Detecting Deception in Online Social Networks

BY

JALAL ALOWIBDI
B.S. (King Abdulaziz University, Jeddah, Saudi Arabia) 2005
M.S. (DePaul University, Chicago, USA) 2008

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois at Chicago, 2014

Chicago, Illinois

Defense Committee:

Dr. Ugo Buy, Computer Science, Chair and Advisor
Professor Philip Yu, Computer Science, Co-advisor
Professor Bing Liu, Computer Science
Dr. Brian Ziebart, Computer Science
Dr. Houshang Darabi, Mechanical and Industrial Engineering

To my father, Suliman & my mother Braikah, and my family

for their constant support and unconditional love.

ACKNOWLEDGMENTS

I would like to convey my deep gratitude and appreciation to my supervisor, Dr. Ugo A. Buy, for his unwavering encouragement, guidance and support during these past five years. Without his supervision and constant help this dissertation would not have been possible. Dr. Ugo is someone you will instantly love and never forget once you meet him. He is a great coach, a professional adviser, always on my side, a positive attitude and one of the smartest people I know. I hope that I could be as lively, enthusiastic, and energetic as Ugo and to someday be able to command an audience and students as well as he can. I am also very grateful to him for his scientific advice and knowledge and many insightful discussions and suggestions about the research. He is my primary resource for getting work done in short, simple and meaningful way, by answering all my scientific questions and concerns to help me crank out this thesis.

In addition, I would like to express my sincere gratitude to my co-advisor, Dr. Philip S. Yu for giving me the opportunity to carry out my research project to be recognized, for his support, and for his continuous guidance. Dr. Philip has been always played a key role in encouraging and coordinating this whole project. Also, I would like to express my sincere appreciation to Dr. Bing Liu, Dr. Brian Ziebart, and Dr. Houshang Darabi for serving on my dissertation committee and providing insightful comments and stimulating thoughtful discussion.

Moreover, I especially thank All the Saudis, starting from the king of the kingdom of Saudi Arabia who is Abdullah Bin Abdulaziz, to every Saudi living in this kingdom for the opportunity of getting the scholarship that is provided by the ministry of higher education and

ACKNOWLEDGMENTS (Continued)

King Abdulaziz University which is organized and represented by the Saudi Arabia Cultural Mission. Without the big support and change in the education system in Saudi Arabia, I would not be here, today, defending my thesis, yet, I would end up with normal life as most Saudis do that passes without fully educated. However, with the courage and support of my country, first, and my family, second, I am today completed my education with highest degree and with full of knowledge in computer science. Therefore, I am going to take this knowledge and education back to my country to be part of developing and educating the Saudis. My hard-working parents and family have sacrificed their lives for me and provided unconditional love and care. I love them so much, and I would not have made it this far without them. I know I always have my family to count on when times are rough. Yet, I succeeded passing those rough times. Also, I am still standing on my feet with many things that make my life different in term of mindedness and the quality of education that I should deliver back to my country. Further, I am always going to pursue self-education to many different things in computer science to keep up with the knowledge and research freshness. Thus, this degree should not stop me from acquiring more knowledge to be part of the change. I just want to mention that when I first came to the states, I have this statement in my mind "My personal vision is to learn more, study hard, work sincerely, search through the information, share my opinion and become a scientist". I wrote this statement in the airplane and I challenged myself to follow those steps in order to succeed and I hope I did!

ACKNOWLEDGMENTS (Continued)

Finally, I would like to end this acknowledgement by emphasizing and expressing my sincere love and gratitude to my great grandfather, Saleem, and my family too. Their love, encouragement and support have been the root of this success.

JSA

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

LIST OF TABLES

LIST OF TABLES (Continued)

LIST OF FIGURES

SUMMARY

Online Social Networks (OSNs) are platforms that contain a variety of services for users to interact over the Internet. OSNs have spread at a stunning speed over the past decade. They are now a part of the lives of dozens of millions of people. 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. Over the past decade OSNs have been helping hundreds of millions of people develop reliable computer-mediated relations. In addition, we believe that OSNs play a significant role in the daily life of hundreds of millions of people. However, many user profiles in OSNs contain inconsistent 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.

To reach this level of trust in OSNs, we need to detect the deceptive profiles by finding misleading, inconsistent, conflicting or false information. Although privacy issues in OSNs have attracted a considerable attention in recent years, currently there is no work on detecting deception based on inconsistencies in a user's profile and posts. In this dissertation, we presented a novel approach for detecting deceptive profiles in OSNs which is our ultimate goal. The problem of detecting deception is important, but extremely challenging and worth to pay more

attention. There are several methods for detecting deceptive profiles, none of which study the deception detection in gender and location, as we explore in this study.

In this dissertation, we explore methods for detecting information about user genders and locations. We first start by addressing the relative strengths of possible indicators for detecting deception about gender and location. Evidently, the indicators that we consider are different for gender and location. Next, we study the effectiveness of these indicators in determining which indicators will help and which will not help. To date, there are no reliable indicators for detecting deception. Therefore, our goal is to find the right indicators for deception about gender and location. In particular, we discuss separately our two approaches for detecting deception about gender and location. We have two different datasets where for each of the two, we have applied two different sets of experiments.

On the one hand, we have studied the effectiveness of profile characteristics for detecting the gender of Twitter users with a dataset that we harvested between January and February 2014. Our approach to deception detection uses our novel approach on gender classification utilizing profile layout color contained in Twitter profiles and first names and user names contained therein. Each profile in our dataset 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. In addition, these characteristic help

find inconsistent information about the gender from different characteristics and flag potential deceptive profiles.

On the other hand, we have studied the effectiveness of spatiotemporal activities for predicting the location of Twitter users with a different dataset that we harvested between March and April 2014. We use publicly available Twitter data of that period to find out where the people spent their vacation in a particular country and a particular holiday. We have explored geo-tagged tweets that come with geo-location activities for a specific group of people. In particular, we have selected Saudi Arabia as a source location and the Spring break holiday in March, 2014 for this study to find conflicting and unrealistic geo-location information. Thus, the outcome of this study is that such spatiotemporal activities by a user can provide reasonably accurate predictions of the users' locations. Also, it does help predict the deceptive profiles based on finding inconsistent spatiotemporal's information.

The long-term objective of this research is to flag automatically deceptive information in user profiles and posts, based on detecting inconsistencies in a user's profile and posts. Here, we focus on detection of inconsistent information involving user gender and conflicting spatiotemporal activities involving user locations. Our method is centered on a Bayesian classifier that takes into account different profile characteristics (i.e., indicators) and returns gender and location *trending factors*, which correlate to the probability in classifying a Twitter user. For detecting the deception about gender, our classifier works in such a way that the computed so-called *male trending* 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 $f = 1 - m$. We use a similar method (i.e., conflict spatiotemporal information) to detect the deception about location. We additionally use manual inspections on a subset of profiles that we identify as potentially deceptive in order to verify the correctness of our predictions (e.g., likely deceptive) since there is no ground truth for us to use in order to validate our predictions.

Our goal is to study and determine which indicators is going to be considered for that purpose. To address the problem, this dissertation has defined a set of analysis methods with observed behavioral footprints for detecting deceptive information about user genders and locations in Twitter. We apply Bayesian classification and K-means clustering algorithms to Twitter profile characteristics (e.g., profile layout colors, first names, user names, spatiotemporal information) to analyze user behavior. Our approach to detect deception is mostly independent of the user's language, efficient, scalable, and computationally tractable, while attaining a good level of accuracy. We establish the overall accuracy of each indicator and the strength of all possible values for each indicator using real-world OSN, namely Twitter, to demonstrate the effectiveness of our model. Our empirical experiments obtained by applying our algorithms to multiple datasets showed promising results. To our knowledge, this is the first work to detect deceptive profiles by employing principled inconsistent and conflicting information classification modeling. Thus, this work study probabilistic and statistical models and algorithms for detecting the deception in OSNs by leveraging concepts and methodologies of the classification techniques.

CHAPTER 1

INTRODUCTION

Online Social Networks (OSNs) have spread at stunning speed over the past decade. They are now a part of dozens of millions of people. The growth in the user base has led to a dramatic increase in the volume of generated data. In addition, 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. The key factor underlying the success of OSN-mediated communities, similar to traditional communities, is the trust that exists among community members. Trust in the community is multi-faceted. Community members must not only trust each other, they must also trust the OSN infrastructure including software clients, network servers and the authorities in charge of OSNs to uphold privacy and confidentiality standards.

Typically, users create profiles on OSNs describing their interests, activities and additional personal information. Then, users often start looking for friendships with other users who might be friends, family members, co-workers, classmates and even perfect strangers, who might happen to share common interests. A recent study by McAfee showed that 95% of young people, with ages between 10-23, in the United States have at least one OSN account; and 86% of these young people believe that OSNs are safe (1). In addition, these profiles may be filled to various extents of completeness depending on the user. Some profiles may have highly-detailed personal information, whilst other profiles may contain just basic information. According to McAfee,

around 90% of young people believe that it is dangerous to post information online such as personal information (i.e., social security number, family member names, home address, phone number, etc.) in OSNs (1). However, they still insert personal information or enough data where personal information can be inferred! Hence, OSNs often hold private and confidential information about people and organizations. Therefore, ensuring integrity, accessibility and confidentiality of such information in OSNs is a major concern to people, organizations and companies who are using OSNs.

In addition, we believe that OSNs play a significant role in the daily life of hundreds of millions of people. Many users in OSNs have been engaged and communicated with each other, but, many user profiles contain false information or omit complete truth. In fact, it is easy to provide false information in someone's profile in order to deceive others. Hence, lying in profiles and posts is apparently widespread. Existing studies have shown that lying in OSNs is quite pervasive, often for protecting a user's privacy, for attracting others attention, for avoiding hurting the feelings of others, for enhancing our image, for maintaining relationships or for avoiding psychological trauma (2). 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. In actual fact, lying is one form among the five forms of deception (2). This lying about some information which is one reason out of many mentioned earlier.

A recent study by the Advertising Standards Authority shows that 42% of children (i.e., people ages under 13) reported that they lie (i.e., provide false information) about their age in order to see products or contents with age restrictions (3). Evidently, users often provide

false information about their age either to deceive others or for personal reasons such as to access content restricted by age. For instance, a survey by the EMedia Group shows that 62% of OSN users in the United Kingdom are worried about the privacy of information they entered in OSNs; as many as 31% had actually entered false information about themselves in OSNs to protect their privacy (4). Another study, posted on the Pew Internet & American Life Project, showed that 56% of teenagers surveyed entered false information in their online profiles primarily "to keep themselves safe from unwanted online attention" (5). These considerations make it all 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 the deception; it is unclear which indicators will help and which will not help. In fact, deceiving people will sometimes require great efforts to disguise their deceit. For instance, a man posing as a woman chooses a false identity (e.g., first name and username) in order to make their false statements believable. Another example is that a man is pretending to be in another country chooses fake locations (e.g., latitude and longitude information) to places he never went to, to make his travel believable. Regardless of different ways that users can disguise their identities, we plan to address these difficulties by considering a wide variety of features indicators for gender and location, in an effort to detect situations of this kind.

To help reach a good level of trust in OSNs, we need to detect the deceptive profiles by finding misleading, inconsistent, conflicting or false information. Although privacy issues in

OSNs have attracted considerable attention in recent years, currently there is no work on detecting the deception based on inconsistencies in a user's profile and posts on OSNs. To achieve that purpose, we have faced three major challenges: (1) finding the relative strength of all possible value indicators for deception detection, (2) studying the effectiveness of those indicators for classifying user profiles, and (3) comparing those indicators in order to flag and detect potential deceptive profiles with inconsistent information (i.e., conflict indications). In this dissertation, we presented a novel approach for detecting deceptive profiles in OSNs, which is our ultimate goal. We succeeded in furthering our essential goal by exploring and finding inconsistent information about user gender and conflicting spatiotemporal information about user location. Automatic detection of deception can serve multiple purposes. For example, commercial organizations may utilize detection of a deception in advertising in order to uniquely deliver the right advertisement messages to the right people. Law enforcement may use the automatic detection of deception as part of legal investigations and to bring witnesses to the court. Others may use it for social reasons.

To detect deception, we use characteristics extracted from user profiles and posts; however, the characteristics that we consider are different for gender and location. In the case of a user's gender, we specifically consider the following profile's characteristics extracted from each user's profile information, which are as follows:

1. First name.
2. User name.
3. Background color.

4. Text color.
5. Link color.
6. Sidebar fill color.
7. Sidebar border color.

Similarly for location, we consider the following characteristics extracted from each user profile and post, namely Tweet, which are as follows:

1. Temporal information.
2. Spatial information.
3. Location.

Other researchers might additionally include age, culture, education, ethnic information or even political views beside gender and location in order to detect deception by studying different others profile's characteristics.

In summary, our approach in detecting the deception compares gender and location indicators obtained from different profile characteristics. We then compare the indicators obtained from each feature probabilistically and statistically; And we flag as *potentially deceptive* users' profiles. Furthermore, there are two different methods produced by the end of this dissertation, which are as follows:

1. A new method and empirical tools for gender classification (i.e., guessing) using novel preprocessing methods for profile characteristics such as colors and names.

2. A new method and empirical tools for deceptive users classification (i.e., guessing) based on inconsistent information involving user gender and conflicting spatiotemporal activities involving user locations.

The remainder of this chapter discusses the dataset collection, challenges, and a contribution of our dissertation.

1.1 Dataset Collection

We chose Twitter profiles as the starting point of our data collection for several reasons. First, Twitter is one of the most popular social networks to date with a user population. Therefore, Twitter has a huge user community, cutting across a great many languages, cultures and age groups. In addition, Twitter has a large amount of available data. In early 2013, Twitter reached 645 million registered users (6). Twitter states that there are more than 255 million active users producing around 500 million tweets per a day (7). Second, Twitter has all the attributes that we need to set up the experiment. Twitter profiles include a full name, username, description, layout colors and posted texts, aptly named "tweets", of each user. These attributes are generally public, meaning that they can be accessed and viewed by anyone who requested them. Third, Twitter provides a rich Application Programming Interface (API), which supports automatic collection of large data sets. Lastly, although this practice is illegal, various companies offer services whereby they will create fake users and make them follow a given client. Followers are people who subscribe to view the tweet of another Twitter user. The number of followers, that a user has, is considered a status symbol within the Twitter community. Thus, many users buy services that create fake followers for them in

order to increase their total follower count. It is nice to be able to catch some of these fake users!

1.2 Challenges

In general, there are several challenges to be considered in detecting deception in OSNs, which are as follows.

- 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's characteristics correctly.
- Deceivers often 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.
- For the technical side, the dataset collected from Twitter contains text; yet, analyzing text leads to high dimensional space and computational complexity. For example, there are around 15.4 Million features extracted from the texts for gender classification by Burger et al. (8). This way leads to face great challenges because of high computational complexity (i.e., scalability and efficiency issues).

- Most OSNs are supporting multilingual texts. Twitter, for instance, contains more than 70 languages (9). Our challenge is to have a language independent algorithms that deal with massive datasets in detection of deception.
- 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 the genuine and fake user profile.
- Lastly, gaps in available information is another challenge. The method of gender guessing is first used as input to the method of deception detection.

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 dataset.

1.3 Contributions of work

In this section, we will highlight the contributions of the this dissertation work. First, we will present the contributions on gender classification using profile layout colors. Next, contributions on other attributes, including profile's first name, user name are provided. Finally, our ultimate contribution to the detection of deception, of inconsistent information involving user gender and conflicting spatiotemporal activities involving user locations, is presented. Our contribution can be summarized as follows:

Say It with Colors: Language-Independent Gender Classification on Twitter. Our preliminary work was published in (10). We explored in depth language independent gender classification. We predicted automatically the gender value of users based on their color

preferences. Unlike text-based approaches, we used a novel method for predicting gender using five color-based features. Our preliminary results with our dataset were quite encouraging. Although we were considering only five color-based features, we predicted the gender with an accuracy of 74.2%, a gain of about 24% with respect to a 50% baseline. A key to the success of our gender guessing with colors is our preprocessing of color features using a quantization technique that we discuss later on.

In brief, 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 applied *color clustering* whereby we reduce the representation of profile colors from the traditional 8-bit RGB representation to a 3-bit RGB representation The traditional 8-bit RGB representation yields a feature set consisting of $2^{8*3} = 2^{24}$ 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 a few units from the original color. Thus, we chose to cluster colors in such a way that colors within a given cluster are perceptually similar to each other. In this manner we reduce the total size of our color set to $2^{3*3} = 2^9$ or about 512 colors. The advantage is that we obtain a statistically significant number of profile users in each color cluster from our dataset. 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 in the sequence.

We compared empirically the performance of gender predictions using raw colors and colors obtained by applying clustering and sorting. The accuracy of our gender predictions improved from 65% to 74.2% when applying the two preprocessing stages.

An advantage of our method is its broad applicability to Twitter users, regardless of their language; we use only color-based features to identify gender. In addition, our color-based analysis shows promising results in term of computational complexity compared to other gender-guessing methods, which use a much larger feature set. Our approach utilizes only five color-based features while Burger et al. (8) and Rao et al. (11) use text sentiment with 1.2 Million and 15.4 Million features respectively. Our results show that colors alone can provide reasonably accurate gender predictions, even though a substantial number of users we analyzed do not change the default colors provided by Twitter in their Twitter profiles or in other websites hosting their profiles (e.g., Twitter App). We conclude that colors are a good gender indicator for users who do change the default colors in their profiles. In these cases, we will be able to use colors alone as part of our gender classification methods.

Empirical Evaluation of Profile Characteristics for Gender Classification on Twitter. This work of ours was published in (12). In this work, we considered a novel approach in order to identify the gender of a user's profile from other attributes than colors, including profile's first name and username, for gender classification. Our work, here, is different from existing methods because of its simplicity and the range of profile characteristics that

we consider. We defined a so-called phoneme-based preprocessing technique for reducing the number of features. Our method typically results in a reduction in feature space size by two to four orders of magnitude.

The phoneme-based preprocessing stage works as follows. 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 harvested from Twitter. 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 (13). 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. For instance, our accuracy for gender prediction for first names has improved from about 71% to 82.5% because of this preprocessing stage. We are quite encouraged that not only we improved the accuracy of our gender predictions; 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.

Detecting Deception in Online Social Networks. 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 inconsistent information. For instance, some of the profiles in OSNs might be fully true, partially true or not true based on the degree of the information provided on each profile. Existing studies have shown that lying in OSNs is quite widespread, often for protecting a user's privacy. In our final works that was published in (14; 15), we presented a novel approach for detecting deceptive profiles in OSNs. Our ultimate goal, here, is to find inconsistent information about user gender and location. There are several methods for detecting deceptive profiles, none of which study the deception detection in gender and location, as we explore our research. In particular, we define a set of analysis methods for detecting deceptive information about user genders and locations in Twitter. We apply Bayesian classification and K-means clustering algorithms to Twitter profile characteristics (e.g., profile layout colors, first names, user names, spatiotemporal information) to analyze user behavior. We establish the overall accuracy of each indicator and the strength of all possible values for each indicator through extensive experimentations with our crawled dataset.

Our preliminary results with our datasets are quite encouraging. On the one hand, we can identify deceptive information about gender and location with reasonable accuracy. On the other hand, our approach uses 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, because our approach combines a good accuracy and language independence with low computational complexity by its simplicity.

Through our analyses, we have identified several thousands, *potentially deceptive* and *likely deceptive* profiles. We manually inspected likely deceptive profiles, as we report below, and found that a large proportion of those profiles were indeed deceptive.

On the one hand, for the gender based approach in detecting the deception, we have identified 4% of the 174,600 profiles collected as potentially deceptive profiles. Therefore, we manually inspected profiles deemed to have higher probabilities to be deceptive, 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 inspect them thoroughly 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 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.

On the other hand, for the location based approach in detecting the deception, we have identified 5% of the 35,000 profiles collected as potentially deceptive profiles. Yet, we manually inspected profiles with higher probabilities to be deceptive, as we report below, 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 potentially deceptive profiles that we identified. We found, in some cases, that's about 90.0% of the potentially deceptive profiles were indeed likely deceptive. In addition, the overall outcome of 5.0% of the users is potentially deceptive

and about 35.0% of those users are likely deceptive. We conclude that our approach can provide reasonably accurate predictions of gender and location feature-based deception.

1.4 Thesis Structure

This dissertation is organized as follows.

Chapter 2: Literature Review. Since our work, detection of deception in OSNs, is unique and we have not known any research with focus on detection of deception in the industry to begin with, we decided in this chapter to divide the literature review into three main subjects. First, we investigated and reviewed some of the proposed methods and approaches in gender classification in both OSNs and micro-blogs. Second, we examined and reviewed location-based classification. Finally, we explored and reviewed some of the proposed methods for detecting spam and fraud in Twitter.

Chapter 3: PChars: Profile Characteristics Gender Classification. We investigated in this chapter different profile's characteristics for gender classification. Our approach predicts gender using profile's characteristics based on features extracted from Twitter profiles and posts (e.g., the background color in a user's profile page).

Chapter 4: Detecting Deceptive Information in Twitter About User Gender. In this chapter we introduced our approach for detecting the deception about gender.

Chapter 5: Detecting Deceptive Information in Twitter About User Location. In this chapter we introduced our approach for detecting the deception about location.

Chapter 6: Conclusion and future studies. This chapter concludes our exploration on detection of deception in OSNs. In addition, we pointed out the possible future studies.

CHAPTER 2

LITERATURE REVIEW

This chapter provides some studies that are relevant to our dissertation, but not about detecting the deception in Online Social Networks (OSNs). As for the author's knowledge, this work is the first of its kind and this field has not yet been explored. Most works do not emphasize the deception in OSNs, however, there are some studies about detecting spam, fraudulent behavior and gender classification in OSNs and other platforms such as blogs and articles. Specifically, in the literature, a wide variety of methods on spamming and only handful works on fraudulent behavior in OSNs have been investigated (i.e., there is no method for detecting the deception in OSNs). Also, there are some general studies for deception that explored deception in media and books but not deception in OSNs as we are doing in this dissertation.

We discuss this chapter into three separate subsections. Gender classification is the first subsection. Location classification is the second subsection. The third subsection reviews deception, fraud and spam in electronic media. The following subsections briefly provide a literature review of different works that are related to our work.

2.1 Gender Classification

In this dissertation, we will implement a classification technique using a feature-based approach, extracted from the user profile's characteristics in the OSNs (e.g.,profile layout colors)

to identify the gender of the profile holder. Therefore, we analyze the gender differences to apply them to our model for detection of deception. Gender classification seeks to identify whether either a male or female author produces the text contents that we analyze. In our case, we are additionally looking for the authors' activities (e.g. profile's metadata, profile's layout style, profile's layout colors, profile's statistic information) to identify the gender. Over the past few decades, most existing work explores gender classification by utilizing a text sentiment approach. Researchers from natural language processing community and data mining community worked together on gender classification in both traditional contents, including articles, novels, news, books and documents, and non-traditional contents, including blogs, forums, emails, chats and OSNs. Despite the challenging feature set of these contents, in this section, we have studied various schemes for defining feature feasibility and stability by different researchers.

Above all, researchers must pay more attention to the structure of those systems. On the one hand, in traditional contents they worked with formal and structured writings, long text, and carefully reviewed materials. On the other hand, in non-traditional contents they worked with mostly informal and unstructured writings, short text, carelessly reviewed materials that contain grammatical and spelling errors, slang phrases, informal sentences, emoticons, and abbreviations. These styles can be examined by different methods such as writing-style, Part-of-Speech (POS) n-gram models, word-frequencies, word-classes, POS patterns, POS contents, POS style metrics, POS tags, and so on. Although most existing authors extracted millions of features from text sentiment based on the structure of the systems, our work shows that reasonably accurate predictions are possible using a few selected elite features. Our work is

going to be introduced in the next chapter. The drawback of using text sentiment is the computational complexity of the generated high dimensional space.

In traditional contents, many researchers have been investigated gender classification (16; 17; 18; 19; 20; 21; 22). Hota et al. (16) explored gender classification in literary Shakespeare's characters by using plays. Also, they applied two features set of words, which are stylistic feature set and content-based feature set. Then, they calculated the frequencies of these sets after linearly weighting the two text features. Thus, we found that their findings are quite interesting since they compared different patterns in literary Shakespeare's gendering of his characters between the early and late plays. Furthermore, Argamon et al. (18) investigated the exact same corpus of literary Shakespeare's characters with a different methodology in identifying the gender. However, they used lexical feature based on taxonomies of lexical items (e.g. words and phrases). Therefore, they applied different types of functional lexical features, which are feature set of words, conjunction, modality, comment, appraisal and various combinations. Accordingly, this work would not be effective and useful without using the theory of systemic functional grammar by Halliday (23), a functional approach to linguistic analysis in term of their semantic function.

Koppel et al. (17) investigated gender classification in formal written documents in British English (e.g. fiction and non-fiction genres). They applied three feature sets, which are function words, POS and combination of both function words and POS. Their findings are quite interesting since they also compared two different patterns on documents (e.g. fiction and non-fiction) to identify the gender. In another work by Argamon et al. (19) explored gender classification

on the same corpus with a different approach. Yet, they simplified their previous approach by applying simple lexical and syntactic features. They found that females are more likely to use pronouns and males are more likely to use noun specifiers. Their findings are quite interesting because they found that female writing exhibits involvedness while male writing exhibits information.

Singh (20) did a study on gender classification using a conversational speech dataset. He also examined the lexical richness measures based on word-frequencies. Subsequently, he used 8 measures for identifying gender, which are noun, pronoun, adjective, verb, type-token ratio, clause-like semantic units, Brunet's index, which introduced by Brunet in his work at (24), and Honoré statistic, which introduced by Honoré (25). In addition, Sarawgi et al. (21) explored a statistical technique to identify the gender of authors from scientific papers, which are formal writing. In that case, they applied three different types of statistical language models. These statistical models are probabilistic context-free grammar, token-level language models and character-level language models. Nowson et al. (22) also introduced gender classification in the British National Corpus (BNC). The BNC includes fiction writing, newspaper and academic papers. They applied different tag-set techniques including CLAWS tag-set which is a mathematical reduction set, MAXPOST tagger and PENN tag-set. Lastly, they examined the result by applying Heylighen and Dewaele's F-measure, which was introduced by Heylighen et al (26). To date, gender classification from formal writing is still considered as a hot topic to investigate and explore.

In non-traditional contents, many other researchers have been investigated gender classification in a wide variety of works in different online platforms. Among many studies in this area, Nowson et al. (22), Herring et al. (27) and Miller et al. (28) were the first researchers to investigate gender classification in online contents (e.g. web-blogs). In addition, Herring et al. (27) worked on classifying gender in Weblogs. Thus, they applied content analysis to identify the structural and functional properties of blogs. Obviously, their approach was based on the structural characteristics from analyzing the contents, which is not applicable now with the huge generated data from blogs. Likewise, Miller et al. (28) investigated ways to identify gender in blogs. Yet, their approach focused on the formal features of the blogs contents such as comments and links that are included with the topic. Therefore, they found gender differences in the structure and contents of blogs. However, their research did not include any accuracy results to be mentioned here.

In the next year, Nowson et al. (22) explored gender classification in both formal contents such as BNC and hybrid of formal and informal content such as blogs and emails. They showed after applying the tag-set technique that blogs are less contextual than the emails which are less contextual than formal contents of the BNC dataset. As they were the first researchers to work on gender classification in Weblogs, their approach resulted a reasonable accuracy of less than 60% on average. In addition, Herring and Paolillo (29) investigated in-depth gendering in weblogs using two different sets of features called dependent variables, which contains function words, preferential features, POS, n-gram and independent variables, which contains qualitatively for indications features such as first name, nickname, explicit gender

statement (e.g., I am a male). They introduced web interface called Gender Genie. In a like manner, Yan and Yan (30) explored gender classification in weblogs using both traditional features such as n-gram and non-traditional features that they called weblog-specific features such as Word fonts, Punctuation marks, Emoticons, Background color (i.e. only one color is included). Moreover, de Vel et al. (31) investigated predicting gender from text contents of emails. Their approach depended heavily on structural features, style markers and structural characteristics on selecting the attributes such as message tags, signatures and the vocabulary richness. In the same way, Kucukyilmaz et al. (32) introduced a study aimed to predict gender in chat messages (e.g. MSN messenger, ICQ, IRCs, newsgroups). They applied term-based and style-based classification techniques that generated many features that are belong to each gender.

Recently, Mukherjee and Liu (33) investigated gender classification in blog. Peersman et al. (34) also explored gender classification in the Netlog, which is a different platform of non-traditional contents. Mukherjee and Liu (33) proposed two new techniques to improve the accuracy of predicting gender. The first technique is variable length POS sequence pattern that is a style-based feature. The second technique is the ensemble feature selection method which compares different classes of features such as stylistic features, gender preferential features, factor analysis and word-frequencies features. Likewise, Peersman et al. (34) proposed text categorization features that depend on n-gram feature, syntactic feature, semantic feature and lexical feature. Regardless of the outcomes, they faced several challenges for automatic linguistic analysis, such as stems and POS, because of the language that they used.

In general, gender classification in OSNs uses informal text contents where users freely write the way they would like. Therefore, most researchers investigated gender classification using sophisticated models to identify gender in OSNs. In addition, the impact of these works depends on exploring features of the attributes of the contents (e.g., pattern-features, stylistic-features, word-frequencies-features, n-gram-features). However, a handful of researchers used simple and easy models for predicting gender. The first work on gender classification using dataset of one of the OSNs (e.g., Twitter) investigated by Rao et al. (11). They proposed a novel classification algorithm called stacked-SVM-based classification. They also provided three different classification models, which are sociolinguistic feature models, n-gram feature models and stacked feature model, which is the result of combining the previous two models. The accuracy result of their work is around 72% with 1.2 Million features. These classifications depend on simple features such as n-gram features, stylistic features and some statistics of the user profile. Another work on Twitter, by Pennacchiotti and Popescu (35), provided different set of features extracted from profile contents. Thus, these features are derived from an in depth analysis of profile contents, such as content structure features, text content sentiment features, lexical features and other explicit links features that are pointing to outside sources.

Mislove et al. (36) addressed a concrete study on understanding the demographics of Twitter users. They explored the user population in Twitter; they were among the first researchers who provided gender classification derived from the first name. However, Burger et al. (8) addressed in more depth the problem of gender classification in Twitter. Admittedly, they applied different features such as n-gram and word-frequencies on profile's names, profile's nicknames, profile's

description and profile's text content. Their results were quite surprising compared to the other works in this area. Their resulting accuracy is around 91% with 15.5 Million features. Rao et al. (37) investigated again gender classification on Facebook this time instead of their first work on Twitter. In fact, they applied the same feature techniques, which they used previously (11), but with different classification models and dataset contents. Thus, their features include n-gram features and sociolinguistic features.

AlZamal et al. (38) explored gender classification on Twitter using a demographic inference classifier on different features (e.g. word-frequencies, n-gram, stems, co-stems and hash tags). Liu et al. (39) addressed how to identify the gender composition of commuter population. They applied a demographic inference classifier, as they did in their previous work (38), to estimate the gender of the commuter (e.g. cars and bikes). The earliest work on gender classification using first names as feature-based is by Liu and Ruths (40). They applied only first names for gender classification. More importantly, we found that gender classification, in both traditional and non-traditional contents, is investigated using different features and classifiers. A common trend is that most existing methods share a few well-known features such as n-gram, stylistics, word frequencies and lexical analysis. Since most of the works shared some of the features, researchers in gender classification often seek to customize their work for different datasets and to build a custom-made classifiers, which make their work unique in identifying the gender.

In summary, to date most existing approaches to gender classification on Twitter depend heavily on an analysis of text in posted messages, aptly called tweets; however, the strength of the profile characteristics such as first names, user names and colors, that we explored in

the next chapter, is currently unknown. Burger et al. (8) use four different characteristics from a users profile and posts (i.e., first name, user name, description and tweets) for gender classification. Their method results in some 15 million features. Liu and Ruths (40) use only first names for gender classification. Other works for gender classification use only user posts in order to identify gender (38; 39; 11). For instance, Alzamal et al. (38) and Liu et al. (39) applied the n- gram feature model to about 400 profiles and their tweets. In addition, Rao et al. (11) employ the sociolinguistic-feature model, n- gram feature model and stacked model to analyze text sentiment in posted tweets. They have about 1.2 million features. Except for our methods (10; 12), all existing approaches to gender classification on Twitter use word-based n-grams resulting in a huge feature space consisting of unique words and word combinations extracted from tweets. The size of the resulting feature sets is often in the order of many million features. Therefore, our work in classifying gender is different from existing methods because of its simplicity and the range of profile characteristics that we consider. We defined a quantization color-based technique and a phoneme-based technique for reducing the number of features. Our method typically results in a reduction in feature space size by two to four orders of magnitude and provide a reasonable accuracy.

2.2 Location Classification

Given a user profile u_i and its characteristics c_i, we need to classify the user profiles u_i according to their locations. Usually, location based classifications can serve multiple purposes such as advertisement, legal investigations and different social reasons. However, we use these classification to find inconsistent information that leads to detect deceptive profiles.

To our knowledge, there are many works for location classification using a dataset extracted from OSNs (e.g., Twitter) such as (41; 42; 43). Their approach utilizing classification algorithms and machine learning techniques using features such as n-grams, stylistic features, and some statistics on a user's profile. These features are derived from an in-depth analysis of profile contents, such as geo-location (spatiotemporal), content structure features and explicit links pointing to outside sources. A general advantage is that those 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 those methods, our approach to detect deceptive profiles using first spatiotemporal classification and then applying statistical methods to find unreasonable geo-location activities resulting in low computational complexity and a high degree of scalability as our similar approach in (14).

2.3 Deception

In the third part of the literature reviews, we discuss some studies on deception and spam in OSNs. Deception is a process of producing a mental state of belief in something that is not actually true. Generally, deception can be in the contents, identities, personal information and many others. Thereby, deception can be in different forms including lies, equivocation, concealment, exaggeration and understatement.

In recent years, the field of the deception has attracted many researchers as stated by Castelfranchi et al. (44). Thomas et al. (45) addressed some issues behind the black market of Twitter accounts. It is considered as one form of deception (i.e., deception in both the content and the personal information about the profiles as well as deception in having the profile

to follow others not because they attracted them but they get paid to do so). In addition, they investigated Twitter accounts 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 officially working at the Twitter company where they have more access privileges than the rest of the research community. They applied their approach to Twitter contents to distinguish spamming messages from authentic ones while our approach is to identify deceptive profiles (i.e., the person responsible for the information). Prior to this work, many researchers studied spam on different platforms (e.g., emails, OSNs and forums). For instance, Gao et al. (46) explored OSNs to detect and characterize spam campaigns.

Spam classification in email has been investigated in different works. Ramachandran et al. (47) proposed SpamTracker to classify spam. Damiani et al. (48) proposed a decentralized privacy preserving approach for spam filtering. Both previous works showed how hard it is to detect spam, although not impossible. In different platform called forum, Niu et al. (49) evaluated the impact of forum spamming using a context-based approach. Moreover, Shin et al. (50) investigated spam in forum and propose real-time classifying forum spam. This work seems to be inspired by Shin et al's. Previous work (51) which was about analyzing different features of popular forum spammer tools.

Due to the popularity of the OSNs including Twitter, many researchers have analyzed the behavior of profiles in OSNs. Castillo et al. (52) proposed an automatic method for assessing the credibility of Twitter contents. Likewise, Yang et al. (53) investigated the cultural differences in Twitter credibility between two countries. Furthermore, Yang et al. (54) performed an empirical analysis of the cyber criminal ecosystem in Twitter. In addition, Yardi et al. (55) examined and analyzed the differences between fake and legitimate Twitter users while, in particular, Chu et al. (56) proposed a model to classify legitimate users, fake users and a combination of both in Twitter. Moreover, Zhang et al. (57) explored and proposed a framework model to classify between promoting and spam campaigns. Furthermore, Wang (58), Benevenuto et al. (59), McCord and Chuah (60) and Wang (61) investigated the spams in Twitter and how to detect the spams using content-based approach. However, due to limitation and the scope of the research, none of the previous researchers investigated and analyzed the deception in OSNs. In fact, no research to date could answer the following question: is the profile fake or legitimate? In this dissertation, we are going to have a model for automatically detect deception and flag it for further investigation.

CHAPTER 3

PCHARS: PROFILE CHARACTERISTICS GENDER CLASSIFICATION

In this chapter, we explore gender classification using profile's characteristics approach extracted from user profiles alone with no posted text involved. This chapter is the foundation to detect deception in Online Social Networks (OSNs). Our approach in this dissertation to deception detection is based on the results of gender classification utilizing colors, first names and user names appearing in Twitter profiles. We investigated two novel approaches using profile's characteristics for gender classification. The first approach applies *a color normalization-based (i.e., quantization and sorting based*) technique to profile layout colors. The second approach applies *a phoneme-based* technique to profile first names and user names. Our goal is to evaluate profile's characteristics with respect to their predictive accuracy and computational complexity. We will show that we can get good accuracy even with simple features. To detect deception we use the knowledge of the previously-investigated gender classification. 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, that can be used to detect the deception, as we explore this in more details in the next chapter. They also help find inconsistent information from different characteristics and flag *potentially deceptive* profiles. We will go in detail about the deception in the next chapter.

3.1 Say It with Colors: Language-Independent Gender Classification

In this section, we introduce our first work on gender classification. Here we explore in depth language independent gender classification. Our approach, predicts gender using five color-based features extracted from Twitter profiles such as the background color in a user's profile page. This is in contrast with most existing methods for gender prediction that are language dependent. Those methods use high-dimensional spaces consisting of unique words extracted from such text fields as postings, user names, and profile descriptions. Our approach is independent of the user's language, efficient, scalable, and computationally tractable, while attaining a good level of accuracy.

3.1.1 Introduction

Online Social Networks (OSNs) generate a huge volume of user-originated texts. OSNs allow users to share knowledge, opinions, interests, activities, relationships and friendships with each other. Gender classification can serve multiple purposes in these settings. Commercial organizations can use gender classification for advertising. Law enforcement may use gender classification as part of legal investigations. Others may use gender information for social reasons. Here we examine gender classification based solely on color preferences. We specifically present a novel approach for predicting gender using five color-based features extracted from Twitter profile colors (e.g., the background color in a user's profile page) that is.

Methods for gender classification are typically language dependent, not scalable, inefficient, and held offline using high-dimensional spaces. A recent study (9) shows that there are around 78 different languages in Twitter with English as the dominant language. Another study by

Wauters (62) shows that only around 50% of Twitter messages are in English. Our Twitter dataset alone contains 34 different languages. An estimate breakdown of language use in our dataset shows that around 69% users are English speaking with the remaining 31% distributed over 33 languages. In addition, around 20% of the 69% users who set their profiles to be English speaking routinely post texts in different languages than English. Thus, about 45% of users in our dataset use languages different than English for their posts and profiles. Our long-term goal is gender identification in OSNs with an emphasis on accuracy, computational efficiency and scalability of gender predictions. We are especially interested in language-independent methods.

To date, most existing approaches to gender classification on Twitter depend heavily on an analysis of text in posted messages, aptly called tweets; however, the strength of profile colors for gender classification is currently unknown. Most existing research for gender classification on Twitter is language dependent. An existing study for gender classification (8) shows that 66% of users in their dataset use English. Other works for gender classification (38), (39), (11) did not mention the language distribution of their Twitter dataset, which we assume to be in English. In contrast, our dataset contains profiles of users of all ages, languages, and cultures. In particular, Burger et al. (8) used four different characteristics from a user's profile and posts (i.e., first name, user name, description and tweets) for gender classification. Liu and Ruths (40) utilized only first names for gender classification. Alowibdi et al. (12) applied a phoneme-based analysis to characteristics extracted from a user's profile (e.g., first names and user names). Other works for gender classification use user posts and other statistical information, such as

friends and followers, in order to identify gender (38), (39), (11), (10). In general, all existing approaches to gender classification on Twitter use word based n-grams resulting in a huge feature space consisting of unique words and word combinations extracted from tweets. The size of the resulting feature sets is often in the order of many million features (8). On the whole, our work to predict gender from profile's colors is unique and different from existing methods in term of its simplicity, language independence and low computational space and time complexity. In addition, our work is different because of the range of profile colors characteristics that we consider.

We predicted automatically the gender value of users based on their color preferences. 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 (63), (64). Unlike text-based approaches, we used a novel method for predicting gender using five color-based features. Our preliminary results with our data set are quite encouraging. Although we are considering only five color-based features, we can predict gender with an accuracy of 74.2%, a gain of about 24% with respect to a 50% baseline. A key to the success of our gender guessing with colors is our preprocessing of color features using a color normalization (i.e., quantization and sorting) technique that we discuss later on. An advantage of our method is its broad applicability to Twitter users regardless of their language; we use only color-based features to identify gender. In addition, our color-based analysis shows promising results in term of computational complexity compared to other gender-guessing methods, which use a much larger feature set. Our approach utilizes only five color-based features while Burger

et al. (8) and Rao et al. (11) use text sentiment with 1.2 million and 15.4 million features. Our results show that colors alone can provide reasonably accurate gender predictions, even though a substantial number of users we analyzed do not change the default colors provided by Twitter in their Twitter profiles or in other web sites hosting their profiles (e.g., Twitter mobile application). We conclude that colors are a good gender indicator for users who do change the default colors in their profiles. In these cases, we will be able to use colors alone as part of our gender classification methods.

Our main contributions to color-based gender classification are outlined below.

1. We defined a novel, language-independent approach for predicting gender using color-based features. Most other existing methods rely on text, which varies by language.
2. We validated our approach by analyzing different classifiers over a large dataset of Twitter profiles. Our results show that colors alone can provide reasonably accurate gender predictions. In some cases, we can predict gender with compatible accuracy of 74.2%, a gain of about 24% with respect to a 50% baseline.
3. We defined a color quantization and sorting technique, which we call color normalization, for preprocessing colors harvested from Twitter profiles. This technique substantially improves prediction accuracy while also reducing dramatically the size of our feature set. As a result, our color-based analysis has much lower computational complexity than most other other gender-guessing methods, which use much larger feature sets based on text features.

4. We concluded that colors alone are not useful features. However, we found that considering a combination of multiple (five) color selections from each Twitter profile leads to a reasonable degree of accuracy for gender prediction.

This section is organized as follows. In Subsection 2, we described our dataset collection. In Subsection 3, we detail our proposed approach. In Subsection 4, we report our empirical results from different classifiers and we analyze these results Finally, in Subsection 5, we give some conclusions.

3.1.2 Dataset Collection for Gender Guessing from Colors

We chose Twitter profiles as the starting point of our data collection for several reasons. First, Twitter is one of the most popular social networks to date with a huge user community cutting across great many languages, cultures and age groups. In early 2013, Twitter reached 555 million registered users (6). As of today, Twitter states that there are more than 200 million active users producing around 400 million tweets per a day (7). Second, Twitter has all the color attributes that we need to set up the experiment. These attributes are generally public, meaning that they can be accessed and viewed by anyone who requests them. Lastly, Twitter provides a rich Application Programming Interface (API), which supports automatic collection of large data sets.

For our experiments, we chose Twitter profiles as the starting point of our data collection. In Twitters terminology, the followers of a given user U are users interested in reading $U's$ tweets. These users will be notified when U posts a new tweet. Also, the friends of a user V are the users following $V's$ tweets. In general, users can register themselves as followers of

any other user; no permission is required unless the user protects her profile using Twitter's protection features. A new Twitter user must first fill a profile form, consisting of about 30 fields containing biographical and other personal information, such personal interests and hobbies. However, many fields in the form are optional, and indeed substantial portions of Twitter users leave many or all of those optional fields blank. In addition, Twitter's profile form does not include a specific "gender" field, which complicates gender identification for Twitter users. One can choose additional fields that are not mentioned above for gender classification, such as posted tweets; however, we decided to perform gender classification using only profile colors for scalability.

Among many other fields in a Twitter profile, here we are interested in the five fields that allow users to choose different colors for the following items:

1. Background color.
2. Text color.
3. Link color.
4. Sidebar fill color.
5. Sidebar border color.

Users choose their own preferences by selecting colors from a color wheel while editing their profiles. Unlike other OSNs, such as Facebook, Twitter allows users to redesign and change their profiles. In some cases, users chose both a background color and a background picture (from a picture file) for their profiles. In these cases, the background picture overrides the

background color, which is not shown. However, our empirical setup will take into account the background color chosen by a user even if that color is overriden by that user.

We ran our crawler between January and February 2014, subject to Twitter's limitation of less than 150 requests per hour. We started our crawler with a set of random profiles and we continuously added any profile that the crawler encountered (e.g., profiles of users whose names were mentioned in tweets we harvested). Subsequently, we filtered all the profiles with valid URLs. The URL is a profile field that lets a Twitter user 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.

In all, the dataset we used at the time of our study consisted of 169,449 profiles, of which 94,251 were classified as male and 75,198 were classified as female. We considered only profiles for which we obtained gender information independently of Twitter content (i.e., by following links to other profiles). For each profile in the dataset, we collected the five profile colors listed above. We also stratified the data by randomly sampling 150,000 profiles, of which about 75,000 are classified as male and about 75,000 are classified as female. In this manner, we obtain an even baseline containing 50% male and female profiles. Twitter offers 19 predefined designs, including a default design, to each new user joining the social network. Each design defines colors for all five fields. Users can select those designs easily. As of this writing, the color (R=192, G=222, B=237), a light shade of blue, is the default background color for any new Twitter user.

In order to account for the existence of predefined designs in the Twitter user setup, we have considered different subsets of our overall dataset, and we studied each subset independently of other subsets. In addition, we stratified each subset by randomly sampling the profiles, from which we obtain even baselines containing 50% male and female profiles. We specifically considered the following subsets:

T1. This is the entire dataset, A, consisting of 150,000 profiles with a 50% male and 50% female breakdown.

T2. This is dataset A-D, which is the subset containing all collected profiles, except for profiles using the default design with the RGB values of (192, 222, 237) as the background color, denoted by D. D represents 11.4% of dataset A while *T2* represents 88.6%. The base condition is a 50% male and 50% female breakdown.

T3. This is dataset is A-C, which is the subset obtained by excluding C, the subset all profiles that use any of the 19 predefined designs including the default design, from A. C represents around 57% of A while *T3* represents 43%. The base condition is a 50% male and 50% female breakdown. Here we report detailed empirical results about *T3*, since it includes only profiles with custom color choices, and we summarize results for the other datasets.

T4. This is dataset A-B, obtained by excluding from the entire dataset, A, all profiles, B, that use any of the 19 predefined designs as well as black or white as background color. B represents 71.8% of A, while *T4* represents 28.2%. The base condition is still a 50% male and 50% female breakdown.

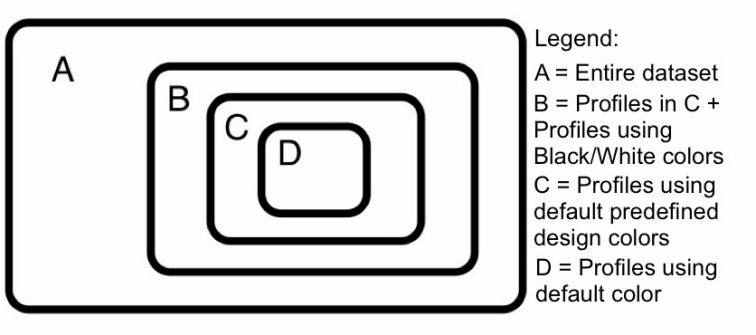

Figure 1. Four Subset of our dataset.

Figure 1 shows the four subsets that we considered for our analyses. Overall, female users are more likely to choose their own layout colors, while male users are more likely to use the default design or one of the other predefined designs.

3.1.3 Proposed Approach for Gender Guessing from Colors

Our algorithm for preprocessing colors before feeding the colors to the classifier is shown in Figure 2 below. First, we harvest colors from user profiles. Next, we apply a color quantization and sorting procedure (i.e., normalization) to reduce the number of colors. The colors are converted from their Red, Green and Blue (RGB) representation to the corresponding HSV (Hue, Saturation, Value) representation. We then sort the colors by their hue and value, and finally we convert them back to RGB. The sorting allows labeling similar colors (e.g., adjacent colors in the sort) by consecutive numbers that we feed to the classifier.

Figure 3 shows the color distribution of profile background colors harvested from profiles in our data set before quantization. Broader stripes denote the relative frequency of background

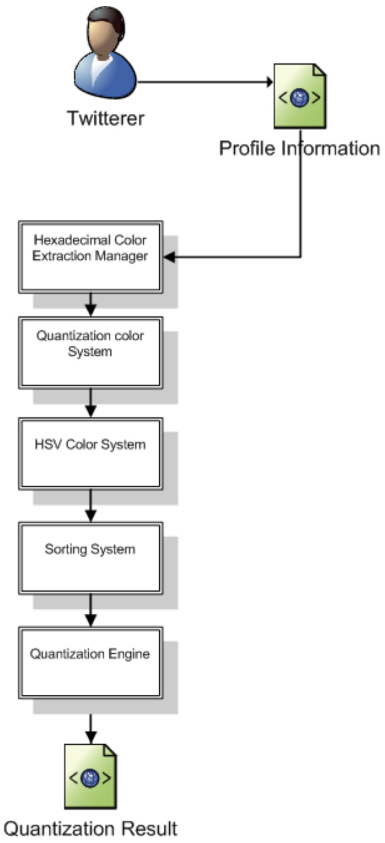

Figure 2. Algorithm for color preprocessing.

color in the profiles that we analyzed. In particular, the broad light blue stripe to the center left of the figure represents the default background color of Twitter profiles. This is the most popular background color, presumably because it is the default color.

Colors harvested from Twitter user profiles are typically specified as a combination of RGB values ranging between 0 and 255. This gives a total of 256^3 colors combinations. Because of the large number of combinations, we use quantization, a compression procedure that substantially

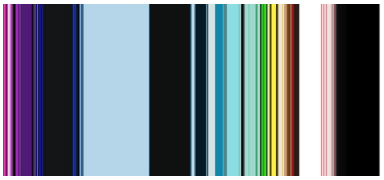

Figure 3. Distribution of profile background colors before applying color quantization in our dataset.

reduces the huge number of colors. Each of the red, green and blue values is shrunk from 8 bits to 4 bits and 3 bits respectively. This technique reduces the total number of color combinations from $256^3 \approx 16 * 10^6$ to just $16^3 = 4096$ colors and $8^3 = 512$ colors, respectively. Each of the original colors we harvested is converted to the compressed color having the least Euclidean distance from the original color. Next, according to the algorithm in Figure 2, we convert each quantized color to the corresponding HSV representation. We use this representation for sorting the colors according to their similarity. First, colors are sorted by their hue; we use values to break ties between colors having identical hues. Figure 4 below shows the 512 colors (i.e., the quantization color procedure of 9-bit RGB) obtained after quantization and sorting.

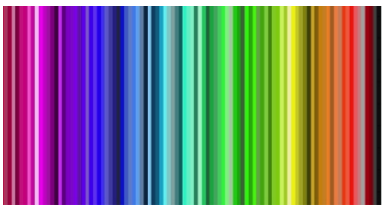

Figure 4. Spectrum of sorted, quantized colors obtained by the color-preprocessing algorithm shown in Figure 2.

The rationale for applying color quantization is that the feature set obtained from straight RGB values would be quite large, a total of $256^{(3*5)}$ cases for 5 color features. A feature set of this size would be mostly unnecessary as most colors are perceptually indistiguishable 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. Next, we investigated the size of each cluster. Larger clusters would lead to smaller features sets; however, larger clusters may also lead to the inclusion of substantially different colors in the same cluster. For this reason, we studied empirically clusters of various sizes and we concluded that clusters grouping 512 colors in each cluster, with 3-bit RGB values per cluster, gave us the highest accuracy results.

We observed empirically that quantization and sorting are beneficial to the accuracy of our gender predictions. In general, our accuracy has improved by up to 15% because of these procedures. Figure 5 shows in 3 dimensions the profile background colors distribution for male and female users, the quantization color centroid and background color distribution for both genders in our data set after applying the quantization color procedure of 9-bit RGB. In brief, our quantization color procedure is a reduction from 24-bit to both 12-bit and 9-bit RGB color representations. We tried both finer and coarser representations for colors and we found that 3 bits per color give us the best prediction accuracy among the options that we considered. We conclude that this representation is a reasonable compromise between the number of colors (i.e., the feature values) that we must consider and the perceptual differences within the resulting color clusters. Color quantization is especially important because we are using a total of 5 color

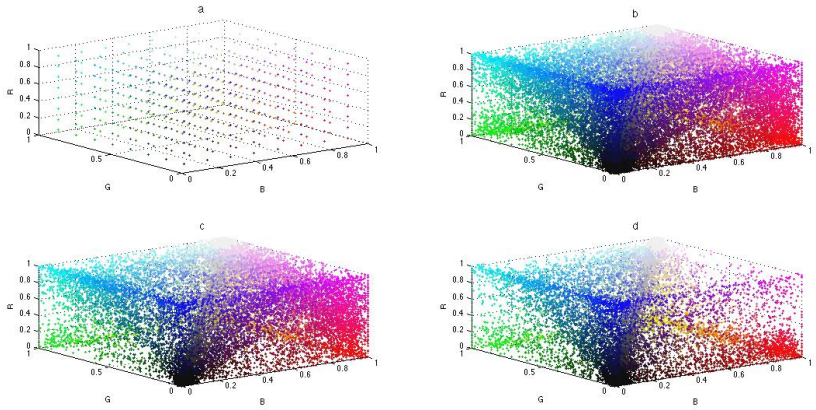

Figure 5. Part (a) shows the centroid of the quantization color procedure; Part (b) shows the color distribution of both genders for the profile background after applying the quantization color procedure to our data set; Part (c) shows the color distribution of the profile background of female users after applying the quantization color procedure to our data set; and Part (d) shows a similar color distribution for male users.

features for each user we analyze. In general, quantization reduces the number of cases (i.e. combinations) for five color-based features from $256^{(3*5)}$ cases to $8^{(3*5)}$ cases.

3.1.4 Empirical Study of Gender Guessing from Colors

In this subsection we evaluate empirically our dataset using different classifiers and we report our findings on gender guessing from Twitter profile colors.

3.1.4.1 Experimental results

We performed four sets of experiments, one for each of the four subsets of our dataset that we mentioned earlier in Figure 1. In each experiment set, we tried many classifiers; different classifiers produced different results. Next, we selected the top classifiers. Here we consider the following four classifiers: Probabilistic Neural Network (PNN), Decision Tree (DT), Naïve

TABLE I

ACCURACY OF GENDER PREDICTIONS FOR DATASET T3 WITH RGB COLORS WITHOUT QUANTIZATION.

	Scores (%)	1 color	2 colors	3 colors	4 colors	5 colors
NB	Precision	59.2	59.1	61.1	62.1	62.2
	Recall	59.2	59.1	61.1	62.1	62.2
	F-score	59.2	59.1	61.1	62.1	62.2
	Accuracy	59.2	59.1	61.1	62.1	62.2
DT	Precision	59.9	61.5	63.7	64.0	64.1
	Recall	58.8	61.5	63.8	64.0	64.1
	F-score	57.9	61.4	63.8	64.0	64.1
	Accuracy	58.8	61.5	63.8	64.0	64.1
PNN	Precision	**62.2**	**65.6**	**66.7**	66.2	**66.9**
	Recall	**61.2**	**65.7**	**66.5**	65.4	65.0
	F-score	**60.5**	**65.7**	**66.4**	63.2	63.9
	Accuracy	**61.3**	**65.7**	**66.6**	64.4	65.0
NB-Tree	Precision	58.6	61.1	64.4	**67.2**	65.2
	Recall	58.3	61.1	64.4	**67.2**	**65.2**
	F-score	57.9	61.1	64.4	**67.1**	**65.2**
	Accuracy	58.2	61.1	64.4	**67.2**	**65.2**

Bayes (NB) and Naïve Bayes/Decision-Tree Hybrid (NB-Tree). We performed a 10-fold cross validation on our data subsets for each classifier (65). In each set of experiments, we trained our classifiers with all five color-based features.

We assessed the effectiveness of color quantization by running experiments with and without color quantization (i.e., using the raw RGB data harvested from Twitter profiles). Table I and Table II report the performance of dataset T3 using different classifiers and color-based features with a 50% baseline. In particular, we choose T3 among the other datasets because T3 is our largest data subset containing only colors chosen by users from the color wheel. The last five columns in the table report results for different numbers of color features. We use the color

TABLE II

ACCURACY OF EMPIRICAL RESULTS FOR DATASET T3 AFTER APPLYING COLOR QUANTIZATION AND SORTING.

	Scores (%)	1 color	2 colors	3 colors	4 colors	5 colors
NB	Precision	59.1	59.0	61.1	61.9	61.9
	Recall	59.1	59.0	61.1	61.9	61.9
	F-score	59.1	59.0	60.9	61.9	61.9
	Accuracy	59.1	59.0	61.1	61.9	61.9
DT	Precision	61.6	67.4	69.1	68.9	68.5
	Recall	61.3	65.7	68.8	68.7	68.3
	F-score	61.2	64.9	68.6	68.6	68.2
	Accuracy	61.3	65.7	68.8	68.7	68.1
PNN	Precision	61.3	66.2	69.1	68.0	66.6
	Recall	61.2	65.4	69.1	66.8	65.5
	F-score	61.1	65.0	69.1	66.2	65.8
	Accuracy	61.1	65.4	69.1	66.8	66.5
NB-Tree	Precision	**68.7**	**69.8**	**72.7**	**72.5**	**73.9**
	Recall	**69.7**	**68.6**	**72.8**	**72.9**	**73.8**
	F-score	**68.7**	**69.9**	**72.9**	**72.5**	**73.9**
	Accuracy	**70.7**	**71.2**	**73.3**	**73.8**	**74.2**

features in the order that we listed previously. Thus, the column with one color feature reports data obtained with the background color alone; the column with two color features reports data for the background color and text color; the next column adds the link color; and the last two columns add sidebar fill color and border color. For each experiment, we report the percentage of correctly identified male users and female users and the overall accuracy.

On the one hand, Table I reports the accuracy of gender prediction without applying the quantization and sorting algorithms discussed above. However, the data in Table II was obtained after applying quantization to Twitter profile colors and sorting the resulting color clusters. As shown in Table I without quantization, the performance of three color-based features roughly

equals the case of four and five features. The best performances in each category are highlighted in boldface. In the case of the PNN classifier, three features actually give better accuracy than four and five features. Also, in the case of the NB-Tree classifier, four features actually give better accuracy than three and five features. In the case of the NB-Tree classifier, four features provide the best accuracy for the RGB Colors. In contrast with Table I, in Table II the accuracy performance increases when using all five color-based features compared to the cases of three and four color-based features.

On the whole, the data in Table I and Table II show that quantization and sorting of colors result in a significant increase in accuracy, especially when all five-color features are used with Naïve Bayes/Decision-Tree Hybrid (NB-Tree) classifier and when three-color features are used with the Probabilistic Neural Network (PNN) classifier. In fact, these two classifiers obtain overall accuracy results of 74.2% and 69.1% with quantization and sorting. Without quantization and sorting these two classifiers achieve only 65.2% and 66.6% accuracy. Modest performance gains are obtained also with the Decision Tree (DT) classifier. In contrast with the other three classifiers, the Naïve Bayes (NB) classifier fails to achieve any gains except the case of the three-color features where it roughly ties its previous performance. In fact, the performance of this classifier drops overall with color quantization and sorting.

Figure 6 shows the accuracy increase obtained by using the color quantization procedure compared to the case of raw RGB colors for each of the four classifiers on dataset *T3*. Part (a) shows the performance of the Naïve Bayes classifier with and without quantization. This is the only classifier that provides slightly better accuracy without quantization than in the case of

quantization. However, the overall performance of the classifier is inferior to that of the other classifiers. Part (b) in Figure 6 shows the performance of the Decision Tree classifier, which yields better accuracy than Naïve Bayes. In this case, color quantization and sorting improve slightly the accuracy of the predictions. The performance of the Probabilistic Neural Network (PNN) and Naïve Bayes/Decision-Tree Hybrid (NB-Tree) classifiers are shown in Part (c) and Part (d) of Figure 6.

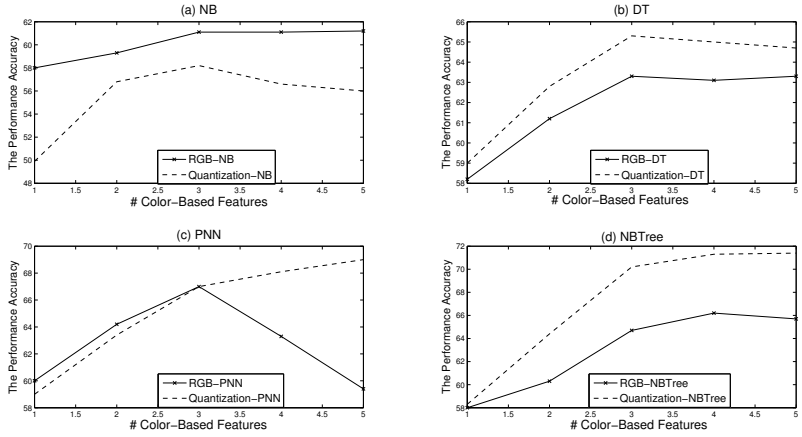

Figure 6. Accuracy of the four classifiers on dataset T3 using different numbers of color-based features.

Table III shows the performance of the four classifiers on all four datasets that we considered after color quantization. Evidently, NB-Tree has the best accuracy on all five datasets with accuracy results consistently above 70% in all four cases. We specifically obtained our best

results with the NB-tree classifier in the *T3* dataset with an accuracy of 74.2% over a 50% baseline of both genders, a gain of about 24.2%. Again, we highlighted in boldface the best classifier.

TABLE III

ACCURACY OF THE EXPERIMENTAL RESULTS FOR THE FOUR DIFFERENT DATASETS WITH COLOR QUANTIZATION AND SORTING.

	Scores (%)	T1	T2	T3	T4
NB	Precision	64.1	63.0	61.9	62.7
	Recall	64.2	63.1	61.9	62.7
	F-score	64.2	63.1	61.9	62.7
	Accuracy	64.3	63.2	61.9	62.6
DT	Precision	69.3.0	69.3	68.5	61.4
	Recall	68.9	69.5	68.3	60
	F-score	69.9	69.4	68.2	60.7
	Accuracy	69.9	69.5	68.1	63.8
PNN	Precision	62.0	67.6	66.6	67.3
	Recall	61.4	65.6	63.5	64.6
	F-score	61.0	64.6	61.8	63.2
	Accuracy	61.4	65.6	63.5	64.6
NB-Tree	Precision	**72.3**	**71.6**	**73.9**	**71.9**
	Recall	**72.0**	**71.4**	**73.8**	**71.4**
	F-score	**72.1**	**71.5**	**73.9**	**71.2**
	Accuracy	**72.3**	**72.0**	**74.2**	**71.4**

An advantage of our approach is that uses only five colors, making it language independent. An additional advantage is that it has a low-dimensional space, resulting in a low computational complexity of our classifiers. In contrast with our method, most existing approaches are language dependent while using high dimensional spaces generated from unique words ex-

tracted from text (i.e. tweets, names, and profile descriptions), and millions of features. As we mentioned before, Burger et al. (8) utilize 15.6 million features with each feature corresponding to a unique word extracted from a tweet. Similarly, Rao et al. (11) use 1.25 million features extracted from tweets.

Figure 7 shows the difference in colors chosen by female vs. male Twitter users. On the top we show popular colors chosen by female users (after clustering); the colors for male users are shown on the bottom of the figure.

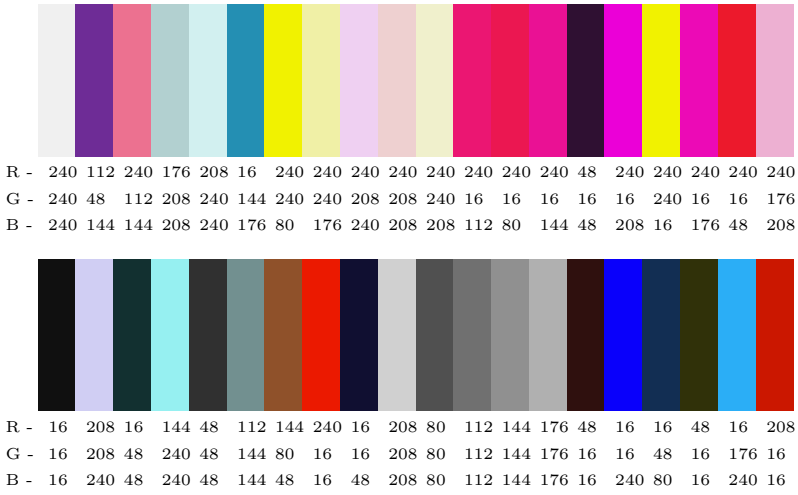

Figure 7. Spectrum of popular colors for female users (top) and male users (bottom).

Figure 8 shows the effects of different training set sizes on the accuracy of the predictions. Similar to Figure 6, the four parts of the figure refer to different classifiers; for each classifier we use color-coded lines to distinguish the number of color features that we consider.

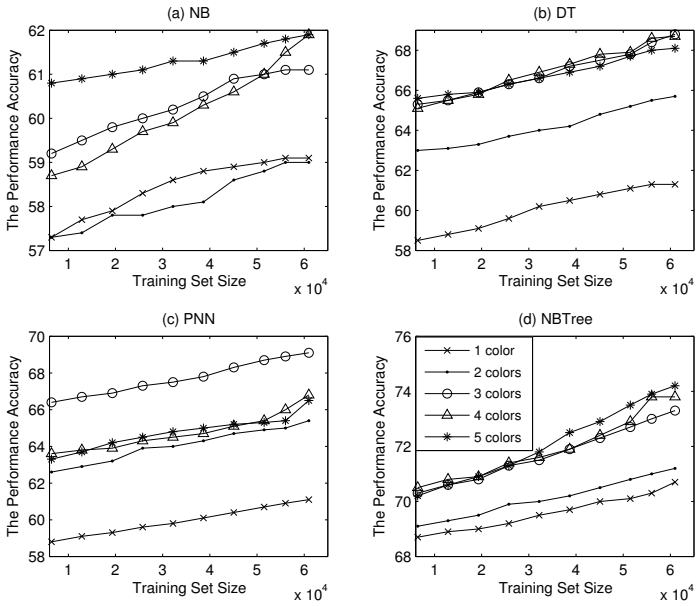

Figure 8. Effects of different training set sizes on accuracy of different classifiers on dataset T3 with different numbers of color-based features.

All diagrams refer to data set T3. In general, the accuracy of our predictions grows linearly in the size of the training sets; larger training sets yield better accuracy results. The four classifiers exhibit similar behaviors with respect to training set size. However, the performance of the classifiers differs depending on the number colors considered. In particular, the PNN classifier does best when three colors are used. Evidently, the inclusion of the sidebar fill color and border color has an adverse effect on the performance of this classifier. The DT and NB-Tree classifiers exhibit similar performance in the case of three, four and five colors. The performance of the DT classifier drops significantly when two colors are used, even more so in the case of one color. The NB-Tree classifier also exhibits a performance drop in the case of two colors and one color; however, this classifier appears to be less sensitive to the number of colors than the DT classifier. Finally, the NB classifier shows the worst performance of the four classifiers we considered; however, this classifier benefits when larger color sets (consisting of 4 and 5 colors) are used. We conclude that the NB-Tree classifier is the most suitable for our gender predictions. Not only does this classifier yield the highest accuracy results; it is also more robust than the other classifier when fewer colors are considered.

3.1.4.2 Threats to validity

There are two main threats to the validity of our study on gender classification. The first threat is our reliance on self-declared gender information entered by Twitter users on external web sites for validation of our predictions. We use this gender information as our ground truth. Evidently, a complete evaluation of all 169,449 Twitter users would be impractical. We manually spot-checked about 10,000 out of the 169,449 profiles in our dataset or about 6.0% of

the dataset. In the cases that we checked by hand, we are confident that the gender information we harvested automatically was indeed correct. The second threat is given by the overall size of the dataset that we could analyze. Although we started from four millions Twitter users, we ended up with just 169,449 users whose gender we could verify independently. This indicates that the size of the training sets was adequate; however, we will continue expanding our data set. Apparently, little will be gained by using larger datasets.

3.1.5 Summary of Gender Guessing from Colors

In summary, we studied gender classification on Twitter. We presented a novel approach for predicting gender utilizing only five color-based features extracted from the profile layout colors. Unlike existing works that use millions of features, we used only five color-based features. Despite the challenging feature-based characteristics for gender classification, we proposed color-based model for gender classification. We applied a color quantization and sorting procedure to the color-based features that compressed the color from 24-bits to 9-bits and produced discrete set of 512 colors. We empirically proved the validity of our approach by examining different classifiers over large Twitter data set collection. Our approach is using an agent with advanced colors preferences to search all profiles and predicting gender. Our empirical studies show that our method is reasonably accurate and highly efficient in terms of computational complexity.

3.2 Pronounce It with Phonemes: Language-Independent Gender Classification

In this section, we introduce our second work on gender classification. Here, we explore profile characteristics for gender classification on Twitter. Unlike existing approaches to gender classification that depend heavily on posted text such as tweets, here, we study the relative strengths of different characteristics extracted from Twitter profiles (e.g., first name and user name in a users profile page). However, our approach is still independent of a user's language because we use a translator on the stream. Our goal is to evaluate profile characteristics with respect to their predictive accuracy and computational complexity. In addition, we provide a novel technique to reduce the number of features of text-based profile characteristics from the order of millions to a few thousands and, in some cases, to only 40 features. We prove the validity of our approach by examining different classifiers over a large dataset of Twitter profiles.

3.2.1 Introduction

In this section, we explore gender identification using only user profiles. Our approach is based on three profile's characteristics for each user profile:

1. Profile first name.
2. Profile user name.
3. Profile colors.

Profile colors include the background color, text color, link color, sidebar fill color and sidebar border color as we discussed earlier. We conducted extensive empirical studies on a large

dataset of Twitter users in order to assess the relative strengths and weakness of these characteristics. In particular, this section will explore gender guessing based on only first name and user name. Our work is different from existing methods because of its simplicity and the range of profile characteristics that we consider. We defined a phoneme-based technique for reducing the number of features. Our method typically results in a reduction in feature space size by two to four orders of magnitude. In the sequel we report our empirical results on different profile's characteristics for gender classification. In particular, we predict automatically the gender value of users based on their profile preferences. Similar to our previous work, 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 (63), (64).

Our main contributions in this section are outlined below:

- We define a new phoneme technique for predicting gender, which sharply reduces feature set size to a few thousands features at the most, and in some cases only 40 features, from several million features.
- We compared empirically different profile's characteristics in order to find the most accurate gender indicators.
- We validated our approach by analyzing different classifiers over a large dataset of Twitter profiles. Our results show that each profile's characteristics can provide reasonably accurate gender predictions.

The remainder of this section is organized as follows. In Subsection 2, we detail our proposed approach. In Subsection 3, we report our empirical results from different classifiers and we analyze these results. Finally, in Subsection 4, we give some conclusions.

3.2.2 Proposed approach for Gender Guessing

Our approach can be summarized as follows:

1. We harvested a large dataset of Twitter profiles.
2. We identified the "ground truth" of a user's gender by following the links from the profiles to other OSNs.
3. We applied the Google Input Tools (GIT) to convert the characters of different languages to characters in English language.
4. We converted first names and usernames to phoneme sequences.
5. We trained, tested and validated our gender predictions using different classifiers.

In Step 1 we harvest first names and username.

Twitter profiles do not include an explicit gender field. Thus, in Step 2 we identify Twitter profiles with an external link to another profile (e.g., a Facebook profile) for the same user. If the other profile includes an explicit gender declaration, we use that declaration as the ground truth for the gender of that user.

In Step 3, we convert the alphabet of different languages than English (e.g., Japanese, Chinese, and Arabic) to characters in English with GIT. For instance, GIT converts such Japanese names as "信浩", "貴志" and "一幸" to Nobuhiro, Takashi and Kazuyuki

respectively. In a similar vein, Arabic names "عبدالرحمن" and "عمر" will be converted to Abdulrahman and Omar.

In Step 4, 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". Our phoneme set contains 40 phonemes that may carry three different lexical stresses, namely no stress, primary stress and secondary stress (13). We employ the LOGIOS lexicon tool for converting names to phonemes (66). In this way, we reduce number of features from the order of millions, as in the work of Burger et al. (8), to only around few thousand features, considering all phoneme combinations, and some cases only 40 features. We apply the n-gram analysis to the resulting phonemes (67). In Subsection 3, we will compare our phoneme-based method with the word-based (traditional) n-gram feature model used by other authors.

Finally, in Step 5 we analyze our feature sets using KNIME. In general, we observed empirically that the phoneme technique is beneficial to the accuracy of our gender predictions. In general, our accuracy has improved by up to 32.5% from a 50% baseline because of this procedure. 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.

3.2.3 Empirical analysis of Gender Guessing

In this subsection we evaluate empirically our dataset using different classifiers and we report our findings.

3.2.3.1 Dataset Collection

Upon registering on the Twitter web site, a new Twitter user is presented with a form requesting various kinds of demographic information. However, many of the fields in the form are optional, and indeed a substantial portion of Twitter users leave many or all of those optional fields blank. In addition, Twitter profiles do not include a specific gender field, which complicates our gender identification efforts.

In a Twitter profile for this experiment we are interested in the following two optional fields:

1. Profile first name.
2. Profile user name.

Users can edit, change and remove their own preferences. We ran our crawler between February and June 2013. Here in this experiment, we follow our previously mentioned approach in building our ground-truth-based knowledge by filtering all the profiles with valid URLs to reach an explicit gender field.

In all, the dataset that we considered for this study consisted of 194,293 Twitter profiles, of which 104,535 are classified as male and 98,758 are classified as female. We considered only profiles for which we have obtained gender information independently of Twitter content (i.e., by following links to other profiles). For each profile in the dataset, we collected the two fields listed above. We also stratified the data by randomly sampling 180,000 profiles, of which about 90,000 are classified as male and about 90,000 are classified as female. In this manner, we obtain an even baseline containing 50% male and female profiles.

Information harvested from Twitter was further processed in various ways. On the one hand, for the colors we used color quantization and sorting as we discussed earlier in this chapter. On the other hand, first names contained in profiles harvested from Twitter undergo a series of pre-processing steps. These steps include the removal of leading and trailing white space, as well as the deletion of last names, numbers, punctuation, and stop words (e.g., Dr, Doc, Mr, Ms). The outcome of this step is first names alone, which can then be used for phoneme sequence generation. The next stage involves computing the phoneme sequences for the preprocessed first names and user names. Phoneme sequences are obtained from LOGIOS and, for profiles in different alphabets, GIT. Next, we generate n-grams of the phoneme sequences. These n-grams and colors are the feature set input to the classifier. The classifiers empirical results are reported below.

3.2.3.2 Empirical results

We performed different sets of experiments in an effort to assess the relative strengths of the various classifiers. As an additional goal, we wished to assess the effectiveness of our techniques for preprocessing our data set. For this reason, we ran experiments in which we did not transform first names and user names into phoneme sequences. In these cases, we generated n-grams directly from the first names and user names harvested from Twitter. We compared the results obtained in this manner with results obtained by transforming those names into phoneme sequences. These results are shown in Table IV and Table V.

We performed different sets of experiments by applying three different classifiers, namely Naïve Bayes (NB), Decision Tree (DT) and Naïve-Bayes Decision-Tree (NB-Tree) hybrid. In

all cases, we performed a 10-fold cross validation on data subsets for each classifier (65). In each set of experiments, we trained our classifiers both with the phoneme-based feature set and with word-frequency based feature set.

We note at the outset that an advantage of the phoneme-based feature set is the reduction in the number of features to a minimum of 40 features, the phoneme set obtained from the LOGIOS lexicon tool, from millions of features in the word-frequency based method. This reduction results in low computational complexity and a high degree of scalability for the phoneme-based feature set. As we will see, we also obtain reasonably high accuracy results—in the best case 78.5%—even with the small feature set (40 features). An additional advantage of the phoneme-based feature set is language independence, as we obtain phonemes from any language and alphabet system in our dataset. In contrast with our phoneme-based method, the n-gram approach based merely on word frequencies is language dependent while using high dimensional spaces with millions of features generated from unique words extracted from text (i.e. first names and user names).

For first names and user names, we compared our phoneme-based technique with the word frequency method. Without phonemes, we reached around half million features. The size of this feature set is consistent with the results re-ported by Burger et al. (8). When using phonemes, the maximum theoretical feature set size for 3-grams is $40^3 = 64,000$ features because there are 40 phonemes. However, the largest feature set size that we have observed in practice is around 16,000 phonemes because many phoneme combinations never occur in a 3-gram. Figure 9 shows

Figure 9. Cloud tagging of phonemes of male users (left-hand side) and female users (right-hand side).

the cloud tagging of phoneme names for both male and female users. Phonemes in the darker shade of blue are used more frequently than the case of the lighter shade.

When using word frequencies, we conducted experiments with 1-gram through 5-gram features. When using phoneme-based features, we conducted experiments with 1-gram through 3-gram features. Table IV shows our empirical results for both cases. Entries labeled "NA" refer to cases that were not applicable in our experimental setup. For instance, the name John can be represented as the 3-phoneme sequence "JH AA N" which supports at most a 3-gram analysis. The highest accuracy we obtained was 82.5% in the case of 3-gram phoneme-based features, an improvement of 32.5% with respect to the baseline. In this case, our feature set size was about 16,000 features. The worst-case accuracy for the phoneme-based feature set was

TABLE IV

ACCURACY OF GENDER PREDICTIONS FOR PROFILES' FIRST NAME.

	1-gram	2-gram	3-gram	4-gram	5-gram
Without phonemes (n-gram applied to characters of first names)					
NB	NA	65.3	67	69.2	75.1
DT	NA	68.2	69.3	72.0	76.3
NB-Tree	NA	**69.3**	**70.7**	**74.0**	**78.3**
With phonemes (n-gram applied to set of phonemes)					
NB	65.2	65.3	66.0	NA	NA
DT	**78.5**	**79.2**	**82.5**	NA	NA

predictably the 1-gram case. Even so, we achieved 78.5% accuracy, an improvement of 28.5% over the baseline with only 40 features.

Our accuracy results for phoneme-based gender classification are in line with the methods of Burger et al. (8) and Liu et al. (40). Those methods obtained an improvement accuracy of 34%, with half a million features, and of 20% with an unknown number of features. Our big advantage is that we obtained accuracy results comparable to their best results with about 16,000 total features. A portion of these features included 10,500 male and female first names available from the US Census Bureau (68).

Table V indicates that phonemes also work well for user names. The best results were obtained with 3-gram phonemes resulting in 75.2% accuracy and a feature set size of 1,235 features. Evidently, our improvement accuracy over the 50% baseline is 25.2%, which is lower than the accuracy of Burger et al. (8) by about 2%. Thus, the method of Burger et al. (8) is slightly superior to ours with respect to accuracy performance whereas our method is superior to theirs in terms of computational complexity.

TABLE V

ACCURACY OF GENDER PREDICTIONS FOR PROFILE'S USER NAMES.

	1-gram	2-gram	3-gram	4-gram	5-gram
Without phonemes (n-gram applied to characters of user names)					
NB	NA	55.3	56	57.2	58
DT	NA	**55.7**	**56.9**	**58.2**	**59.6**
NB-Tree	NA	53.2	54	56	58
With phonemes (n-gram applied to set of phonemes)					
NB	55.2	56	55	NA	NA
DT	**68.5**	**70.2**	**75.2**	NA	NA

Similar to Table IV, the data in Table V shows a significant improvement in accuracy for the phoneme-based feature set with respect to the word-frequency based set. The improvement in accuracy is quite significant considering also the lower computational complexity and language independence of the phoneme-based feature set.

On the whole, the accuracy results achieved with first names are higher than the accuracy results obtained with colors and user names. The accuracy of colors and user names are comparable to each other. In the future, we plan to explore accuracy results obtained by combining all three profile characteristics. In addition, we observe that our phoneme-based n-gram analysis benefits from the addition of features in the 2-gram and especially the 3-gram analysis with respect to the 1-gram analysis. We also note that phonetic analysis of first names and user names can significantly increase our accuracy results. See, for instance, the data relative to the DT classifier in Table IV and Table V. Finally, we observe that different classifiers work best with different feature characteristics. In the case of colors, the NB-Tree classifier shows the

highest accuracy results. In the case of first and last names, it is the DT classifier that shows the highest accuracy results.

3.2.3.3 Threats to validity

There are two main threats to the validity of this study. The first threat is our reliance on self-declared gender information entered by Twitter users on external web sites for validation of our predictions. We use this gender information as our ground truth. Evidently, a complete evaluation of all 194,293 Twitter users would be impractical. We manually spot-checked about 5,000 out of the 194,293 profiles in our dataset or about 2.5% of the dataset. In the cases that we checked by hand, we are confident that the gender information we harvested automatically was indeed correct. The second threat is given by the overall size of the dataset that we could analyze. Although we started from four millions Twitter users, we ended up with just 194,293 users whose gender we could verify independently. This indicates that the size of the training sets was adequate; however, we will continue expanding our data set.

3.2.4 Summary for Gender Guessing Utilizing First Names and User Names

In summary, we empirically studied gender classification on Twitter using different profile characteristics such as first name and user name. Also, we presented a novel approach to predict gender utilizing phoneme-based features extracted from profile first names and user names. In addition, we applied both finer and coarser representations for first names and user names. The main advantage of our gender-classification methods is that they achieve good accuracy results, despite sharp reductions in computational complexity with respect to alternative approaches. Our methods also have broad applicability to different languages and alphabet sets than English.

CHAPTER 4

DETECTING DECEPTIVE INFORMATION IN TWITTER ABOUT USER GENDER

In this chapter, we propose a novel approach for detecting deceptive profiles in OSNs. Our ultimate goal is to find deceptive information about user gender. We specifically define a set of analysis methods for detecting deceptive information about user genders 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. Our methods are quite scalable because we avoid the analysis of text messages, which typically involves high computational complexity; however, we cleverly applied a number of preprocessing methods to raw Twitter data in ways that significantly enhanced the accuracy of our predictions. Subsequently, we apply Bayesian classification algorithms to Twitter profile characteristics (e.g., profile layout colors, first names, user names) to analyze user behavior. We established the overall accuracy of each gender indicator through extensive experimentations with our crawled dataset. Based on the outcomes of our approach, in many cases we are able to detect deceptive profiles about gender with a reasonable level accuracy.

4.1 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 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 us to be confident about having a healthy and trusted relation at 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, currently there is no work on detecting deception in gender information, as in this chapter, and location information, as in the next chapter, posted in OSNs.

The long-term objective of this research is to flag automatically deceptive information in user profiles and posts, based on detecting inconsistencies in those profiles and posts. In this dissertation, we focused specifically on the detection of inconsistent information involving user gender and the detection of conflicting spatiotemporal information involving user location. In the sequel, we discuss separately our two approaches for detecting deception about gender and location. We have two distinct datasets where for each of the two methods and we have applied different analysis methods to the two datasets.

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

1. We harvested 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) a number of colors chosen by Twitter users for their profiles, (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 harvested from Twitter profiles. Profile preprocessing significantly improved our ability to predict the gender of a Twitter users from the features that we had harvested.

3. We independently established the accuracy for each feature (i.e., profile colors, first names, and user names) at predicting the gender of a Twitter user by conducting extensive experimentations with Twitter profiles.

4. We defined a Bayesian classifier seeking to identify Twitter users whose profile characteristics conflict with the self-declared gender information harvested from Facebook. We have 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 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. In addition, these characteristics also help find 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 are 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.

On the whole, our preliminary results with our datasets are quite encouraging. We can identify deceptive information about gender with reasonable accuracy. In addition, our methods use a relatively modest number of profile characteristics 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 to 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.

The remainder of this chapter is organized as follows. Section 2 gives some background information. In Section 3 and 4, we describe our dataset collection. In Section 5 and 6, we extensively describe the deceptive profiles about gender and also we report our empirical results. Finally, in Section 7, we give some conclusions.

4.2 Background and Rationale

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 other's attention. Therefore, dealing with lying people has become part of our daily life. According to one study, as many as 31% of users in OSNs provided false information to be safe online (4). Also, in another study, only 20% of people surveyed declared to be honest about information that they provided online (69). According to yet another study, 56% of teenager provided false information in their profiles in order to protect themselves from undesirable attention (5). 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 (3). The interested reader is referred elsewhere for additional detail about the forms of deception (2). 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.

The foundation of our approach for detecting deception about gender was previous works in gender classification (12; 10). 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 VI, as we explained in chapter 3 of this dissertation.

4.3 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 data collection for several reasons that were mentioned in chapter 1. 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. 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.

4.4 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 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.

4.5 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; however, in this dissertation 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.

4.5.1 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.As we described in chapter 3, 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 (64; 63).

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 $2^{8*3} = 2^{24}$ 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 $2^{5*3} = 2^{15}$ 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 harvested. 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 (13). 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% to 82.5% because of this preprocessing stage. We are quite encouraged that not only we improved the accuracy of our gender predictions; 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 VI 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 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 (12). Although accuracy results vary depending on the characteristics being used, the data in Table VI show significant improvements over the 50% baseline for all the characteristics, which is quite encouraging.

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

$$m = \frac{w_f \cdot s_f + w_u \cdot s_u + w_c \cdot s_c}{w_f + w_u + w_c} \tag{4.1}$$

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 (Equation 4.1) 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,

TABLE VI

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%

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 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. We plan to explore alternative approaches to deception detection within our framework in order to include these users in our analyses.

4.6 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 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).

4.6.1 Empirical evaluation of feature relevance in Twitter

Table VII 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 $\mu = 0.5013$ 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 Equation (Equation 4.1) above.

TABLE VII. Accuracy results in gender predictions obtained by using different profile characteristics from Twitter profiles.

	Strong female	Weak female	Neutral	Weak male	Strong male
Index range	$0 \leq m \leq 12.3$	$12.3 < m \leq 31.1$	$31.1 < m \leq 68.9$	$68.9 < m \leq 87.7$	$87.7 < m \leq 1$
No. of profiles	2,673	30,493	109,562	30,717	1,155
Pot. deceptive	—	2,677	—	3,779	—
Likely deceptive	59	—	—	—	18

Table VII 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 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 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, about 33.3%, 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 (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. These are clearly strong results.

4.6.2 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 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.

Previously we defined a phoneme-based method for enhancing the reliability of first names and usernames as predictors of gender (12). 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 VI. These results further confirm the greater accuracy of Facebook names as gender predictors with respect to first names extracted from Twitter.

4.6.3 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 data set. These profiles cover approximately 10% of

all potentially deceptive profiles in our dataset, excluding profiles that were deleted between the time the profiles were harvested 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 harvested 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 VIII 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 cor-

TABLE VIII

OUTCOMES RETURNED BY EACH EVALUATOR FOR POTENTIALLY DECEPTIVE, TRENDING MALE PROFILES.

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

responding 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.

Similarly, Table IX 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

TABLE IX

OUTCOMES RETURNED BY EACH EVALUATOR FOR POTENTIALLY DECEPTIVE, TRENDING FEMALE PROFILES.

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

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.

On the whole, 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 X

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	

Table X 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 the 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, whereas evaluator D tended to use the "unclear" outcome in such cases. (See Table VIII and Table IX.)

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.

4.7 Summary for Detecting Deceptive Information About User Gender

Our ultimate goal is to find inconsistent information in online social networks about user gender in order to detect deception. In particular, we defined a set of analysis methods for that purpose in Twitter. Also, we apply Bayesian classification algorithms to Twitter profile characteristics (e.g., profile layout colors, first names, user names) to analyze user behavior. Therefore, in this study, we presented frameworks for detecting deception about gender information. In addition, we reported preliminary empirical results with a strategy for attaining this goal within the frameworks. Through extensive experiments, our current results show considerable promise for our frameworks. Our empirical experiments obtained by applying our algorithms to multiple datasets showed promising results.

CHAPTER 5

DETECTING DECEPTIVE INFORMATION IN TWITTER ABOUT USER LOCATION

In this chapter, we propose another novel approach for detecting deceptive profiles in OSNs. Our goal is to find deceptive information about user locations. In particular, we define a set of analysis methods for detecting deceptive information about user locations in Twitter. First, we collected a large dataset of Twitter profiles and tweets. Next, we cleverly applied a preprocessing method to raw Twitter data in order to facilitate analysis of location information extracted from geo-tagged tweets. This method consists of a K-means clustering algorithm applied to the geographical coordinates of tweets from each user. Subsequently, we apply a Bayesian classification algorithm to preprocessed spatiotemporal information to analyze user behavior. We established the overall accuracy of each location indicator through extensive experimentations with our crawled dataset. Based on the outcomes of our approach, in many cases we are able to detect deceptive profiles about location with a reasonable level accuracy.

5.1 Introduction

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

1. We harvested 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, (2) spatial information, and (3) location.

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 at predicting the location of a Twitter user by conducting extensive experimentations with Twitter profiles.
4. We defined a Bayesian classifier seeking to identify Twitter users whose profile tweets characteristics contain conflicting information. We have 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 used a dataset of about 35,000 profiles that we harvested between March and April 2014. We conducted a spatiotemporal analysis of postings (i.e., tweets) containing geotagging 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 spatiotemporal information extracted from tweets can provide reasonably accurate predictions of the users' locations. 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 quite encouraging. We can identify deceptive information about 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 about location.
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 location predictions.
4. We found that considering a combination of multiple profile's characteristics and posts from each Twitter profile leads to a reasonable degree of accuracy for detecting the deception about location.

5. We defined methods for identifying Twitter users containing deceptive information about location.

The remainder of this chapter is organized as follows. Section 2 gives some background information. In Section 3 and 4, we extensively describe the deceptive profiles about location and also we report our empirical results. Finally, in Section 5, we give some conclusions.

5.2 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 activities and profile spatiotemporal activities). This can be done by using knowledge from users' activities. People nowadays periodically edit, change and post their information using geo-tagged tweets. Thus, analysis of user information and geo-tagged tweets that come with spatiotemporal information can provide trends of behavior leading to the detection of deception (i.e., lying about gender and location). Our goal is to investigate and determine which indicators is going to be considered for that purpose. In this chapter, we provide novel gender-based and location-based approaches that rely on both publicly-available information contained in Twitter user profiles and on geo-tagged tweets with spatiotemporal infomration.

In the case of geotagged tweets, we know the exact coordinates (i.e., longitude and latitude) of the user from the posted tweets. Therefore, we treat any posted tweet that comes with coordinates as a visit made by this user since this recorded coordinates information come directly from the client (e.g., a mobile device) used to post the tweet. We then apply classification algorithms (e.g., Bayesian classifier) and machine learning techniques (e.g., K-means algorithm)

to analyze user activities in order to flag for potential deception users' profiles with unreasonable visits.

5.2.1 Why Does Detecting Deception About Location Matter?

In a previous published report (14), we explained in detail the background and rationale for detecting deception about gender. Here, we discuss the background and rationale for detecting deception about location. Posting geo-tagged text with geo-location (i.e., spatiotemporal information) is considered as part of communication to others. First, it is relatively easy to disguise someone's location using such services as *Hotspot Shield* (70). Second, 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, a legitimate concern, others may lie about their location to buttress lies about trips that they took or their physical whereabouts. Analyzing geotagged tweet 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.

5.3 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.

For instance, while conducting this research we discovered that some Twitter users lied about visiting exotic places to gain popularity among their Twitter readership. One such user gave his Twitter account information to a friend visiting a foreign location in order to make it believable that the original user was actually traveling!

We treat any efforts at 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); about 50% of the collected tweets are geo-tagged. In the case of geo-tagged tweets we know the exact coordinates (i.e., longitude and latitude) of the Twitter client (e.g., a user's mobile phone) and the time when the tweet was posted. Therefore, checking all pairs of 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 (Saudi Arabia) for our study because this author was 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. This approach gave us an additional advantage in that the Saudi Tourist Information and Research Centre (STIRC) regularly posts information about travel destinations of Saudi residents during their holiday break (71). Thus, we were able to validate our findings about travel destinations of Saudi residents during Spring break against information provided by the STIRC. We found that the destinations we mined in our population in fact match the information posted by the STIRC.

5.4 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

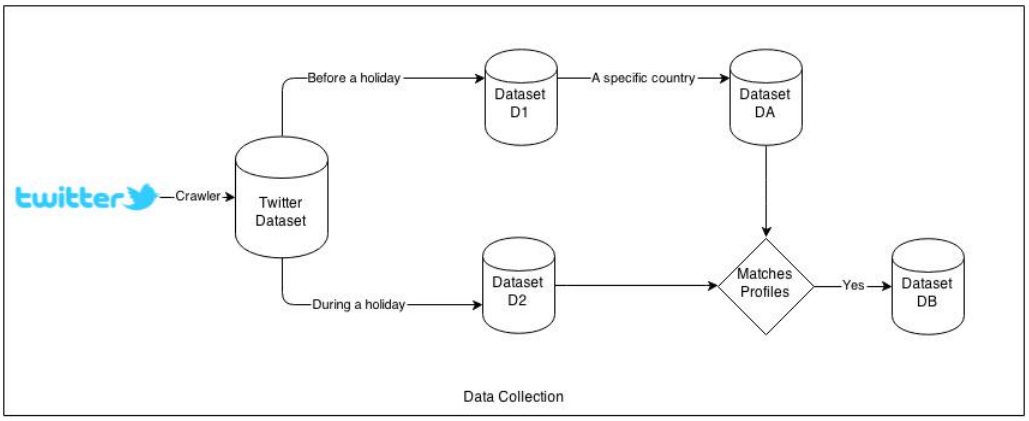

Figure 10. The flow information for the dataset collection

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.

In all, 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.

As with our study on gender, we used only publicly available Twitter data, in this case, to find out where people spent their vacations for a particular country and holiday. 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 geotagged 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 analyses 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 XI 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

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

than Saudi Arabia and the undefined country. Figure 10 shows the flow information that we followed in creating datasets *D1* and *D2*.

TABLE XI

THE TABLE SHOWS NUMBER OF USERS VISITS TO EACH COUNTRY DURING THE SPRING BREAK OF MARCH 2014.

# 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

5.4.1 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 consisting of 600 Million tweets. We specifically considered about 150 Million geo-tagged tweets world-wide. We collected all the coordinate locations of those tweets (i.e., latitude and longitude). Next, we applied k-means clustering to locations in Arabic speaking countries. 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 results for travel originating in Saudi Arabia in Figure 11. 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 analyses is based on these two definitions.

Figure 11. Where did the Saudis Spent the Spring Break of 2014.

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 back our findings about deception into our classifier to train the classifier for future analyses 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.

5.4.2 Empirical results

There are two ways for computing the trending factors that leads to detect deceptive profiles with respect to location. In this subsection, we explore the two approaches to detect deceptive profiles about location.

5.4.2.1 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 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 XII. Table XII shows the user profiles who visiting 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.

TABLE XII

THE TABLE SHOWS THE NUMBER OF PROFILES VISITING DIFFERENT COUNTRIES WITHIN A SHORT PERIOD OF TIME.

# of Countries	# of Profile Visits
1	34132
2	1487
3	143
4	8
40	2
47	2
5	2
206	1
8	1
10	1
6	1
25	1
55	1
7	1
14	1
22	1
86	1
51	1
13	1

From Table XII, we have 1,656 users out of 35,788 unique users having visited more than one country during the holiday break. These 1,656 users made 4,142 visits total. In some cases, those users showed conflicting and impossible geo-location activities, for instance, requiring travel at impossible speeds.

Furthermore, Table XIII 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 1,656. Those flagged potentially and likely deceptive profiles, shown in the Table XIII, were further investigated manually by following the approach we explained earlier.

In addition, for those 1,656 users, we crawled more of their geo-tagged tweets and information. We have collected around 3 Million tweets of which around 2.5 Million tweets contain geo-location information and the others come without geo-location information. Then, we checked if there is any conflicting spatiotemporal information. After that, we compare the

TABLE XIII

ACCURACY RESULTS IN DETECTING DECEPTIVE PROFILES OBTAINED BY USING SPATIOTEMPORAL LOCATION-BASED APPROACH THAT APPLIED TO TRAVELER WHO TRAVEL TO MULTIPLE FOREIGN COUNTRIES.

	Neutral	Potential deceptive	Likely deceptive
Number of profiles	34,132	1487	169
Not deceptive	—	751	2
Not sure deceptive	—	308	15
Deceptive	—	428	152
The precision	—	28.8%	89.9%

detected locations with all the locations that the profile holder visited at another time. One Naïve way to compute a statistical representation of deceptive profile is to compute speed and time as: Euclidean distance as the following formula:

$$Deceptive_{location} = \frac{Distance(location1, location2)}{Interval(time2 - time1)} \tag{5.1}$$

Given this computed path speed if it is conflicting (i.e., if it is 500 miles/hour as it is in our case), then, it is a potential deceptive profile about location. Indeed, we verified the profiles that we identified and reported them in Table XIII.

5.4.2.2 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 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 because spending the holiday break in such countries is highly unlikely. 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 $DC_{discourage-countries}$. 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 XIV shows the accuracy results in detecting deception by using the top-10 discouraged countries.

In conclusion, we are only including the top-10 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.

TABLE XIV

ACCURACY RESULTS IN DETECTING DECEPTIVE PROFILES OBTAINED BY USING SPATIOTEMPORAL LOCATION-BASED APPROACH THAT APPLIED TO TRAVELER WHO TRAVEL TO DISCOURAGED COUNTRIES.

	Neutral	Potential deceptive
Number of profiles	35,756	32
Not deceptive	—	1
Not sure deceptive	—	2
Deceptive	—	29
The precision	—	90.0%

5.4.3 Discussion

In this section, we have investigated the deception about location. 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 we faced during this investigation.

5.4.3.1 Validation compares to official

Before setting up our experiments, we need to find a way in which we are able to validate our findings about travel destinations of Saudi residents during Spring break. This way should justify our mining travel destinations information about the Saudi residents during their holiday break.

We have confirmed travel destinations of users in Saudi Arabia based on a study conducted by the Saudi Tourist Information and Research Centre (71). This study was published by the SABQ Online Newspaper (72). 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 (73), 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.

5.4.3.2 Validation uses profiles 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 XIII.

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 ither 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 minutes. Figure 12 shows the mentioned profiles after we deleted the profile identity for privacy protection.

5.4.3.3 Challenges

There are many challenges, here, we are experiencing with the dataset collection and validation. One of the challenging is that some of geo-tagged tweets have not enough geo-location information which make it a bit difficult decision for the weighted spatiotemporal features indications. For example, some geo-tagged tweets information linked to undefined coordinates information. Thus, the spatiotemporal features indications must be interchanged dynamically

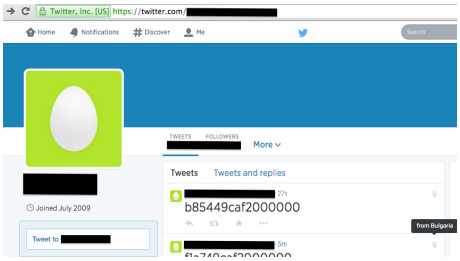

Figure 12. an example of potential deceptive profile about location.

based on the available information. Another challenge is that some of the profile's settings are edited periodically by the owners. Therefore, we collected enough geo-tagged tweets information for a profile at the one time, but, on the other time, we have different geo-tagged tweets information that belong to the same profile. For example, in some cases, we found geo-tagged tweets information that we collected and stored in our database *DB* come with geo-location information. However, because some profile users are editing their profile settings from time to time, we found some of geo-tagged tweets information that we collected and stored in different time in our database *DB* come without geo-location information. Therefore, we have excluded those types of profiles.

5.5 Conclusion

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 in Twitter. Also, we apply Bayesian classification and K-means clustering algorithms to Twitter profile characteristics (e.g., profile layout colors, first names,

user names, 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 frameworks. Through extensive experiments, our current results show considerable promise for our frameworks. 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.

CHAPTER 6

SUMMARY AND FUTURE WORK

6.1 Summary

The long-term objective of this dissertation is to flag automatically deceptive information in user profiles and posts, based on detecting inconsistencies in a user's profile and posts. In particular, we focus on detection of inconsistent information involving user gender and conflicting spatiotemporal activities involving user locations. Therefore, we discuss separately our two approaches for detecting deception about gender and location. We have two different datasets where for each of the two, we have applied two different set of experiments.

On the one hand, we studied the effectiveness of the profiles characteristics for detecting the gender of Twitter users with dataset that we harvested between January and February 2014. Our approach to deception detection is based on our previous results on gender classification utilizing color quatization and sorting (i.e., normalization), phoneme based analysis for first names and user names. To detect deception about gender, we use results from gender classification whereby each user profile in our dataset has a link to a Facebook page in which users declare explicitly their gender. Therefore, 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. In addition,

these characteristics help in finding inconsistent information about the gender from different characteristics and flag for potential deceptive profiles.

On the other hand, we have studied the effectiveness of spatiotemporal activities for predicting the location of Twitter users with a different dataset that we harvested between March and April 2014. We used publicly available Twitter data from that time period to find out where the people spent their vacation for particular country and particular holiday. We have explored geo-tagged tweets that come with geo-location activities for a specific group of people. In particular, we selected Saudi Arabia as a source location and the spring break holiday in March, 2014 as a holiday for this study to find unreasonable geo-location information. The outcome of this study is that such spatiotemporal activities by a user for her profile can provide reasonably accurate predictions of the users' locations. These activities also help predict the deceptive profiles based on spatiotemporal features.

Our preliminary results with our datasets are quite encouraging. On the one hand, when used in gender and location features combination, we can identify deceptive information about gender and location with reasonable accuracy. On the other hand, our approach uses 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, because our approach combines a good accuracy and language independence with low computational complexity. Through our analyses, we have identified several thousands, *potentially deceptive* and *likely deceptive* profiles. We manually inspected likely de-

ceptive profiles, as we report below, and found that a large proportion of those profiles were indeed deceptive.

On the one hand, for the gender based approach in detecting the deception, we have identified 4% of the 174,600 profiles collected as potentially deceptive profiles. Also, we have identified 77 profiles of the 174,600 profiles analyzed as likely deceptive profiles. Therefore, we manually inspected the likely deceptive profiles that deemed to have higher probabilities to be deceptive and found that a large proportion of those profiles (about 42.85%) were indeed deceptive. Manual inspection was inconclusive also to the potentially deceptive profiles and we found that an additional 7.8% of profiles, as those profiles were either deleted before we could inspect them thoroughly 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.

On the other hand, for the location based approach in detecting the deception, we have identified 5% of the 35,000 profiles collected as potentially deceptive profiles. We manually inspected profiles with a higher probability to be deceptive, as we report below, 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 potentially deceptive profiles that we identified. We found, in some cases, that about 90.0% of the potentially deceptive profiles were indeed

deceptive. In addition, the overall outcome of 5.0% of the users are potentially deceptive and about 35.0% of those users are likely deceptive.

6.2 Future Work

In the future we will continue exploring alternative strategies in an effort to improve the accuracy of our predictions even further. 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. Combining 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 supported by our main frameworks.

CITED LITERATURE

1. 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.

2. Guerrero, L. K. K., Andersen, P. A., and Afifi, W. A.: Close encounters: Communication in relationships. USA, Sage Publications, 2012.

3. Authority, A. S.: Children and advertising on social media websites. http://goo.gl/qswXGe.

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

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

6. Brain, S.: Twitter statistics. http://www.statisticbrain.com/twitter-statistics.

7. Business, t.: Who is on Twitter? https://business.twitter.com/whos-twitter.

8. Burger, J. D., Henderson, J., Kim, G., and Zarrella, G.: Discriminating gender on Twitter. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 1301–1309, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

9. Mocanu, D., Baronchelli, A., Perra, N., Gonçalves, B., Zhang, Q., and Vespignani, A.: The Twitter of babel: Mapping world languages through microblogging platforms. PloS one, 8(4):1–9, 2013.

10. Alowibdi, J. S., Buy, U. A., and Yu, P. S.: Language independent gender classification on Twitter. In IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM'13, pages 739–743, Aug 2013.

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

12. Alowibdi, J. S., Buy, U. A., and Yu, P. S.: Empirical evaluation of profile characteristics gender classification on Twitter. In The 12th International Conference on Machine Learning and Applications (ICMLA), volume 1, pages 365–369, Dec 2013.

13. Speech at CMU: The CMU pronouncing dictionary. `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`.

14. Alowibdi, J. S., Buy, U. A., Yu, P. S., and Stenneth, L.: Detecting deception in online social networks. In Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on, 2014.

15. Alowibdi, J. S., Buy, U. A., and Yu, P. S.: Deception detection in Twitter. In Journal-Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on, 2014.

16. Hota, S. R., Argamon, S., Koppel, M., and Zigdon, I.: Performing gender: Automatic stylistic analysis of Shakespeare's characters. In Proceedings of Digital Humanities 2006, pages 100–104, 2006.

17. Koppel, M., Argamon, S., and Shimoni, A. R.: Automatically categorizing written texts by author gender. Literary and Linguistic Computing, 17(4):401–412, 2002.

18. Argamon, S., Whitelaw, C., Chase, P., Hota, S. R., Garg, N., and Levitan, S.: Stylistic text classification using functional lexical features: Research articles. J. Am. Soc. Inf. Sci. Technol., 58(6):802–822, 2007.

19. Argamon, S., Koppel, M., Fine, J., and Shimoni, A. R.: Gender, genre, and writing style in formal written texts. Text, 23(3):321–346, 2003.

20. Singh, S.: A pilot study on gender differences in conversational speech on lexical richness measures. Literary and Linguistic Computing, 16(3):251–264, 2001.

21. Sarawgi, R., Gajulapalli, K., and Choi, Y.: Gender attribution: Tracing stylometric evidence beyond topic and genre. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning, pages 78–86, Portland, Oregon, USA, June 2011.

22. Nowson, S., Oberlander, J., and Gill, A.: Weblogs, genres and individual differences. In Proceedings of the 27th Annual Meeting of the Cognitive Science Society, pages 1666–1671, Stresa, Italy, 2005.

23. Halliday, M., Matthiessen, C. M., and Matthiessen, C.: An introduction to functional grammar. London, England, Routledge, 2014.

24. Brunet, É.: Le Vocabulaire de Jean Giraudoux: structure et évolution : statistique et informatique appliquées à l'étude des textes à partir des données du Trésor de la langue française. Travaux de linguistique quantitative. Geneva, Switzerland, Slatkine, 1978.

25. Honoré, A.: Some simple measures of richness of vocabulary. Association for Literary and Linguistic Computing Bulletin, 7(2):172–177, 1979.

26. Heylighen, F. and Dewaele, J.-M.: Variation in the contextuality of language: An empirical measure. Foundations of Science, 7(3):293–340, 2002.

27. Herring, S. C., Scheidt, L. A., Bonus, S., and Wright, E.: Bridging the gap: A genre analysis of weblogs. In System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on, 2004.

28. Miller, C. R. and Shepherd, D.: Blogging as social action: A genre analysis of the weblog. Into the blogosphere: Rhetoric, community, and culture of Weblogs, 2004.

29. Herring, S. C. and Paolillo, J. C.: Gender and genre variation in weblogs. Journal of Sociolinguistics, 10(4):439–459, 2006.

30. Yan, X. and Yan, L.: Gender classification of weblog authors. In AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, pages 228–230, 2006.

31. de Vel, O., Corney, M., Anderson, A., and Mohay, G.: Language and gender author cohort analysis of e-mail for computer forensics. In In proceeding of the Second Digital Forensics Research Workshop. DFRWS, 2002.

32. Kucukyilmaz, T., Cambazoglu, B. B., Aykanat, C., and Can, F.: Chat mining for gender prediction. In Advances in Information Systems, pages 274–283. Springer, 2006.

33. Mukherjee, A. and Liu, B.: Improving gender classification of blog authors. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language

Processing, pages 207–217, Cambridge, MA, October 2010. Association for Computational Linguistics.

34. Peersman, C., Daelemans, W., and Van Vaerenbergh, L.: Predicting age and gender in online social networks. In Proceedings of the 3rd international workshop on Search and mining user-generated contents, pages 37–44, 2011.

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

36. Mislove, A., Jørgensen, S. L., Ahn, Y.-Y., Onnela, J.-P., and Rosenquist, J. N.: Understanding the demographics of Twitter users. In 5th International AAAI Conference on Weblogs and Social Media (ICWSM'11), pages 554–557, 2011.

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

38. Al Zamal, F., Liu, W., and Ruths, D.: 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.

39. Liu, W., Al Zamal, F., and Ruths, D.: 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.

40. Liu, W. and Ruths, D.: 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.

41. Cheng, Z., Caverlee, J., and Lee, K.: 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, pages 759–768, 2010.

42. Jurgens, D.: 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.

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

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

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

46. Gao, H., Hu, J., Wilson, C., Li, Z., Chen, Y., and Zhao, B. Y.: Detecting and characterizing social spam campaigns. In Proceedings of the 10th ACM SIGCOMM conference on Internet measurement, pages 35–47, 2010.

47. Ramachandran, A., Feamster, N., and Vempala, S.: Filtering spam with behavioral blacklisting. In Proceedings of the 14th ACM conference on Computer and communications security, pages 342–351, 2007.

48. Damiani, E., De Capitani di Vimercati, S., Paraboschi, S., and Samarati, P.: P2p-based collaborative spam detection and filtering. In Peer-to-Peer Computing, 2004. Proceedings. Proceedings. Fourth International Conference on, pages 176–183, 2004.

49. Niu, Y., Chen, H., Hsu, F., Wang, Y.-M., and Ma, M.: A quantitative study of forum spamming using context-based analysis. In Internet Society Annual Network and Distributed System Security Symposium (NDSS), 2007.

50. Shin, Y., Gupta, M., and Myers, S.: Prevalence and mitigation of forum spamming. In INFOCOM, 2011 Proceedings IEEE, pages 2309–2317, 2011.

51. Shin, Y., Gupta, M., and Myers, S.: The nuts and bolts of a forum spam automator. In Proceedings of the 4th USENIX conference on Large-scale exploits and emergent threats, pages 3–3, 2011.

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

53. Yang, J., Counts, S., Morris, M. R., and Hoff, A.: Microblog credibility perceptions: comparing the usa and china. In Proceedings of the 2013 conference on Computer supported cooperative work, pages 575–586, 2013.

54. Yang, C., Harkreader, R., Zhang, J., Shin, S., and Gu, G.: Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on Twitter. In Proceedings of the 21st international conference on World Wide Web, pages 71–80, 2012.

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

56. Chu, Z., Gianvecchio, S., Wang, H., and Jajodia, S.: Detecting automation of Twitter accounts: Are you a human, bot, or cyborg? IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, pages 811–824, 2012.

57. Zhang, X., Zhu, S., and Liang, W.: Detecting spam and promoting campaigns in the Twitter social network. In Data Mining (ICDM), 2012 IEEE 12th International Conference on, pages 1194–1199, 2012.

58. Wang, A. H.: Don't follow me: Spam detection in Twitter. In Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on, pages 1–10, 2010.

59. Benevenuto, F., Magno, G., Rodrigues, T., and Almeida, V.: Detecting spammers on Twitter. In Collaboration, electronic messaging, anti-abuse and spam conference (CEAS), volume 6, 2010.

60. McCord, M. and Chuah, M.: Spam detection on Twitter using traditional classifiers. In Autonomic and Trusted Computing, pages 175–186. Springer, 2011.

61. Wang, A. H.: Machine learning for the detection of spam in Twitter networks. In e-Business and Telecommunications, pages 319–333. Springer, 2012.

62. Wauters, R.: Only 50% of Twitter messages are in english, study says. `http://techcrunch.com/2010/02/24/twitter-languages/`.

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

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

65. Kohavi, R. et al.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In IJCAI, volume 14, pages 1137–1145, 1995.

66. Speech at CMU: LOGIOS lexicon tool. http://www.speech.cs.cmu.edu/tools/lextool.html.

67. Cavnar, W. B., Trenkle, J. M., et al.: N-gram-based text categorization. Ann Arbor MI, 48113(2):161–175, 1994.

68. Bureau, U. C.: Frequently occurring first names from the census. http://www.census.gov/2010census/.

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

70. AnchorFree-Inc.: Hotspot shield. http://www.hotspotshield.com/.

71. information, T. and research centre: Tourism information. http://www.mas.gov.sa/.

72. Newspaper, S. O.: Saudi top destinations abroad. http://sabq.org/yX6fde.

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

74. Westin, K.: Confessions of a linkedin imposter: We are probably connected. http://www.tripwire.com/state-of-security/security-awareness/confessions-of-a-linkedin-imposter-we-are-probably-connected/.

75. Wood Jr, G. S. and Judikis, J. C.: Conversations on community theory. Purdue University Press, 2002.

76. Castelfranchi, C. and Tan, Y.-H.: Trust and deception in virtual societies. Kluwer Academic Publishers, 2001.

77. Ward, D. and Hexmoor, H.: Towards deception in agents. In Proceedings of the second international joint conference on Autonomous agents and multiagent systems, pages 1154–1155, 2003.

78. Vrij, A., Edward, K., Roberts, K. P., and Bull, R.: Detecting deceit via analysis of verbal and nonverbal behavior. Journal of Nonverbal Behavior, 24(4):239–263, 2000.

79. Griffin, E.: A first look at communication theory 4th edition. Boston: McGraw-Hill, 2003.

80. Guerrero, L. K., Andersen, P. A., and Afifi, W. A.: Close encounters: Communication in relationships. Sage, 2013.

81. Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and Bhattacharjee, B.: Measurement and analysis of online social networks. In Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, pages 29–42, 2007.

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

83. Caspi, A. and Gorsky, P.: Online deception: Prevalence, motivation, and emotion. CyberPsychology & Behavior, 9(1):54–59, 2006.

84. Burgoon, J. K., Blair, J., Qin, T., and Nunamaker Jr, J. F.: Detecting deception through linguistic analysis. In Intelligence and Security Informatics, pages 91–101. Springer, 2003.

85. Zhou, L., Burgoon, J. K., Nunamaker, J. F., and Twitchell, D.: Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. Group decision and negotiation, 13(1):81–106, 2004.

86. Hancock, J. T., Curry, L., Goorha, S., and Woodworth, M. T.: Lies in conversation: An examination of deception using automated linguistic analysis. In Annual Conference of the Cognitive Science Society, volume 26, pages 534–540, 2004.

87. Fuller, C. M., Biros, D. P., and Wilson, R. L.: Decision support for determining veracity via linguistic-based cues. Decision Support Systems, 46(3):695–703, 2009.

88. Newman, M. L., Pennebaker, J. W., Berry, D. S., and Richards, J. M.: Lying words: Predicting deception from linguistic styles. Personality and social psychology bulletin, 29(5):665–675, 2003.

89. Galanxhi, H. and Nah, F. F.-H.: Deception in cyberspace: A comparison of text-only vs. avatar-supported medium. International journal of human-computer studies, 65(9):770–783, 2007.

90. Twitchell, D. P., Nunamaker Jr, J. F., and Burgoon, J. K.: Using speech act profiling for deception detection. In Intelligence and Security Informatics, pages 403–410. Springer, 2004.

91. Pak, J. and Zhou, L.: A social network based analysis of deceptive communication in online chat. In E-Life: Web-Enabled Convergence of Commerce, Work, and Social Life, pages 55–65. Springer, 2012.

92. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., and Inches, G.: Overview of the author profiling task at pan 2013. In Notebook Papers of CLEF 2013 LABs and Workshops, CLEF-2013, Valencia, Spain, September, pages 23–26, 2013.

93. Nguyen, D., Gravel, R., Trieschnigg, D., and Meder, T.: Tweetgenie: automatic age prediction from tweets by d. nguyen, r. gravel, d. trieschnigg, and t. meder; with ching-man au yeung as coordinator. ACM SIGWEB Newsletter, (Autumn):4, 2013.

94. Nguyen, D., Gravel, R., Trieschnigg, D., and Meder, T.: " how old do you think i am?"; a study of language and age in Twitter. In Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, 2013.

95. Holten, D.: Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. IEEE Transactions on Visualization and Computer Graphics, 12(5):741–748, 2006.

VITA

Research Interests

Centered primarily around Social Networks, Data Mining and Software Engineering. In particular, I am interested in Privacy, Security, Analysis and Information Retrieval on Social Networks, Security on Software Systems.

Education

2009–2014 **Ph.D. in Computer Science**, *University of Illinois, College of Engineering, Department of Computer Science*, Chicago, USA.
Thesis: "Detecting Deception in Online Social Networks."
Advisors: Dr. Ugo Buy and Professor Philip Yu.

2007–2008 **Master of Science in Software Engineering**, *DePaul University, College of Computing and Digital Media*, Chicago, USA.
Managed and Developed a part of Collaborative System for Requirement Elicitation.

2001–2005 **Bachelor of Science Degree in Computer**, *King Abdulaziz University, Faculty of Computing and Information Technology*, Jeddah, Saudi Arabia.
Developed an E-learning System (FOSOL)

Certifications

2008 Certificate of Academic Requirements Completion: **Java Developer Program**, **IT Project Management Program** and **SQL Server Database Administration Program**, *The Institute for Professional Development, DePaul University, Chicago*

2010 Certificate of Course Completion: **CCNA Exploration: Network Fundamentals, Routing Protocols and Concepts**, **LAN Switching and Wireless and Accessing the WAN**, *Cisco Networking Academy*

Experience

2009–2010 **Research Assistant**, *Distributed Real-Time Intelligent Systems Laboratory*, University of Illinois, College of Engineering, Department of Computer Science, Chicago.

2011–2012 **Graduate Assistant**, *Web Developer at Office of Information Technology*, University of Illinois, College of Liberal Arts and Sciences, Chicago.

2012–2012 **Project Officer**, International Monetary Fund, Washington D.C..

2010–2014 **Research Assistant**, University of Illinois, College of Engineering, Department of Computer Science, Chicago.

VITA

Computer skills

Languages	C++, C, Matlab, Zoom, Java, Perl	**Scripts**	JavaScript, AJAX, PHP, JSP, LaTeX
Web Apps	PHPNuke, XOOPS, Wordpress	**Databases**	Oracle, SQL Server, MySQL, SQLite
Technical	Network administrator, Maintenance, NetLOGO, Photoshop, Dreamweaver, MSI Packaging		

Languages

- **Arabic**: Native Tongue
- **English**: Fluent

Awards and Memberships

Graduate Level

2011 **Yahoo!'s 2011 Key Scientific Challenges (KSC) Program in Security and Privacy**, *An International program for choosing the most promising research to promote the future of the Internet in different topics*, Yahoo, Sunnyvale, CA.

2011–2014 **President**, *Saudi Students Association*, University of Illinois, Chicago, IL.

2011–2014 **President**, *Saudi Students Union*, Chicago, IL.

2009–2014 **Scholarship for Doctorate Degree**, *King Abdullah Scholarship Program*, Ministry of Higher Education, KSA.

2007–2008 **Scholarship for Master Degree**, *King Abdullah Scholarship Program*, Ministry of Higher Education, KSA.

20005–2006 **Scholarship for Diploma in English Language**, *King Abdullah Scholarship Program*, Ministry of Higher Education, KSA.

2010–Lifetime **Golden Key International Honour Society**, *An international Honour Society is given to the top 3% students in the university*, University of Illinois, Chicago, IL.

2010–2011 **Treasurer**, *Computer Science Graduate Student Association*, University of Illinois, Chicago, IL.

2008–Lifetime **Upsilon Pi Epsilon (UPE)**, *An international Honor Society for the Computing and Information Disciplines*, DePaul University, Chicago, IL.

2007–Lifetime **Golden Key International Honour Society**, *An international Honour Society is given to the top 3% students in the university*, DePaul University, Chicago, IL.

VITA

Undergraduate Level

2002–2005 **Dean's Honor List**, *Received six certifications and awarded 1000$*, King Abdulaziz University, KSA.

2005 **Outstanding Student**, *Honored from the Prince of Makkah Province, the Minister of Higher Education and the President of King Abdulaziz University for outstanding student with second honor degree in BSc.*, King Abdulaziz University, KSA.

2005 **Travel Award**, *Awarded to travel to participate in the first leaders work students in cooperation council for the Gulf States in Bahrain*, King Abdulaziz University, KSA.

2003–2005 **Golden Card**, *Awarded a golden card from the Deanship of student affairs for my activities in the university to be VIP student*, King Abdulaziz University, KSA.

2004–2005 **Student Counsel**, *Student counsel to represent the Faculty of Science*, King Abdulaziz University, KSA.

2003–2005 **Gifted Center**, *A center for talented student*, King Abdulaziz University, KSA.

2003–2005 **President**, *Computer Club*, King Abdulaziz University, KSA.

Other

2007–Present **Association for Computing Machinery (ACM)**, *Learning society for computing*, USA.

2007–Present **Institute of Electrical and Electronics Engineers (IEEE)**, *Learning society for engineering sciences, research, and technology*, USA.

2002–2008 **Saudi Computer Society**, *Learning society for computing*, KSA.

Interests

Reading, Research, Programming, Soccer, Swimming, Puzzle games and Surfing the Internet

VITA

Publications

Conference Publications

- **J. Alowibdi**. "Detecting Deception in Online Social Networks," Yahoo! 2011 Key Scientific Challenges Program (KSC'11), Sunnyvale, CA, USA, August 2011
- **J. Alowibdi**, U. Buy, and P. Yu. "Language Independent Gender Classification on Twitter," Proc. 2013 IEEE/ACM Int. Conf. on Advances in Social Networks Analysis and Mining (ASONAM'13), Niagara Falls, Ontario, Canada, August 2013
- **J. Alowibdi**, U. Buy, and P. Yu. "Empirical Evaluation of Profile Characteristics Gender Classification on Twitter," Proc. IEEE 12th International Conference on Machine Learning and Applications (ICMLA 2013), Miami, Florida, Dec. 2013
- **J. Alowibdi**, U. Buy, and P. Yu. "Say It with Colors: Language-Independent Gender Classification on Twitter," Online Social Media Analysis and Visualization, Lecture Notes on Social Networks, (LNSN) Series by Springer-Verlag, Eds. Jalal Kawash, New York, 2014
- **J. Alowibdi**, U. Buy, P. Yu, and L. Stenneth. "Detecting Deception in Online Social Networks," Proc. 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM'14, Beijing, China, 2014
- **J. Alowibdi**, S. Ghani, M. Mokbel. "VacationFinder: A Tool for Collecting, Analyzing, and Visualizing Geotagged Twitter Data to Find Top Vacation Spots," 6th ACM SIGSPATIAL International Workshop on Location-Based Social Networks, LBSN'14, Dallas, Texas, USA, 2014
- **J. Alowibdi**, U. Buy, and P. Yu. "Detecting Deception in Twitter," Journal of Social Network Analysis and Mining, Series by Springer-Verlag, Eds. Reda Alhajj, 2014
- **J. Alowibdi** and L. Stenneth, *"An Empirical Study of Data Race Detector Tools"*, Proceedings of the 25th IEEE Chinese International Conference on Control and Decision, CCDC'13, Guiyang, China
- L. Stenneth, K. Thompson, **J. Alowibdi** and W Stone, *"Transportation transfers from GPS sensor trace"*, 15th IEEE International Intelligent Transportation Systems Conference, ITSC'12, Anchorage, Alaska, USA
- **J. Alowibdi**, *"Adopting Knowledge Based Security System for Software Development Life Cycle"*, The 2011 International Conference on Software Engineering Research and Practice, SERP'11, WORLDCOMP 2011, Las Vegas, Nevada, USA
- W. Stone, L., and **J. Alowibdi**, *"Reducing Travel Time by Incident Reporting via CrowdSourcing"*, The 2011 International Conference on Internet Computing, ICOMP'11, WORLDCOMP 2011, Las Vegas, Nevada, USA
- **J. Alowibdi**, *"Artificial Neural Network for Recognizing Handwritten and Machine Printed Type Alphanumeric Arabic Characters"*, UIC Student Research Forum 2011, University of Illinois at Chicago, USA
- **J. Alowibdi**, *"Managing Online Requirements Elicitation in Ultra-Large Software System"*, CDMRS/CTIRS Research Symposium, DePaul University, Chicago, IL, USA(2008)