Soc. Netw. Anal. Min. (2015) 5:32
DOI 10.1007/s13278-015-0273-1

ORIGINAL ARTICLE

Deception detection in Twitter

Jalal S. Alowibdi[1] · Ugo A. Buy[2] · Philip S. Yu[2] · Sohaib Ghani[3] · Mohamed Mokbel[3]

Received: 27 November 2014 / Revised: 25 April 2015 / Accepted: 16 June 2015 / Published online: 30 June 2015

Abstract Online Social Networks (OSNs) play a significant role in the daily life of hundreds of millions of people. However, many user profiles in OSNs contain deceptive information. Existing studies have shown that lying in OSNs is quite widespread, often for protecting a user's privacy. In this paper, we propose a novel approach for detecting deceptive profiles in OSNs. We specifically define a set of analysis methods for detecting deceptive information about user genders and locations in Twitter. First, we collected a large dataset of Twitter profiles and tweets. Next, we defined methods for gender guessing from Twitter profile colors and names. Subsequently, we apply Bayesian classification and K-means clustering algorithms to Twitter profile characteristics (e.g., profile layout colors, first names, user names, and spatiotemporal information) and geolocations to analyze the user behavior. We establish the overall accuracy of each indicator through extensive experimentation with our crawled dataset. Based on the outcomes of our approach, we are able to detect deceptive profiles about gender and location with a reasonable accuracy.

Keywords Deception detection · Gender classification · Profile indicators · Profile characteristics · Profile classification · Location classification · Twitter

✉ Jalal S. Alowibdi
jalal.alowibdi@gmail.com; jalowibdi@uj.edu.sa; jalowibd@cs.uic.edu

Ugo A. Buy
buy@cs.uic.edu

Philip S. Yu
psyu@cs.uic.edu

Sohaib Ghani
sghani@gistic.org

Mohamed Mokbel
mokbel@gistic.org

[1] Faculty of Computing and Information Technology, University of Jeddah, Jeddah, Saudi Arabia

[2] Department of Computer Science, University of Illinois at Chicago, Chicago, USA

[3] KACST GIS Technology Innovation Center, Umm Al-Qura University, Makkah, Saudi Arabia

1 Introduction

Online Social Networks (OSNs) are part of the daily life of hundreds of millions of people. However, many user profiles in OSNs contain misleading, inconsistent, or false information. Existing studies have shown that lying in OSNs is 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 us to be confident about having a healthy and trusted relationships in OSNs. Trust is an important factor in OSNs. However, information posted in OSNs is often not trusted because lying is so widespread. Although privacy issues in OSNs have attracted a considerable attention in recent years, yet currently there is no work on detecting deception in gender and location information posted in OSNs.

The long-term objective of this research is to automatically flag deceptive information in user profiles and posts, based on detecting inconsistencies in those profiles and posts. Here we focus specifically on the detection of inconsistent information involving user gender and the detection of conflicting spatiotemporal information—information about space and time—involving user location. In the sequel, we separately discuss our two approaches for

detecting deception about gender and location. We collected two distinct Twitter datasets and applied different analysis methods to the two datasets.

We applied the following paradigm for detecting deceptive information about gender.

1. We collected a dataset consisting of about 174,600 Twitter profiles by running a crawler on Twitter's programmable interfaces between January and February 2014. We were specifically interested in the following features for each Twitter user profile: (1) number of colors chosen by Twitter users for their profile, (2) the user name, and (3) the user's first name. We selected profiles containing an external link to a Facebook page specifying the gender of the Twitter user.
2. We applied a number of preprocessing methods to colors and names features of Twitter profiles. Profile preprocessing significantly improved our ability to predict the gender of a Twitter users from the collected features.
3. We independently established the accuracy for each feature (i.e., profile colors, first names, and user names) for predicting the gender of a Twitter user by conducting extensive experimentation with Twitter profiles.
4. We defined a Bayesian classifier seeking to identify Twitter users whose profile characteristics conflict with the self-declared gender information collected from Facebook. We identified several thousands profiles as being *potentially deceptive* and a smaller subset of profiles as being *likely deceptive*.
5. We manually checked the profiles and postings of Twitter users that the Bayesian classifier had identified as being potentially deceptive.

The outcome of these studies is that the first name, user name, and background color chosen by a user for his/her profile can provide reasonably accurate predictions of the user's gender. In addition, these characteristics can also help finding deceptive information. We specifically identified 4 % of the 174,600 profiles analyzed as *potentially deceptive*. Manual inspection was inconclusive in an additional 7.8 % of profiles, as those profiles were either deleted before we could manually inspect 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 *potentially deceptive* profiles that we identified. We found that about 8.7 % of these potentially deceptive profiles were indeed likely 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. In addition, there were 77 profiles of the 174,600 profiles analyzed as *likely deceptive*. We manually inspected these likely deceptive profiles and found that a large proportion of those profiles (about 42.85 %) were indeed deceptive.

Furthermore, we applied the following paradigm for detecting deceptive information about location.

1. We collected a dataset consisting of about 35,000 Twitter profiles by running a crawler on Twitter's programmable interfaces between March and April 2014. We were specifically interested in the following features for each Twitter user profile: (1) temporal information and (2) spatial information.
2. We validated our findings by comparing them with information about travel destinations of Saudi residents posted by the Saudi Tourist Information and Research Centre.
3. We independently established the accuracy for each feature by predicting the location of a Twitter user by conducting extensive experimentation with Twitter profiles.
4. We defined a Bayesian classifier seeking to identify Twitter users whose profile tweets characteristics contain conflicting information. We identified several thousands profiles as being *potentially deceptive* and as being *likely deceptive*.
5. We manually checked the profiles and postings of Twitter users that the Bayesian classifier had identified as being potentially deceptive.

To detect deception about user location, we conducted a spatiotemporal analysis of postings (i.e., tweets) containing geo-tagged information (i.e., latitude and longitude of the client from which a tweet originated). We used publicly available Twitter data of that period to find out where the people spent their vacation for a particular country, Saudi Arabia, and a particular holiday (Spring break, 2014). The outcome of this study is that analysis of spatiotemporal information extracted from tweets can provide reasonably accurate predictions of the users' locations accuracy. We specifically identified 5 % of the 35,000 profiles in the dataset as potentially deceptive profiles. We manually inspected potentially deceptive profiles and found that a large proportion of those profiles (about 35.0 %) were indeed deceptive. We also manually inspected a statistically significant sample of the likely deceptive profiles that we identified. We found, in some cases, that about 90.0 % of the identified potentially deceptive profiles were indeed likely deceptive. We conclude that our approach can provide reasonably accurate predictions of gender and location feature-based deception.

On the whole, our preliminary results with our datasets are very encouraging. We can identify deceptive information about gender and location with reasonable accuracy. In

addition, our methods use a relatively modest number of profile characteristics and spatiotemporal features, resulting in a low-dimensional feature space. We have deliberately excluded any other profile characteristics, such as posted texts (tweets), because our approach combines a good accuracy and language independence with low computational complexity.

Our main contributions are outlined below.

1. We defined a novel framework for detecting deception in user profiles using different profile characteristics with inconsistent information (i.e., conflict indications). Our framework supports multiple approaches for deception detection.
2. We created a large dataset of Twitter users, and we applied our approaches to the dataset in an effort to assess the performance of the approaches.
3. We applied novel preprocessing methods to our datasets to enhance the accuracy of our gender predictions.
4. We found that considering a combination of multiple profile's characteristics from each Twitter profile leads to a reasonable degree of accuracy for detecting the deception about gender and location.
5. We defined methods for identifying Twitter users containing deceptive information about gender and location.
6. Although we discuss the deception about gender before Alowibdi et al. (2014), here in this research, we added one more novel technique about deception which is the location-based approach in detecting the deception.

The remainder of this paper is organized as follows. Section 2 gives background information and summary of related work about deception, gender classification, and location classification. In Sect. 3, we explain the motivation behind the work. In Sects. 4 and 5, we extensively describe the deceptive profiles about gender and location and also we report our empirical results. Finally, in Sect. 6, we give some conclusions and outline future work directions.

2 Related work

In this section, we are going to cite the related work in detecting the deception of users profiles. We are specifically interested in detecting deception about user's geo-location and gender classification by utilizing spatiotemporal activities and posts. Related work to ours falls into three categories, namely detection of deception in different fields, deception based on location information of user's spatiotemporal activity, and deception based on gender classification.

2.1 Deception

The field of the deception has recently attracted many researchers. Alowibdi et al. (2014) proposed a novel approach to detect the deceptive profiles utilizing inconsistent information about the gender. They compared different gender indicators in order to find deceptive profiles. Then, they applied statistical algorithms to find inconsistent information about the gender. After that, they flag for potential deceptive profiles. Here, in this paper, we extend that approach to find deceptive profiles using location-based approach.

Currently, beside Alowibdi et al. (2014), there is another research close to ours, which has been investigated by Thomas et al. (2013). They investigated about 120,000 Twitter profiles to explore how fake profiles generally behave. Also, there are many other researchers investigating the behavior of profiles in OSNs such as Castillo et al. (2011) and Yardi et al. (2009). Most of these works investigated spamming which is totally different than the field of deception. On the other hand, our work defines a model for automatically detecting deception and flag it for further investigation.

In addition, many other researchers generally investigated the deception in various applications such as chat, email, and opinion applications (Castelfranchi and Tan 2001; Warkentin et al. 2010; Caspi and Gorsky 2006; Hancock et al. 2004; Newman et al. 2003). These researchers proposed linguistic feature classification approach of text. Unlike most existing approaches, our approach is different in term of its originality, simplicity, and targeted platform (e.g., OSNs) and novelty. Our approach involved detecting deceptive profile with unreasonable and suspicious geo-location activities in OSN profiles.

2.2 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. (2010). 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 (2011), 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. There are various other works as well that have investigated gender classification (Al Zamal et al. 2012; Burger et al. 2011; Liu and Ruths 2013; Liu et al. 2012; Mislove et al. 2011; Rao et al. 2011). These works achieved different accuracy

results depending on the method used. A general disadvantage with these works is they use text-based characteristics for gender classification, resulting in an explosion in the resulting number of features (in the order of tens of millions of features). In contrast with these methods, our approach uses only a few hundred features, resulting in low computational complexity and a high degree of scalability (Alowibdi et al. 2013a, b).

2.3 Location classification

There are many works for location-based classification using a dataset extracted from OSNs (e.g., Twitter) such as Cheng et al. (2010), Jurgens (2013), and Sakaki et al. (2010). Their approach utilize classification algorithms and machine learning techniques using many profile features. These features are derived from an in-depth analysis of profile contents, such as geo-location, content structure features, and explicit links pointing to outside sources. A general advantage is that these works can be implemented in our approach for geo-location classification, but with different goal which is to find inconsistent information that lead to detect deceptive profiles. In contrast with these methods, our approach to detect deceptive profiles first use spatiotemporal classification and then apply statistical methods to find unreasonable geo-location activities resulting in low computational complexity and a high degree of scalability similar to the approach in Alowibdi et al. (2014).

3 Motivation

Lying in OSNs is apparently quite widespread. In OSNs, people lie for different reasons by posting information that is not actually true about themselves. For example, children may lie because they want to register for an OSN with age restrictions. Adults may lie because they want to attract others attention. According to one study, as many as 31 % of users in OSNs provide false information to be safe online (RealWire.com 2007). Also, in another study, only 20 % of people surveyed declared to be honest about the information they provide online (Turner 2010). According to yet another study, 56 % of teenager provided false information in their profiles in order to protect themselves from undesirable attention (Lenhart and Madden 2007). As many as 42 % of children under the age of 13 reported that they lie about their age in order to be able to see content with age restrictions (Authority 2013). The interested reader is referred elsewhere for additional detail about the different forms of deception (Guerrero et al. 2012). Here, we define deception as providing false information about one's own gender or location, regardless of the reasons for providing such information.

The above surveys on deception in OSNs make it 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. Deceitful people will sometimes use great efforts to disguise their deceit. Thus, the problem of detecting the deception is important, but extremely challenging and worthy of attention. To our knowledge, there is no previous work on detecting the deception based on finding conflicting information in a user's profile on an OSN. This work, which extends our previous results (Alowibdi et al. 2014), is part of a long-standing project aimed at enhancing the trust among members of OSNs.

4 Detect deception about gender

In this section, we explain in detail our method for detecting deception about the gender of a Twitter user and we report empirical results for detecting the corresponding profiles.

4.1 Background

The foundation of our approach for detecting deception about gender was previous works in gender classification (Alowibdi et al. 2013a, b). We sought to identify a Twitter user's gender based on the user's profile characteristics independently from a ground truth. In those reports, we studied three kinds of profile characteristics, namely profile layout colors, names, and user names. We preprocessed profile colors with a novel color quantization (i.e., normalization) method and we applied phoneme-based preprocessing to the profile names and user names. Thanks in part to our preprocessing methods, we obtained good accuracy classification results with low computational complexity and high scalability as shown in Table 1.

4.2 Dataset collection

Typically, in OSNs users create profiles describing their interests, activities, and additional personal information. Thus, we chose Twitter profiles as the starting point of our

Table 1 Accuracy results in deceptive profiles about gender obtained by comparing inconsistent information of different profile characteristics from Twitter profiles

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

data collection for several reasons that were mentioned in our previous work (Alowibdi et al. 2013a, b). In general, users choose their own preferences for many fields (e.g., name, username, description, and colors) while editing their profiles. Here, we are specifically interested in the following seven fields from the profile of each Twitter user, namely, name, username, background color, text color, link color, and sidebar fill color.

We collected information about user profiles on Twitter by running our crawler between January and February 2014. 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.

4.3 Dataset collection validation

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 of the profiles in our dataset that is about 6.6 % of the dataset. In the cases that we checked by hand, we are confident that the gender information we collected 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.

4.4 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 dataset 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; however, in this research, we focused on 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. The second method requires access to other OSNs than Twitters, which is much more difficult to obtain.

4.4.1 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 and on first names and user names contained therein. 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 (Berthold et al. 2009; Hall et al. 2008).

Consequently, for profile colors, we obtained our best results when we considered the following five color-based features in combination: (1) profile background color, (2) text color, (3) link color, (4) sidebar fill color, and (5) sidebar border color. We employed two preprocessing stages in order to enhance the accuracy of our gender predictions using profile colors. First, we apply *color clustering* whereby we reduce the representation of profile colors from the traditional 8-bit RGB representation to a 5-bit RGB representation, by discarding the three least significant bits from each of the red, green, and blue values.

The traditional 8-bit RGB representation yields a feature set consisting of $2^{8\times3} = 2^{24}$ or about 16 Million colors. A feature set of this size would be mostly unnecessary as most colors are perceptually indistinguishable from neighboring colors with R, G, and B values differing only by few units from the original color. Thus, we chose to cluster colors in such a way that colors with a given cluster are perceptually similar to each other. In this manner, we reduce the total size of our color set to $2^{5\times3} = 2^{15}$ or about 32 thousand colors. The advantage is that we obtain a statistically significant number of profile users in each color cluster. The second preprocessing stage is a *color sorting* technique by which we arrange colors according to their hue. In this manner, we create a sequence in which similar colors are close to one another.

We compared empirically the performance of gender predictions using raw colors and colors obtained by applying clustering and sorting. In general, the accuracy of our gender predictions improved from 65 to 74 % when applying the two preprocessing stages.

With respect gender predictions using first names and user names, we applied a phoneme-based preprocessing stage. In brief, we first transformed names in a variety of alphabets to Latin characters used in the English alphabet by applying the Google Input Tool (GIT) to the first names and user names we had collected. GIT converts the alphabet of different languages than English (e.g., Japanese, Chinese, and Arabic) to characters in English. Next, we transform English-alphabet names into phoneme sequences. A phoneme is the smallest set of a language's phonology. For example, John can be represented as the 3-phoneme sequence "JH AA N," while Mary can be represented as "M EH R IY." We use a phoneme set from Carnegie Mellon University that contains exactly 40 phonemes (Speech at CMU 2013). Each phoneme may carry three different lexical stresses, namely no stress, primary stress, and secondary stress. This transformation resulted in a substantial reduction in the feature space of our classifier with evident performance benefits. In general, our accuracy has improved from about 71–82.5 % because of this preprocessing stage. We are quite encouraged that not only we improved the accuracy of our gender predictions, but we also discovered a world-wide trend whereby similar sounding names are associated with the same gender across language, cultural, and ethnic barriers. We tried both finer and coarser representations for names, and we found that phonemes give us the best prediction accuracy among the options that we considered, along with a dramatic reduction in the size of our feature spaces.

In particular, 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 1 shows a summary of overall accuracy results obtained by applying the NB-tree classification algorithm in the KNIME machine learning package to our entire dataset. 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 of 82 % obtained with first names alone; Column 3 reports accuracy results of 70 % obtained with user names alone; Column 4 reports accuracy results of 75 % obtained with the combination of five profile colors we studied; and Column 5 reports accuracy results of 85 % 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 (Alowibdi et al. 2013a). Although accuracy results vary depending on the characteristics being used, the data in Table 1 show significant improvements over the 50 % baseline for all the characteristics, which are quite encouraging.

We compute the male trending factor m of each user profile in our dataset 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 five 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 Eq. (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 $\mu = 0.5013$ with a standard deviation $\sigma = 0.1887$. 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.

After computing the male trending index for each profile in our dataset, we divide the profiles in the dataset 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 collected 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. We plan to explore alternative approaches to deception detection within our framework in order to include these users in our analysis.

4.5 Empirical results

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

4.5.1 Empirical evaluation of feature relevance in Twitter

Table 2 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 dataset are $\mu = 0.5013$ and $\sigma = 0.1887$, 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 Eq. (1) above.

Table 2 shows that there are 59 (18) likely deceptive profiles among strongly trending female (male) profiles. Also, we have 2677 (3779) potentially deceptive profiles among weakly trending female (male) 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 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

Table 2 Accuracy results in gender predictions obtained using different profile characteristics from Twitter profiles

	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	2673	30,493	109,562	30,717	1155
Potentially deceptive	–	2677	–	3779	–
Likely deceptive	59	–	–	–	18

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 dataset. A surprisingly high number of such profiles showed a name change. In particular, 892, about 33.3 %, of the 2,677 weak female, potentially deceptive profiles showed a name change between the time of our dataset collection (January and February 2014) and this writing (September 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.

4.5.2 *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 dataset contains only profiles in which this link is present. Thus, we ran the Support Vector Machine (SVM) classifier on our all of our stratified dataset, 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.

Previously we defined a phoneme-based method for enhancing the reliability of first names and usernames as predictors of gender (Alowibdi et al. 2013a). 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 1. These results further confirm the greater accuracy of Facebook names as gender predictors with respect to first names extracted from Twitter.

4.5.3 *Evaluation of predictions by multiple blind review*

We further evaluated the accuracy of our predictions on gender deception by a multiple blind review of a statistically significant sample of potentially deceptive profiles. We used the following procedure. First, we randomly selected 400 potentially deceptive profiles, with a 50 % male and female breakdown, from our dataset. These profiles cover approximately 10 % of all potentially deceptive profiles in our dataset, excluding profiles that were deleted between the time the profiles were collected and the time we evaluated the profiles. As we mentioned earlier, about 19 % of potentially deceptive profiles in our dataset were in fact deleted before we could analyze them manually.

Second, we asked 5 evaluators to determine the gender of the profile holders for each of the 400 potentially deceptive profiles. Each evaluator was instructed to follow a sequence of examination steps. First, each evaluator was instructed to examine profile characteristics such as profile colors, user name, and first name. Next, each evaluator was to examine the self-description of the profile's user. Next, each evaluator was to examine profile postings (i.e., tweets), avatar and pictures in reverse chronological order. However, evaluators were not told the self-declared gender collected from Facebook for each of the 400 randomly chosen profiles. In addition, evaluators were required to work independently of other evaluators, without communicating with each other.

Each evaluator could return, for each of the 400 profiles, one of four possible outcomes: (1) Male, meaning that the profile was thought to belong to a male user with a high degree of confidence; (2) Female, meaning that the profile was thought to belong to a female user with a high degree of confidence; (3) Male/Female, meaning that the profile was thought to belong to multiple people of different genders; and (4) Unclear, meaning that the gender of the profile's holder could not be established from the profile's characteristics.

Table 3 shows the outcomes returned by each evaluator in the case of the 200 potentially deceptive, trending male profiles. These profiles had a self-declared female gender in the corresponding Facebook profile. All evaluators identified a number of profiles as being deceptive, although the

total number of such profiles varied by each evaluator. For instance, evaluator B identified 36 profiles as being deceptive, with a further 22 profiles belonging to multiple users. At the opposite end, evaluator D identified 15 profiles as deceptive with 42 further profiles being unclear. Clearly, evaluator D followed a more conservative approach to gender verification than evaluator B.

On the whole, the five evaluators found that on average 11.3 % of the profiles belong to male users. Thus, they were indeed deceptive. Also, about 9.2 % of profiles belong to multiple people of different genders, arguably a deceptive condition. In addition, on average 11.9 % of profiles were unclear whether belonging to a male or a female user.

Table 3 Outcomes returned by each evaluator for potentially deceptive, trending male profiles

	Female	Male	Female/male	Unclear	Total profiles
A	134	25	10	31	200
B	129	36	22	13	200
C	130	21	49	0	200
D	142	15	1	42	200
E	141	16	10	33	200

Table 4 Outcomes returned by each evaluator for potentially deceptive, trending female profiles

	Female	Male	Female/male	Unclear	Total profiles
A	29	108	42	21	200
B	30	148	12	10	200
C	26	122	52	0	200
D	22	140	0	38	200
E	20	125	26	29	200

Similarly, Table 4 shows the outcomes returned by each evaluator in the case of the 200 potentially deceptive, trending female profiles. These profiles had a self-declared male gender in the corresponding Facebook profile. All evaluators identified a number of profiles as being deceptive, although the total number of such profiles varied by each evaluator. Again, evaluator B identified the highest number of profiles as being deceptive, with 30 such profiles and a further 12 profiles belonging to multiple users. This time, evaluator E identified the lowest number of deceptive profiles with 20 deceptive profiles, 26 multiple-user profiles, and 29 undecidable profiles.

Overall, the five evaluators found that on average 12.0 % of the 200 profiles belonged to female users with a high degree of confidence, meaning that these profiles were indeed deceptive. Also, there were a further 13.2 % of profiles belonging to multiple people of different genders. Finally, 9.8 % of profiles were unclear as to whether they belonged to male or female users.

Table 5 shows the degree of agreement on the gender of each profile examined among our five evaluators. We measured the frequency with which our five evaluators reached a consensus on the gender of each profile they examined. We defined different levels of consensus as three, four, or five evaluators returning the same outcome on a given profile. As the data in table shows, in the overwhelming majority of cases (90 % of the profiles) at least three evaluators of five evaluators returned the same outcome. Moreover, in 42 % of the profiles, our evaluators reached a unanimous agreement. While the number of cases in which consensus was not reached is relatively modest, 40 profiles or 10 % of the total, we believe this number is inflated by different interpretations of two of the outcomes by our evaluators. In particular, evaluator C tended to use the outcome male/female when a profile could not conclusively identified with either gender,

Table 5 Consensus results from the evaluators for all potentially deceptive profiles

	No consensus	3 consensus	4 consensus	5 consensus	Total
Trending male					
No. of Pro.	20	35	56	89	200
Female		18	40	83	
Male		4	3	6	
F/M		3	1	0	
Unclear		10	12	0	
Trending female					
No. of Pro.	20	40	61	79	200
Female		2	10	9	
Male		22	46	70	
F/M		6	0	0	
Unclear		10	5	0	

whereas evaluator D tended to use the "unclear" outcome in such cases (see Tables 3 ,4).

In summary, we are satisfied that our evaluators tended to agree quite often. Of course, an exact determination of a user's gender is impossible without access to confidential demographic information. While some individual errors in the identification a user's gender were possibly made during our verification process, we are confident that the gender of profile users was generally identified correctly by our evaluators. We concluded that about 11–12 % of potentially deceptive profiles on average are indeed deceptive with a further 11 % of profiles belonging to multiple users of different genders.

5 Detecting deception about location

In this section, we explain in detail our method for detecting deception about the location of a Twitter user and we report empirical results for detecting the corresponding profiles.

5.1 Background and rationale

To leverage the level of trust in OSNs, we need to detect the deceptive profiles by finding misleading, inconsistent, or false information using the user profiles (i.e., profile characteristics and spatiotemporal activities). This can be done using knowledge from users' activities. People nowadays periodically edit, change, and post their information using geo-tagged tweets. Thus, analysis of the user information and geo-tagged tweets that come with spatiotemporal information can provide trends of behavior leading to the detection of deception. In this work, we provide novel location-based approach that rely on publicly available information contained in Twitter user profiles and on posted geo-tagged tweets with spatiotemporal information.

5.1.1 Why does detecting deception about location matter?

Posting tweet with geo-location is now a common part of communication on Twitter. However, in geo-tagged tweets, it is relatively easy to disguise someone's location using services such as *Hotspot Shield* (AnchorFree-Inc 2014). Deception about location is sometimes indicative of a broader pattern of deception. While some Twitter members may disguise their location in order to protect their privacy, others may lie about their location to buttress lies about trips that they took or their physical whereabouts.

Analyzing geo-tagged tweets can serve a variety of stakeholders, including OSN users, governmental tourism agencies, law enforcement agencies for legal investigations, commercial advertisement agencies, and various kinds of businesses—such as restaurants and retailers—seeking to learn about the behavior of their customers.

5.2 Goals and assumptions

We are detecting deceptive profiles about locations based on finding inconsistent, misleading, unreasonable, and conflicting spatiotemporal information from a given user. For example, when a user posts multiple tweets with different locations within a short period of time, it is possible that the tweets may be fake. Twitter users may wish to conceal their locations for multiple reasons, such as to protect their privacy or to buttress additional lies about their personal life. While conducting this research, we discovered that some Twitter users lied about visiting exotic places to gain popularity among their Twitter readership. One of the user gave his Twitter account information to a friend, who is visiting a foreign location, in order to show that the original user was actually traveling!

We treat any efforts of disguising someone's location or lying about their location as deceptive. This kind of analysis faces two main challenges. First, the huge amount of tweets generated world-wide prevents us from performing a pairwise comparison of all tweets from every user whose information we crawled. For example, Twitter generates about 500 Million tweets daily. Moreover, Twitter allows us to collect around 2.5 % of tweets generated daily (or 13 Million tweets). Also, Twitter allows to collect about 50 % of the geo-tagged tweets. In the case of geo-tagged tweets, we know the exact coordinates of the Twitter client and the time when the tweet was posted. Therefore, checking all tweets from all users that we crawled would lead to an insurmountable computational complexity. In addition, validation of potentially deceiving and likely deceiving user tweets would be impossible. Second, most Twitter users do not travel most of the time. In order to conduct meaningful experiments, we must choose a time of the year when people are likely to travel.

We address the two challenges above by restricting our analysis to one specific country, Saudi Arabia, and a holiday period when many people in that country are likely to travel for vacation. The Spring break holiday period ran from March 20–27, 2014. We chose this target location for our study because three authors were in fact located in Saudi Arabia during the chosen holiday period. In this manner, we could study the activities of a set of users whose behavior we are familiar with. The uniformity and the size of the population that we studied made it easier for us to validate our empirical findings through manual examinations of tweets that we flagged as potentially deceptive.

5.3 Dataset collection

Twitter generates daily a massive amount of data that can be analyzed and classified for different reasons. Here, we use Twitter data to detect profiles containing deceptive location information using spatiotemporal features of posted tweets. We ran our crawler between March 1st, 2014 and April 30th, 2014. We started our crawler with a set of random tweets using Twitter streaming APIs. We continuously added any tweets that the crawler encountered either with or without geo-tagged information. Subsequently, we filtered out all tweets without geo-tagged information. The geo-tagged information, here, is important because it contains explicit spatial and temporal information that we use to detect deceptive profiles.

Overall, the dataset consists of around 600 Million tweets world-wide crawled between March and April 2014, including tweets without geo-tagging. We analyzed a portion of this dataset and identified about 2.5 Million unique users.

For each tweet in the dataset, we collected the spatial and temporal information, the posted tweet's text information, and the profile holder's profile information. These are the key information items needed for our study. The indicators, we considered here, for detecting deceptive profiles about location, differ from other approaches in detecting the deception, such as detecting deceptive profiles about gender, age, culture, education, ethnic information, or even political views.

Our goal was to extract users' activities two weeks before the Spring holiday as well as users' activities during the Spring holiday for the selected country. Therefore, we filtered the dataset according to spatial and temporal criteria. First, we selected geo-tagged tweets issued between March 10th and 19th, 2014. This selection yielded a dataset *D1* containing about 100 Million tweets. We further selected tweets with coordinates located in Saudi Arabia out of *D1*, resulting in tweet subset *DA* containing 1.3 Million tweets. We defined Saudi Arabia as a geographical area enclosed by a polygon with 36 sides. We identified the corners of the polygon by carefully selecting locations on the borders of that country. The tweets in dataset *DA* originated from 81,116 unique users, thought to be Saudi residents because the corresponding tweets were geo-tagged within Saudi Arabia. We denote this user set by *SU*.

Next, we selected tweets issued between March 20, 2014 and March 27, 2014—the holiday break—from our entire dataset consisting of 600 Million tweets. We obtained a dataset, *D2*, containing about 40 Million tweets. We further selected tweets originating from *SU* users from *D2*. The resulting set *DB* contains tweets created by Saudi residents and issued during the holiday break. We used the set *DB* for our analysis below. Dataset *DB* contains 293,443 geo-tagged tweets. Out of that dataset *DB*, we have 35,788 unique user profiles and 222,524 unique visited coordinates. There are 215 unique countries, including the undefined country for tweets issued from oceans or other locations not belonging to any country. Table 6 shows the countries with over 100 visits during the Spring break of March 2014 in our dataset *DB*[1]. In *DB*, there are 270,504 visits (i.e., tweets) made within the source country of Saudi Arabia. In addition, there are 6,104 visits (i.e., tweets) from the undefined country and 16,835 visits (i.e., tweets) from other defined countries than the original source country. There are 38,254 unique visits made to the 215 countries (i.e., repeated visits to the same country are not counted). There are 2,466 users who apparently visited more than one country. These visits can be conflicting visits and might be potentially deceptive profiles. In addition, there are 1,482 unique visits made to an undefined country. Furthermore, there are 2,866 unique visits to 213 different countries than Saudi Arabia and the undefined country.

Table 6 Table shows number of users visits to each country during the Spring break of March 2014

No. of visits	Country code	Country name
209490	sa	Saudi Arabia
2174	ae	United Arab Emirates
1914	kw	Kuwait
842	gb	Great Britain
716	us	United States
658	tr	Turkey
559	my	Malaysia
541	id	Indonesia
503	eg	Egypt
425	qa	Qatar
415	br	Brazil
394	fr	France
369	bh	Bahrain
298	jo	Jordan
256	de	Germany
239	es	Spain
214	aq	Antarctica
157	sd	Sudan
133	jp	Japan
132	cn	China
123	ru	Russian Federation
114	in	India
104	ca	Canada
103	it	Italy

[1] For the purpose of this research, we treat Antarctica as a country.

Figure 1 shows the flow information that we followed in creating datasets *D1* and *D2*.

5.4 Approach

In order to detect unusual behaviors by Saudi travelers during the holiday break in March, 2014, we first analyzed the prevailing behavior of those travelers during the period. Our goal was to identify and examine manually behaviors deviating from the norm before deciding our criteria for flagging potentially deceptive profiles.

We started our analysis with the whole crawled dataset. We specifically considered about 150 Million geo-tagged tweets world-wide. Next, we applied k-means clustering to locations. We experimented with various values of k, the number of clustered locations. We found that $k = 30$ was a reasonable compromise between the number of clusters and the accuracy needed to support our further analysis steps.

Next, we considered all tweets from each user in dataset *DB*. Each user is represented as a graph whose nodes convey location and temporal information (i.e., coordinates and time) of each geo-tagged tweet from that user while the edges capture the chronological movement of the user. We then mapped the nodes of each graph (corresponding to the movements of each user in *DB*) to the nearest cluster points.

We observed chronological movement patterns from the aggregated graphs (i.e., chronological movements originated from each country in the region of interest). We further simplified the graphs by choosing one location from many locations in the same country visited by a Saudi holiday traveler. We show the visualization for travel originating in Saudi Arabia in Fig. 2. Evidently, most Saudis traveling abroad during the holiday break visited exactly one country. For this reason, we decided to flag travelers visiting two or more countries as *potentially deceptive*. We counted the undefined country as well as identified countries when applying this criterion. We also flagged as potentially deceptive travelers to countries where travel is discouraged, such as countries in a state of war, since travel to such countries is highly unlikely. Moreover, we decided to flag travelers visiting three or more countries (including the undefined country) as *likely deceptive*. The remainder of our analysis is based on these two definitions.

We manually examined all potentially deceptive and likely deceptive profiles in order to determine whether the tweets from those profiles appeared consistent with real travel to the locations of the tweets. We used this analysis to determine whether a user profile was either truly deceptive or not. For all these users, we had to crawl additional data within the limitations allowed by Twitter in order to make an accurate determination. We used various kinds of information to make the determination. For example, we used inconsistent spatiotemporal information, such as tweets from disparate locations within a short period of time, to determine that a user's profile was deceptive. We plan to feed our findings about deception into our classifier to train the classifier for future analysis of this kind. Our long-term goal is to avoid manual examination of user profiles altogether by building a fully automated, ground-truth-based classifier system.

5.5 Empirical results

There are two ways in computing the deceptive location trending factors that leads to detect deceptive profiles. In this subsection, we are exploring the two approaches to detect deceptive profiles about locations.

5.5.1 Traveling to multiple foreign countries

Following the approach above, we checked profiles of users visiting multiple countries, including the undefined country, during Spring break. We found that there are 2466

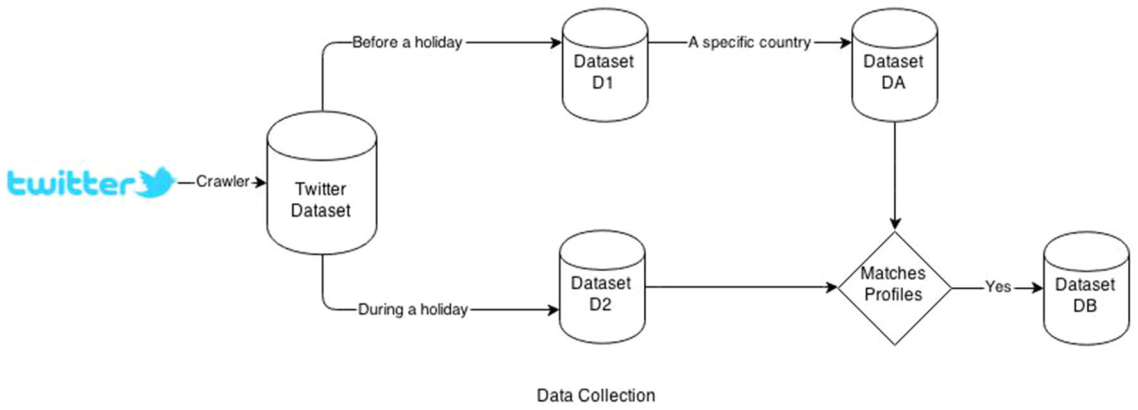

Fig. 1 The flow information for the dataset collection

Fig. 2 Where did the Saudis Spent the Spring Break of 2014

user profiles from dataset *DB* that meet this condition. This was computed by comparing the number of unique users, which is 35,788, to the number of the total visits made by those unique users as shown in Table 7. It shows the user profiles who visited either one country or more than one country during the spring break. We ignore any additional visits made inside the border of destination (e.g., if the user visits two or more locations within the same country, those explored visits are not counted, but, considered as one visit). For the purpose of this analysis, we divide the 35,788 identified user profiles into three disjoint sets. Therefore, in this subsection, we discuss *potentially deceptive* users as well as *likely deceptive* users based on vacation activities.

From Table 7, we have 1656 users out of 35,788 unique users having visited more than one country during the holiday break. Also, there are 4,142 visited countries made by these 1,656 users. Thus, in some cases, those users showed conflicting and impossible geo-location activities.

Furthermore, Table 8 shows that there are around 1,656 users identified to be as either *potentially* or *likely deceptive*. In addition, we have identified, 323 users, about 19.5 %, as potentially deceptive and 580 users, about 35.0 %, as likely deceptive, out of the 1656. Those flagged potentially and likely deceptive profiles, which shown in Table 8, were further investigated manually by following the approach we explained above.

One Naïve way to compute a statistical representation of deceptive profile is to compute speed and time as Euclidean distance:

$$\text{Deceptive}_{\text{location}} = \frac{(\text{location1}, \text{location2})}{(\text{time2} - \text{time1})} \tag{2}$$

Given this computed path speed if it is conflicting as compare to normal speed , then, it is a potential deceptive profile about location. Indeed, we verified the profiles that we identified and reported them in Table 8.

5.5.2 *Traveling to discouraged countries*

For the purpose of this analysis, we divide the 35,788 identified user profiles into two disjoint sets. Therefore, in this subsection, we discuss potentially deceptive users based on their visits to discouraged countries. We checked

Table 7 Table shows the number of profiles visiting different countries within a short period of time

No. of countries	No. of profile visits
1	34132
2	1487
3	143
4	8
5	2
6	1
7	1
8	1
10	1
22	1
25	1
40	2
47	2
51	1
55	1
86	1
206	1

Table 8 Accuracy results in detecting deceptive profiles obtained using spatiotemporal location-based approach that applied to traveler who travel to multiple foreign countries

	Neutral	Potential	Likely
Number of profiles	34,132	1487	169
Neutral	–	751	2
Potentially deceptive	–	308	15
Likely deceptive	–	428	152
The accuracy	–	28.8 %	89.9 %

profiles of users visiting discouraged countries during Spring break. We follow a simple and greedy statistical method that uses *DB*.

First, we identified a list of discouraged countries, such as countries in a state of war. It is highly unlikely that someone will go for a vacation in such a country. We then flagged any profiles that spent the holiday break in such countries. The list of the discouraged countries are different from a country to another. For our study, we selected the 10-top discouraged countries provided by the government of Canada to their citizen since the government of Saudi Arabia does not provide any list of discouraged countries. We detailed this list of discouraged countries in the discussion subsection.

Assume that *DC* is the list of discouraged countries. This list should be subset of the country list that extracted from dataset *DB*. Therefore, any profile from the dataset *DB* to countries that meets this condition, is flagged as *potentially deceptive*. We identified 62 visits that are subset of $DC_{\text{discourage}-\text{countries}}$. Also, from the 62 visits, we identified 32 unique users. Thus, those 32 users are flagged as *potentially deceptive*. We manually further inspected those users and identified 29 users, about 90.0 %, as *likely deceptive*. In fact, All 29 users are indeed identified earlier in the subsection of traveling to multiple foreign countries (i.e., 29 users match the list of likely deceptive profiles that we identified in the previous subsection). Table 9 shows the accuracy results in detecting deception using the top-10 discouraged countries.

Table 9 Accuracy results in detecting deceptive profiles obtained using spatiotemporal location-based approach that applied to traveler who travel to discouraged countries

	Neutral	Potential
Number of profiles	35,756	32
Neutral	–	1
Potentially deceptive	–	2
Likely deceptive	–	29
The accuracy	–	90.0 %

In conclusion, we are only including the top-0 discouraged countries. However, if we have including more discouraged countries or the least visited countries to this approach, we may identify more profiles to be as potential deceptive.

5.6 Discussion

In this section, we explain how we validated our findings by comparing them with information about travel destinations of Saudi residents posted by the Saudi Tourist Information and Research Centre (STIRC). We also validated our findings by manually inspecting potentially and likely deceptive profiles. Also, we include some challenges faced during this investigation.

5.6.1 Validation by comparing with official data

We confirmed travel destinations of users in Saudi Arabia based on a study conducted by the Saudi Tourist Information and Research Centre (TIARC 2014). This study was published by the SABQ Online Newspaper (Newspaper 2014). According to the study, the top 10 destinations for about 6 Million Saudis are United States of America, United Kingdom, Malaysia, Gulf Cooperation Council Countries excluding Saudi Arabia, Indonesia, Philippines, Turkey, Morocco, Australia, and Switzerland. Similarly, our dataset shows our findings match the study by the Saudi Tourist Information and Research Centre. Our findings show that the top 10 destinations are Gulf Cooperation Council Countries excluding Saudi Arabia, United

Kingdom, Indonesia, Turkey, United States of America, Egypt, Jordan, Malaysia, France, and Spain. It also shows more than expected visits to such countries as Brazil, Germany, and India. This validation leads us to have a better understanding about where the Saudis are spending their vacations as normally expected according to the government data. Therefore, any conflicting or unexpected destination locations information to the Saudis must be checked for further investigation.

According to the government of Canada (Gofcanada 2014), there are 12 discouraged destinations. The citizens of Canada are warned not to visit the following countries: Niger, Chad, South Sudan, Somalia, Yemen, Central Africa Republic, Syria, Iraq, Iran, Afghanistan, and North Korea. Our experiments have shown that there are 62 unique Saudis who visited these countries.

5.6.2 Validation of the user profiles who made more than 50 visits

Here, we validate our dataset by randomly selected any profile who meets the following condition. The condition is to select any profile in our dataset *DB* who visits more than 50 locations during the holiday break. Given the fact about where the Saudis spent their vacations, we have identified around 880 unique users, about 2.4 % of the population, who visited more than 50 locations (i.e., more than 50 checked-in) during the holiday. In this case, we counted all the visits the user made—inside and outside—the countries that the user explored. In fact, we crawled those 880 users again to get more tweets information, and found that they generated more than 1.3 Million tweets in which around 1.1 Million tweets contain geo-location information and the others come without geo-location information. As a result, we further investigated those profiles by applying our manual approach to check whether those geo-tagged tweets are inconsistent with the spatiotemporal information. We found, yet, that 523 out of the 880, about 59.4 %, users are likely deceptive profiles and we report that in Table 8.

Moreover, in another way in selecting random profiles to be investigated manually, we have listed all the countries that were visited by Saudis in ascending order of their visits. We have around 215 unique points (i.e., countries). Also, there are around 34 countries have been visited by at least 10 unique Saudis. In contradiction, there are 1482 Saudis who visited the undefined country. In addition, there are around 180 countries have been visited by at most 9 unique Saudis. From the bottom of the list, we randomly selected 200 profiles with visits to discouraged countries to be manually inspected for deception. We found 34 profiles, about 17 %, are likely deceptive after manually inspected them. Through this investigation, we also randomly selected one profile out of the 34 likely deceptive users to deeply manually inspected. The chosen profile visited a discouraged country in which located in Africa. This kind of visit is considered as unusual, and, to be as *potentially deceptive* at the one hand. On the other hand, we manually further inspected this profile. Therefore, we found that the profile generates random geo-tagged tweets that come with random geo-location and random posting text in every 5 min.

5.6.3 Challenges

We experienced many challenges during dataset collection and validation. One of the major challenge was that some of geo-tagged tweets do not have accurate or complete geo-location information which make it a bit difficult decision for the weighted spatiotemporal features indication. Thus, the spatiotemporal features indications must be inter-changed dynamically based on the available information.

Another challenge is that some of the profile's settings are edited periodically by the owners. For example, we collected enough geo-tagged tweets information for a profile at one point of time, but, on the other time, we have different geo-tagged tweets information that belong to the same profile. Currently we just excluded those types of profiles.

6 Conclusions and future work

Our ultimate goal is to find inconsistent information in online social networks about user gender and location in order to detect deception. In particular, we defined a set of analysis methods for that purpose on Twitter. Also, we apply Bayesian classification and K-means clustering algorithms to Twitter profile characteristics (e.g., profile layout colors, first names, user names, and spatiotemporal information) to analyze user behavior. Therefore, in this study, we presented frameworks for detecting deception about gender and location information. In addition, we reported preliminary empirical results with a strategy for attaining this goal within the framework. Through extensive experiments, our current results show considerable promise for our framework. Based on the outcomes of our approach, we are able to detect deceptive profiles with an accuracy of around 90.0 % in some cases. Our empirical experiments obtained by applying our algorithms to multiple datasets showed promising results.

In the future, we will continue exploring alternative strategies in an effort to improve the accuracy of our predictions. Although, our two approaches in detecting the deception, namely detecting the deception about *gender* and *location*, are independent and different in term of their depth, properties, structures, and novelties, yet, the

synthesis of the two approaches are going to be implemented and going to provide a powerful tool in detecting the deceptive profiles. We will also consider additional features, such as the genders of Twitter friends and followers, as part of gender predictions as well as more features in the location. We will also explore text-based features factors for both approaches, such as user postings, and we will include these features if their advantages outweigh their cost in terms of language dependence and increased computational complexity. Finally, we plan to explore more novel approaches in detecting the deception such as age and other factors that are supported by our main framework.

References

Al Zamal F, Liu W, Ruths D (2012) 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), 2012

Alowibdi JS, Buy UA, Yu PS (2013) Empirical evaluation of profile characteristics gender classification on Twitter. In: The 12th International Conference on Machine Learning and Applications (ICMLA), vol. 1, Dec 2013, pp 365–369

Alowibdi JS, Buy UA, Yu PS (2013) Language independent gender classification on Twitter. In: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM'13, Aug 2013, pp 739–743

Alowibdi JS, Buy UA, Yu PS, Stenneth L (2014) Detecting deception in online social networks. In: Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on, 2014

AnchorFree-Inc. (2014) Hotspot shield, http://www.hotspotshield.com/

Authority AS (2013) Children and advertising on social media websites, http://goo.gl/qswXGe

Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, Ohl P, Thiel K, Wiswedel B (2009) Knime-the konstanz information miner: version 2.0 and beyond. ACM SIGKDD Exp Newslett 11(1):26–31

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

Caspi A, Gorsky P (2006) Online deception: prevalence, motivation, and emotion. CyberPsychol Behav 9(1):54–59

Castelfranchi C, Tan YH (2001) The role of trust and deception in virtual societies. In: Proceedings of the 34th Annual Hawaii International Conference on System Sciences

Castillo C, Mendoza M, Poblete B (2011) Information credibility on Twitter. In: Proceedings of the 20th ACM international conference on World wide web, 2011, pp. 675–684

Cheng Z, Caverlee J, Lee K (2010) You are where you tweet: a content-based approach to geo-locating Twitter users. In: Proceedings of the 19th ACM international conference on Information and knowledge management, 2010, pp 759–768

E.-M. of RealWire.com (2007) Social networking sites: Almost two thirds of users enter false information to protect identity, http://goo.gl/ERtNdA

Guerrero LKK, Andersen PA, Afifi WA (2012) Close encounters: communication in relationships. Sage Publications, USA

G. of canada (2014) Country travel advice and advisories, http://travel.gc.ca/travelling/advisories

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The weka data mining software: an update. ACM SIGKDD Exp Newslett 11(1):10–18

Hancock JT, Curry L, Goorha S, Woodworth MT (2004) Lies in conversation: an examination of deception using automated linguistic analysis. In: Annual Conference of the Cognitive Science Society, vol. 26, 2004, pp 534–540

Jurgens D (2013) Thats what friends are for: Inferring location in online social media platforms based on social relationships. In: Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, 2013

Lenhart A, Madden M (2007) Teens, privacy & online social networks, http://www.pewinternet.org/Reports/2007/Teens-Privacy-and-Online-Social-Networks.aspx

Liu W, Al Zamal F, Ruths D (2012) 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

Liu W, Ruths D (2013) 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

Mislove A, Jørgensen SL, Ahn YY, Onnela JP, Rosenquist JN (2011) Understanding the demographics of Twitter users. In; 5th International AAAI Conference on Weblogs and Social Media (ICWSM'11), 2011, pp 554–557

Newman ML, Pennebaker JW, Berry DS, Richards JM (2003) Lying words: predicting deception from linguistic styles. Personal Social Psychol Bulletin 29(5):665–675

Pennacchiotti M, Popescu AM (2011) A machine learning approach to Twitter user classification. In: proceedings of the International Conference on Weblogs and Social Media, 2011

Rao D, Paul MJ, Fink C, Yarowsky D, Oates T, Coppersmith G (2011) Hierarchical bayesian models for latent attribute detection in social media. In: 5th International AAAI Conference on Weblogs and Social Media (ICWSM'11)

Rao D, Yarowsky D, Shreevats A, Gupta M (2010) Classifying latent user attributes in Twitter. In: Proceedings of the 2nd international workshop on Search and mining user-generated contents, 2010, pp 37–44

S. O. Newspaper (2014) Saudi top destinations abroad, http://sabq.org/yX6fde

Sakaki T, Okazaki M, Matsuo Y (2010) Earthquake shakes Twitter users: real-time event detection by social sensors. In Proceedings of the 19th international conference on World wide web, 2010, pp 851–860

Speech at CMU (2013) The CMU pronouncing dictionary http://www.speech.cs.cmu.edu/cgi-bin/cmudict

T. Information and Research Centre (2014) Tourism information, http://www.mas.gov.sa/

Thomas K, McCoy D, Grier C, Kolcz A, Paxson V (2013) Trafficking fraudulent accounts: the role of the underground market in Twitter spam and abuse. In: USENIX Security Symposium, 2013

Turner B (2010) Do people often lie on social networks? http://curiosity.discovery.com/question/do-people-lie-social-networks/

Warkentin D, Woodworth M, Hancock JT, Cormier N (2010) Warrants and deception in computer mediated communication. In: Proceedings of the 2010 ACM conference on Computer supported cooperative work, 2010, pp 9–12

Yardi S, Romero D, Schoenebeck G et al (2009) Detecting spam in a Twitter network, First Monday 15(1)