to the validity of this research is our reliance on self-declared gender information entered by Twitter users on external web sites for validation of our predictions. We believe that deceptive people sometimes do make mistakes by entering conflicting information in different OSNs. In this study we rely on gender information from external links posted by profile owners. We use this gender information as our ground truth. Evidently, a complete evaluation of 174,600 Twitter users would be impractical. However, we manually "spot checked" about 10,000 out of the profiles in our dataset or about 6.6% of the dataset. In the cases that we checked by hand, we are confident that the gender information we harvested automatically was indeed correct over 90% of the time. In the majority of the remaining cases we could not determine the accuracy of our ground truth.

V. Proposed Approach

Detecting deception involving the gender of OSN users is quite challenging. To date, there are no reliable indicators for detecting deception of this kind. Our research is aimed at detecting automatically deceptive profiles from profile characteristics in OSNs. We are specifically interested in detecting deception about user's gender by utilizing profile characteristics.

In general, there are multiple approaches for detecting deception in OSNs depending on how one uses information from profile characteristics. Here are some examples.

1) Detecting deception by comparing different characteristics for each user in a data set obtained from a single OSN (e.g., first names and colors in a given OSN).
2) Detecting deception by comparing characteristics from different OSNs (e.g, Twitter and Facebook) for the same user.
3) Detecting deception by comparing a combination of characteristics from a user's profile in a given OSN (e.g., first name, user name and colors in a Twitter profile) with a ground truth obtained from external source.

In the first case, one would compare gender characteristics obtained from each user and flag for potential deception profiles with conflicting indications. In the second case, one would

flag for potential deception users whose gender indications from different OSNs conflict with each other. In the third case, profiles whose characteristics conflict with the ground truth are flagged for potential deception.

Our framework for detecting deception supports all three approaches. We are in fact currently investigating all these approaches; however, here we focus only the third method. In the sequel we describe an implementation using a Bayesian classifier and we report on preliminary empirical results with the method. We also started investigating the second approach above; below we report data comparing the accuracy of gender predictions using first names from Twitter vs. Facebook. The first method above requires a broader set of characteristics than we have considered so far, including posted texts and user descriptions, which are language dependent. We are currently investigating those additional characteristics.

A. Detecting the Deception

Our approach to deception detection is based on our previous results on gender classification based on color features contained in Twitter profiles [4] and on first names and user names contained therein [3]. In brief, we analyzed user profiles with different classifiers in the Konstanz Information Miner (KNIME), which uses the Waikato Environment for Knowledge Analysis (WEKA) machine learning package [7], [17]. For profile colors, we obtained our best results when we considered the following five color features in combination: (1) profile background color, (2) text color, (3) link color, (4) sidebar fill color, and (5) sidebar border color. We also applied a color quantization and sorting procedure that not only improved the accuracy of our gender classification, but also reduced the size of our feature space by several orders of magnitude, to a few hundred features. We conducted extensive experimentations with various kinds of classifiers; the hybrid Naïve-Bayes Decision-Tree (NB-tree) classifier yielded the highest accuracy results. The interested reader is referred elsewhere for additional details [4].

We also defined gender classifiers based on first names and user names harvested from Twitter profiles [3]. In that case, we applied a phonetic analysis to first names and user names, transforming names into language-independent phoneme sequences. This transformation resulted in a substantial reduction in the feature space of our classifier with evident performance benefits. The empirical results we obtained with our various classifiers are reported in the next section below.

We compute the male trending factor m of each user profile in our data set with a Bayesian classifier that uses the following formula.

$$m = \frac{w_f \cdot s_f + w_u \cdot s_u + w_c \cdot s_c}{w_f + w_u + w_c} \quad (1)$$

In the above formula w_f, w_u and w_c denote the relative weight of the three gender indicators we consider, namely first names, user names and the 5 color characteristics combined. The weight of an indicator is given by the difference between the measured accuracy of that indicator, as a percentage, and the baseline value of 50%. Thus, if first names have an accuracy of 82%, the weight, w_f of the first name indicator is 32. Moreover, s_f, s_u and s_c indicate the sensitivity of a user's feature for a given indicator. For instance, the first name "Mary" has a high sensitivity, close to 1, for the female trending index, and a low sensitivity, close to 0, for the male trending index. We assign sensitivity values depending on the proportion of female vs. male users who have the given feature. Thus, the female and male sensitivity for a given value complement each other with respect to the unit value. Evidently, the male trending index computed with Equation (1) and the female trending index computed by the corresponding formula for f are also complementary with respect to one. The average value of the male trending index over our stratified data set is $\mu = 50.13$ with a standard deviation $\sigma = 18.87$. These are encouraging numbers. The average falls quite close to the middle of the range for m, that is, between 0 and 1 (as a percentage). Also, the standard deviation is sufficiently high in order for m to be a significant factor in distinguishing male from female profiles.

TABLE I
ACCURACY RESULTS IN GENDER PREDICTIONS OBTAINED BY USING DIFFERENT PROFILE CHARACTERISTICS FROM TWITTER PROFILES.

Characteristics	First names	User names	Colors	All
Accuracy	82%	70%	75%	85%

After computing the male trending index for each profile in our data set, we divide the profiles in the data set into 5 groups depending on the computed male index m. We define profiles with m values falling in the range $0 \leq m \leq \mu - 2\sigma$ as strongly trending female. Profiles whose m value falls in the range $\mu - 2\sigma < m \leq \mu - \sigma$ are classified as weakly trending female. Conversely, we classify profiles with m values falling in the range $\mu + 2\sigma \leq m \leq 1$ as strongly trending male. Profiles whose m value falls in the range $\mu + \sigma \leq m < \mu + 2\sigma$ are classified as weakly trending male. The remaining profiles are not deemed trending either way (neutral profiles).

Last, we compare user profiles trending male or female with the ground truth harvested from Facebook profiles. Profiles of strongly trending users whose computed trend conflicts with the corresponding ground truth are flagged for likely deception. Profiles of weakly trending users whose computed trend conflicts with the corresponding ground truth are flagged for potential deception. Note that our analysis is inconclusive in the case of users whose computed m value differs from average μ by less than the standard deviation σ. In this case, our analysis is inconclusive. We plan to explore alternative approaches to deception detection within our framework in order to include these users in our analyses.

VI. EMPIRICAL RESULTS

Here we report the results of the empirical studies on our data set. We first report our current results in the identification of deceptive profiles contained in our data set. We generated these results by linearly weighing gender indicators obtained from different Twitter profile characteristics and by

TABLE II
BREAKDOWN OF TWITTER USER PROFILES BY GENDER TRENDING FACTORS WITH DECEPTION PREDICTIONS.

	Strong female	Weak female	Neutral	Weak male	Strong male
Index range	$0 \leq m \leq 12.3$	$12.3 < m \leq 31.1$	$31.1 < m \leq 68.9$	$68.9 < m \leq 87.7$	$87.7 < m \leq 1$
Number of profiles	2,673	30,493	109,562	30,717	1,155
Potentially deceptive	—	2,677	—	3,779	—
Likely deceptive	59	—	—	—	18

comparing the resulting male trending factors with the self-declared genders in the corresponding Facebook profiles. Next, we report preliminary results on comparing the same type of characteristic (i.e., first names) from two different OSNs (Facebook vs. Twitter).

A. Empirical evaluation of feature relevance in Twitter

We first report the accuracy of gender predictions obtained with the three kinds of profile characteristics that we considered so far for Twitter users, namely first name, user name, and profile colors. Table I shows a summary of overall accuracy results obtained by applying the the NB-tree classification algorithm in the KNIME machine learning package to our entire data set. Table entries show the overall percentage of user profiles whose gender was predicted correctly using the characteristics under consideration. In particular, Column 2 reports accuracy results obtained with first names alone; Column 3 reports accuracy results obtained with user names alone; Column 4 reports accuracy results obtained with the combination of five profile colors we studied; and Column 5 reports accuracy results obtained when applying all characteristics (i.e., first names, user names, and colors) in combination. As explained above, we preprocessed first names and user names using our phoneme-based method [3]. Although accuracy results vary depending on the characteristics being used, the data in Table I show significant improvements over the 50% baseline for all the characteristics, which is quite encouraging.

Table II reports the size of the five subsets of our Twitter profiles resulting from partitioning based on the computed male trending factor m of each user. Recall that the average and standard deviation of m over our entire data set are $\mu = 50.13$ and $\sigma = 18.87$ respectively. Table columns report data for Twitter profiles classified as strongly trending female, weakly trending female, neutral, weakly trending male, and strongly trending male. The rows give the following information for each group of profiles: (1) the ranges of m values, (2) the total number of profiles in each group, (3) the number of potentially deceptive profiles among weakly trending profiles, and (4) the number of likely deceptive profiles among the strongly trending profiles. Groups are defined according to the standard deviation formula given earlier. The values of m are determined according to Equation (1) above.

Table II shows that there are 59 (18) likely deceptive profiles among strongly trending female (male) profiles. Also, we have 2,677 (3,779) potentially deceptive profiles among weakly trending female (male) profiles. We are currently evaluating the accuracy of these predictions, starting from the strongly trending profiles.

We were able to determine that 28 of the 59 strongly trending female profiles declaring a male gender indication on Facebook in fact belonged to female users by a manual inspection of those profiles. For the remaining 31 profiles, we were either unable to determine the user's gender by a visual examination of the profiles in question, or we determined that those profiles in fact belonged to male users, as declared in Facebook. Likewise, for the 18 strongly-trending male profiles declaring a female gender, we were able to determine that 5 profiles indeed belonged to male users, with 11 profiles belonging to female users. We were unable to determine the gender of the remaining two profiles.

We performed our manual inspections of a profile by examining data contained in the profile, by reading posted tweets, and by examining posted pictures. Whenever possible, we also followed any available links to other social network pages for a given profile user and repeated our inspections on those pages. Throughout these examinations, we sought to identify the profile owner and the owner's gender. When we could identify both the owner and the owner's gender beyond reasonable doubt, we used this gender as a our ground truth for the given profile.

We manually inspected a randomized sample of the potentially deceptive profiles in order to verify the accuracy of our predictions in this case. We specifically examined 133 weakly trending female profiles and 188 weakly trending male profiles, or about 5% of each group. We found that 17 of 133 female-trending potentially deceptive profiles were indeed deceptive (i.e., female users declaring to be male). We also found that 24 of these 133 profiles had been deleted or belonged to groups of people. Out of the 188 weak-male, potentially deceptive profiles, we found 11 profiles to be clearly deceptive, while a further 39 profiles had been deleted or belonged to groups of people. On the whole, we found that about 8.7% of potentially deceptive profiles that we examined were indeed deceptive. We also found that many more potentially deceptive profiles, about 19.6% of the total, had been deleted before we could examine them or belonged to groups of people.

Finally, we conducted a longitudinal study on first names of potentially deceptive profiles in our data set. A surprisingly high number of such profiles showed a name change. In particular, 892 of the 2,677 weak female, potentially deceptive profiles showed a name change between the time of our data set collection (January and February 2014) and this writing (April 2014). In 399 cases, the two first names in question

were fully incompatible with each other (i.e., the two names were not a nickname or short version of one another.) This is indicative of deception on a user's first name contained in Twitter profiles; at least one of the original name or the new name must have been incorrect for 399 of 2,677 profiles or 25.6% of these profiles. Likewise, we found that 968 of 3,779 weak-male, potentially deceptive profiles showed a name change, with inconsistent names in 491 cases, or 13.0% of the total. We are clearly encouraged by these results.

B. Comparing first names in different OSNs

Now we report on empirical comparisons of first names extracted from two different OSNs, namely Twitter and Facebook. Our goal is to determine which of the two indicators is a more reliable predictor of gender for the same user when used independently of other characteristics. Recall that some Twitter profiles contain a link to a Facebook page for the same user. In fact, our data set contains only profiles in which this link is present. Thus, we ran the Support Vector Machine (SVM) classifier on our all of our stratified data set, consisting of 174,600 profiles with a 50% male and female breakdown. No characteristics in addition to first names were included in these experiments.

We noted a significant difference in the reliability of first names from Facebook vs. Twitter as gender predictors. In particular, we report an accuracy of 87% for Facebook names, and an accuracy of 75% only for Twitter names. This result seems to indicate that the greater degree of structure and formality imposed by a Facebook profile with respect to a Twitter profile has resulted in a higher degree of trustworthiness for the former profiles than the latter profiles. For instance, a Facebook profile includes a gender field, first-name field, last-name field and a nickname field. A Twitter profile has a single field for a user's full name. We speculate that the ability for a user to define a nickname in Facebook may induce users to report their true first names in the first-name field, whereas Twitter users may be tempted to casually report their nicknames in the full name field of their Twitter profiles. We plan to investigate the discrepancy in accuracy results further in order to validate or disprove our conjecture.

Previously we defined a phoneme-based method for enhancing the reliability of first names and usernames as predictors of gender [3]. We also applied this technique to Facebook names and Twitter names. When this technique is used, our accuracy results improve to 91% for Facebook first names and to 82% for Twitter names, as reported in Table I. These results further confirm the greater accuracy of Facebook names as gender predictors with respect to first names extracted from Twitter. The interested reader is referred elsewhere for details on our phoneme-based method [3].

VII. Conclusions and Future Work

We presented a framework for detecting deception about gender information in online social networks and we reported preliminary empirical results with a strategy for attaining this goal within the framework. Our current results show considerable promise for our framework.

There are two main threats to the validity of this study. First, as we noted earlier some social network users misrepresenting their gender may go to great lengths in hiding their deceit. Thus, these users may deliberately choose profile colors and user names that are stereotypical of their purported gender. Also, these users may misrepresent their gender in their Facebook profiles, which we use as a ground truth. We plan to address this issue by expanding our notion of ground truth to include additional sources of information reachable by web crawling. Second, we have not explored how cultural preferences affect our predictions. For instance, a certain profile color could be associated with different gender identities in different cultures. We plan to address this issue by breaking down our data set based on cultural and ethnical factors and by repeating our analyses on each subset we consider.

In the future we also will continue exploring alternative strategies in an effort to further improve the accuracy of our predictions. We specifically plan to address detection of deception in neutral profiles, that is, profiles that fall within a standard deviation of the average female and male trending factors. For instance, we may reduce the range of trending factors for these profiles (i.e., to one half of the standard deviation from the average). We will also consider additional features, such as the genders of Twitter friends and followers, as part of gender predictions. We will explore text-based features factors, such as user postings, and we will include these features if their advantages outweigh their cost in terms of language dependence and increased computational complexit. Finally, we plan to explore the first two approaches supported by our framework, as we outlined in Section V above.

References

[1] Advertising Standards Authority, "Children and advertising on social media websites," http://www.asa.org.uk/News-resources/Media-Centre/2013/ASA-research-shows-children-are-registering-on-social-media-under-false-ages.aspx.

[2] F. Al Zamal, W. Liu, and D. Ruths, "Homophily and latent attribute inference: Inferring latent attributes of Twitter users from neighbors." in *6th International AAAI Conference on Weblogs and Social Media (ICWSM'12)*, pp. 387–390, Dublin, Ireland, 2012.

[3] J. Alowibdi, U. Buy, and P. Yu, "Empirical evaluation of profile characteristics gender classification on Twitter," in *12th International Conference onMachine Learning and Applications (ICMLA)*, pp. 365–369, Miami, Florida, Dec. 2013. IEEE, 2013.

[4] J. Alowibdi, U. Buy, and P. Yu, "Language independent gender classification on Twitter," in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM'13*, pp. 739–743, Niagara Falls, Ontario, Aug. 2013.

[5] D. Bamman, J. Eisenstein, and T. Schnoebelen, "Gender identity and lexical variation in social media," in *Journal of Sociolinguistics,* Vol. 18(2), pp. 135–160, April 2014.

[6] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on Twitter," in *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, vol. 6, Redmond, Washington, July 2010.

[7] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, K. Thiel, and B. Wiswedel, "KNIME–The Konstanz information miner: version 2.0 and beyond," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 26–31, 2009.

[8] J. D. Burger, J. Henderson, G. Kim, and G. Zarrella, "Discriminating gender on Twitter," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK.: Association for Computational Linguistics, July 2011, pp. 1301–1309. [Online]. Available: http://www.aclweb.org/anthology/D11-1120

[9] C. Castelfranchi and Y.-H. Tan, "The role of trust and deception in virtual societies," in *Proc. 34th Annual Hawaii Int. Conf. on System Sciences,* pp. 8, Wailea, Hawaii, Jan. 2001.

[10] C. Castelfranchi and Y.-H. Tan, *Trust and deception in virtual societies*, Kluwer Academic Publishers, 2001.

[11] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on Twitter," in *Proceedings of the 20th International Conference on the World Wide Web*. ACM, 2011, pp. 675–684.

[12] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Detecting automation of Twitter accounts: Are you a human, bot, or cyborg?" *IEEE Transactions on Dependable and Secure Computing*, pp. 811–824, 2012.

[13] E. Damiani, S. De Capitani di Vimercati, S. Paraboschi, and P. Samarati, "P2p-based collaborative spam detection and filtering," in *Peer-to-Peer Computing, 2004. Proceedings. Proceedings. Fourth International Conference on*. IEEE, 2004, pp. 176–183.

[14] eMedia, "Social networking sites: Almost two thirds of users enter false information to protect identity," http://www.realwire.com/releases/social-networking-sites-almost-two-thirds-of-users-enter-false-information-to-protect-identity.

[15] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao, "Detecting and characterizing social spam campaigns," in *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*. ACM, 2010, pp. 35–47.

[16] L. K. Guerrero, P. A. Andersen, and W. A. Afifi, *Close encounters: Communication in relationships*. Sage, 2013.

[17] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[18] A. Lenhart and M. Madden, "Teens, Privacy & Online Social Networks," http://www.pewinternet.org/Reports/2007/Teens-Privacy-and-Online-Social-Networks.aspx.

[19] W. Liu and D. Ruths, "Whats in a name? using first names as features for gender inference in Twitter," in *2013 AAAI Spring Symposium Series, In Symposium on Analyzing Microtext*, 2013.

[20] W. Liu, F. Al Zamal, and D. Ruths, "Using social media to infer gender composition of commuter populations," in *Proceedings of the When the City Meets the Citizen Workshop, the International Conference on Weblogs and Social Media*, 2012.

[21] McAfee, "Mcafee digital deception study 2013: Exploring the online disconnect between parents & pre-teens, teens and young adults," http://www.mcafee.com/us/resources/reports/rp-digital-deception-survey.pdf.

[22] M. McCord and M. Chuah, "Spam detection on Twitter using traditional classifiers," in *Autonomic and Trusted Computing*. Springer, 2011, pp. 175–186.

[23] A. Mislove, S. L. Jørgensen, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist, "Understanding the demographics of Twitter users," in *5th International AAAI Conference on Weblogs and Social Media (ICWSM'11)*, 2011, pp. 554–557.

[24] D. Mocanu, A. Baronchelli, N. Perra, B. Gonçalves, Q. Zhang, and A. Vespignani, "The Twitter of babel: Mapping world languages through microblogging platforms," *PloS one*, vol. 8, no. 4, p. e61981, 2013.

[25] M. Pennacchiotti and A.-M. Popescu, "A machine learning approach to Twitter user classification." in *proceedings of the International Conference on Weblogs and Social Media*, 2011.

[26] A. Ramachandran, N. Feamster, and S. Vempala, "Filtering spam with behavioral blacklisting," in *Proceedings of the 14th ACM conference on Computer and communications security*. ACM, 2007, pp. 342–351.

[27] D. Rao, M. J. Paul, C. Fink, D. Yarowsky, T. Oates, and G. Coppersmith, "Hierarchical bayesian models for latent attribute detection in social media." in *5th International AAAI Conference on Weblogs and Social Media (ICWSM'11)*, 2011.

[28] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, "Classifying latent user attributes in Twitter," in *Proceedings of the 2nd international workshop on Search and mining user-generated contents*. ACM, 2010, pp. 37–44.

[29] Y. Shin, M. Gupta, and S. Myers, "Prevalence and mitigation of forum spamming," in *INFOCOM, 2011 Proceedings IEEE*. IEEE, 2011, pp. 2309–2317.

[30] K. Thomas, D. McCoy, C. Grier, A. Kolcz, and V. Paxson, "Trafficking fraudulent accounts: The role of the underground market in Twitter spam and abuse," in *USENIX Security Symposium*, 2013.

[31] A. H. Wang, "Machine learning for the detection of spam in Twitter networks," in *e-Business and Telecommunications*. Springer, 2012, pp. 319–333.

[32] A. H. Wang, "Don't follow me: Spam detection in Twitter," in *Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on*. IEEE, 2010, pp. 1–10.

[33] D. Ward and H. Hexmoor, "Towards deception in agents," in *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*. ACM, 2003, pp. 1154–1155.

[34] S. Yardi, D. Romero, G. Schoenebeck, *et al.*, "Detecting spam in a Twitter network," *First Monday*, vol. 15, no. 1, 2009.