Identity Deception Detection on Social Media Platforms

Estée van der Walt and J. H. P. Eloff
Department of Computer Science, University of Pretoria, Pretoria, South Africa
estee.vanderwalt@gmail.com, eloff@cs.up.ac.za

Keywords: Deception, Big Data, Cyber Security, Social Media, Cyber Crime.

Abstract: The bulk of currently available research in identity deception focuses on understanding the psychological motive behind persons lying about their identity. However, apart from understanding the psychological aspects of such a mindset, it is also important to consider identity deception in the context of the technologically integrated society in which we live today. With the proliferation of social media, it has become the norm for many people to present a false identity for various purposes, whether for anonymity or for something more harmful like committing paedophilia. Social media platforms (SMPs) are known to deal with massive volumes of big data. Big data characteristics such as volume, velocity and variety make it not only easier for people to deceive others about their identity, but also harder to prevent or detect identity deception. This paper describes the challenges of identity deception detection on SMPs. It also presents attributes that can play a role in identity deception detection, as well as the results of an experiment to develop a so-called Identity Deception Indicator (IDI). It is believed that such an IDI can assist law enforcement with the early detection of potentially harmful behaviour on SMPs.

1 INTRODUCTION

Many cyber-security threats can be found in big data platforms such as social media, for instance grooming (Dedkova, 2015), paedophilia (Schulz et al., 2015), cyber bullying (Al-garadi et al., 2016) and cyber terrorism (Von Solms and Van Niekerk, 2013), to name but a few. These threats generally affect people and some of them apply identity deception as a means to an end. Consider for example the case study of identity deception where a paedophile in Cape Town (South Africa) was found guilty of 148 sex crimes. These crimes involved children between the age of 12 and 16 who had been befriended on Facebook via a fake profile depicting that of a minor (Peterson, 2016).

Current research in this field focuses mainly on understanding the psychological motive behind identity deception. Research shows why people lie about their identity (Kashy and DePaulo, 1996) and gives some cues about deception (DePaulo et al., 2003).

However, very little has been done so far to describe the influence that big data platforms like SMPs have. (Back et al., 2010), as well as (Guillory and Hancock, 2012) found that people were less likely to lie when other people could verify and hold them accountable for the facts stated on social media. Peer pressure also had an effect on the tendency to tell lies (Squicciarini and Griffin, 2014). However, none of the above addressed the detection of identity deception in SMPs. SMPs have changed not only our perception and use of Identity Deception, but also challenge its detection. A plethora of identity attributes exist, of which some pertain only to SMPs. An example of such an attribute would be the Twitter 'handle' or username, which is a personal identifier of a specific person on the SMP.

Identity attributes on SMPs are potentially vulnerable to online deception. In this context, the work done by Alowibdi et al. describes a set of methods to detect online gender or location deception (Alowibdi et al., 2015). The problem with this approach is that the methods and attributes used are considered in isolation and independent from others, which creates a challenge for determining a deceptive person across multiple attributes. Deception in one attribute could be deemed harmless, but in combination it may well tell another story. A person could for example have no profile image. This does not necessarily indicate potential deception and may only mean that the person wants to remain anonymous. However, if this knowledge is combined with the fact that the person's online activity differs

Walt, E. and Eloff, J.
Identity Deception Detection on Social Media Platforms.
DOI: 10.5220/0006271105730578
In *Proceedings of the 3rd International Conference on Information Systems Security and Privacy (ICISSP 2017)*, pages 573-578
ISBN: 978-989-758-209-7

from the norm with regard to daily activity, suspicions should be raised.

The first part of this paper investigates identity deception and the challenges specifically introduced by SMPs. The subsequent part presents experimental research to evaluate identified identity attributes that can potentially be used to detect identity deception on SMPs. The paper concludes with discussing the first steps towards the detection of identity deception on SMPs.

2 CHALLENGES OF IDENTITY DECEPTION ON SMPS

In previous research (Van der Walt and Eloff, 2015), the authors did not focus on identity deception, but extracted attributes available in social media that can now be used to identify those attributes that are vulnerable to identity deception.

The attributes in SMPs differ from those in normal record-keeping systems (e.g. a police database for offenders). SMP attributes, for example, username, birth date and address, can often not be trusted. There is no accountability for the accuracy of these attributes (Duranti and Rogers, 2016) and only human intervention (Sloan et al., 2015) can help to ascertain their accuracy. Since SMPs constitute a big data platform, human intervention is not plausible at the expected scale.

To understand how SMPs support identity deception, it is imperative to understand what it is by investigating the underlying concepts. The next paragraphs introduce and describe the concepts of identity, deception, and identity deception.

2.1 Identity

Identity attributes define who you are or any qualities that you display that can be used to distinguish you from another person. The following attributes in SMPs are known to be indicative of deception:

- The friend, follower ratio (Quercia et al., 2011).
- The type of images used as profile and background (Sharma, 2013).
- The distance between the geo-location recorded on the SMP and the location as stated by the person (Alowibdi et al., 2015).
- The sentiment, i.e. whether the overbearing language usage conveys a positive or negative feeling (Drasch et al., 2015).
- The number of devices used (Robinson, 2016).
- The timespan of activities on SMPs for a given person, compared to the corpus (Radziwill and Benton, 2016).

(Wang et al., 2006) split identity attributes into three groups, namely personal information provided, biometrical attributes that belong to an identity, and biographical attributes that build up over a period of time, for example the credit history of a person. (Clarke, 1994) defines identity attributes as one's appearance, name, the code you are identified by, your social behaviour, knowledge, what you have, what you do, what you are and your physical characteristics.

Taking cognisance of the above definitions, the authors of this paper identified the following groups of identity attributes related to SMPs for consideration in this study:

- Attributes that can change, for example location.
- Attributes that cannot change, for example birth date, ID, name.
- Attributes that change over time, for example image.
- Attributes that indirectly define who you are, for example friends, number of devices used.

These attribute groups help us to understand what strategies can be applied towards deception and are discussed in the next section.

2.2 Deception

The Oxford English Dictionary defines 'deception' as "the action of deceiving someone", and 'deceit' as "the action or practice of deceiving someone by concealing or misrepresenting the truth" (Oxford, 2012). In the authors' opinion, deception is when a fact is presented that is contrary to the truth.

Deception can present itself in many forms. For example, lies can be spread about the outcome of an election (Cook et al., 2014) or fake news can be published to spread angst (Conroy et al., 2015).

A variety of research has been presented to define deception based on its purpose. (Wang et al., 2006) classified deception based on three main purpose groups, namely concealment, theft and forgery. According to (Kashy and DePaulo, 1996), people deceive for the purpose of manipulation, impression management, insecurity, socialisation, sociability or relationship management.

Interpersonal Deception Theory (IDT) defines the following different strategies towards deception (Buller and Burgoon, 1996):

- *Falsification* or changing of the facts.
- *Exaggeration* of facts.
- *Omission* of important information.

- *Equivocation* or presentation of vague information to leave a false impression.

Truth Deception Theory (TDT), on the other hand, defines the following motives towards deception and closely resembles IDT: *Lies*, *omission*, *evasion*, *equivocation* and generating *false conclusions* from true information (Levine et al., 2016).

The current paper focuses on the IDT strategies used for deception as defined by (Buller and Burgoon, 1996). These strategies, such as the omission of information, easily relate to the identity attribute groups defined for SMPs. An example would be the omission of a birth date on a SMP.

Deception in SMPs is common and sometimes even expected (Liu et al., 2014). This paper focuses only on identity deception. The latter is but one example where deceit can be used to harm others on SMPs.

2.3 Identity Deception

Identity deception occurs where the truth is misrepresented to assume another identity. Identity deception has been recorded as early as in the Old Testament of the Bible where Jacob donned his brother's clothes to deceive their father into giving him the inheritance that rightfully belonged to his brother Esau (Bond and DePaulo, 2006).

Identity deception can be seen from different viewpoints, i.e. finding similar identities for people where none should exist or finding deceptive identities where similar ones can exist. Many studies are done towards detecting similar identities. The most common strategy is to detect similar identities based on specific common identity attributes, such as birth date or ID number (Li and Wang, 2015).

Finding similar identities for people in SMPs based on common identity attributes are however challenging, as it is difficult to confirm the accuracy of identity attributes if there is no accountability (Duranti and Rogers, 2016). People are furthermore known to frequently change their social profiles (Liu et al., 2014). Therefore, even though matched identities can be identified, the results cannot always be trusted.

This paper therefore focuses on finding deceptive identities in a corpus where their profiles could be very similar and make distinction between them difficult. A so-called Identity Deception Indicator (IDI) is proposed and discussed in the next section.

3 AN EXPERIMENT TO EVALUATE RELEVANT ATTRIBUTES IDENTIFIED FOR IDENTITY DECEPTION DETECTION

Previous research and experiments by the authors of this paper defined a process for building an IDI for SMPs (Van der Walt and Eloff, 2015). The goal of the previous experiments was to gather social media data, clean the data and understand what attributes in general are available for further exploration on SMPs.

To help determine what social media data to gather, a specific case study for potential identity deception was proposed. The experiments focused on gathering data for minors as they are particularly vulnerable to identity deception from various online threats, such as from paedophiles (Schulz et al., 2015). The experiments used data collected from Twitter for tweets that mention the words 'school' and 'homework'. (Schwartz et al., 2013) believe that these are the top two words used by minors in social media. The data also included tweets from their friends and followers. Overall the corpus collected consisted of 4,764,733 tweets from 6,846 accounts.

This paper continues with the process of building an IDI for SMPs by scoring each Twitter user's identity attributes and by understanding how these contribute towards establishing the user's perceived deceptiveness on SMPs. The aim is not to categorise the Twitter user as being deceptive, but rather to understand whether the attribute could be used towards detecting deception.

In the remainder of the paper, the score given to each Twitter user's identity attributes are also referred to as the Deception Score (DS). A DS is the result of calculations used to ascertain a user's perceived deceptiveness, given the identity attribute. The DS is defined as being in a range between 0 and 1, with 1 being more deceptive.

The algorithms used to calculate the DS per identity attribute are discussed next.

3.1 Distance as an Identity Attribute

Twitter allows a person to state their time zone as part of their profile. The latitude and longitude of each time zone are retrieved using the online Geonames dataset (Wick, 2016). Twitter also automatically stores the sender's latitude and longitude when a tweet is sent from a device. Since people can disable the feature to protect their privacy, the experiment excludes all Twitter users with the geo-location

feature disabled. This feature allows for the calculation of the distance between the geo-location and the time zone as stated by the user. The Haversine method is used to calculate the distance in kilometres between two points on a sphere, given their latitude and longitude (Van Liere, 2010).

The assumption is that outliers indicate a better likelihood of deception as most people are believed to be good and not to tell lies on SMPs (Back et al., 2010) (Dedkova, 2015). Outliers would thus denote those Twitter users who deviate from the norm. Since the experiment recognised that distances can differ between continents where land coverage could vary within a given time zone, outliers were determined per continent.

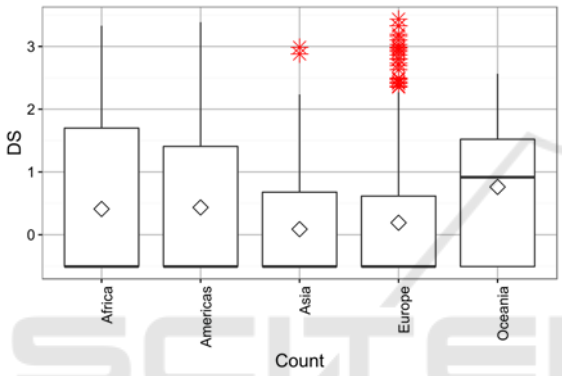

Figure 1: Outliers based on distance per continent.

The results in Figure 1 show that outliers existed in the corpus for Europe and Asia as depicted by the red markers. Outliers are those data points that fall outside one of the following:

- Quartile 1 – (IQR x 1.5)
- Quartile 3 + (IQR x 1.5)

where IQR refers to the Inter Quartile Range. (The above is also known as Tukey's method or Tukey's honest significance test (Navidi, 2006).)

If the distance calculated for a user was one of the outlier data points, the attribute was given a DS of 1 or greater (being deceptive). If the data point was outside the IQR but not defined as an outlier, a DS of 0.5 was awarded. Otherwise the user was regarded as trustworthy.

3.2 Sentiment as an Identity Attribute

Another identity attribute investigates the sentiment of a Twitter user. Each tweet on the Twitter platform can contain up to 140 characters. Various studies have also shown that certain words can be associated with the sentiment of the person (Ghiassi et al., 2013) (Haque and Rahman, 2014). For the experiment at hand, the words in the tweets were matched with the NRC Emotion Lexicon dataset (Mohammad, 2016) to extract and count those words signifying positive or negative sentiment per user.

Figure 2 shows that most Twitter users' overall sentiment is positive. The assumption was made that if a user's sentiment was negative, this signified an outlier to the norm and could be a potential indicator of Identity Deception. A DS of 1 (being deceptive) was awarded to all users displaying an overall negative sentiment.

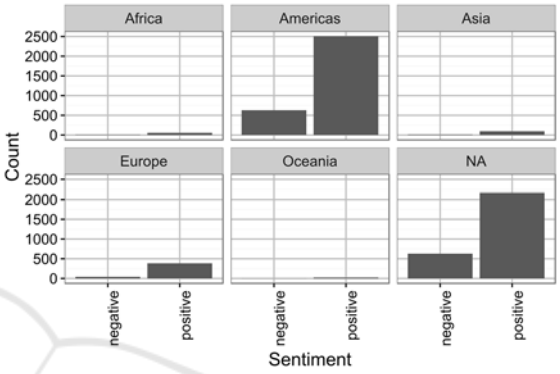

Figure 2: Overall sentiment per continent

3.3 Other Identity Attributes

A DS was also calculated for each of the listed identity attributes as part of the experiment. The calculations are depicted in Table 1.

Each of the DS scores produced results, which are discussed next.

Table 1: DS calculations for other identity attributes.

Attribute	DS calculation
Number of devices	When number of devices = 1, then DS= 0; when 2, then 0.5; else 0.5 + (number of devices * 0.05)
Friends vs followers	Normalised friend-to-follower ratio
Type of images	When image is unique, DS is 0; else 1
Average tweet time	Within IQR it is 0; between min and max, DS is 0.5; when outlier, then 1

3.4 Results

Potentially deceptive Twitter users were identified for each of the identity attributes based on their DS. This result shows that all identity attributes evaluated were successful contributors towards the detection of deception and good candidates towards building an IDI for SMPs.

4 FIRST STEPS TOWARDS THE DETECTION OF IDENTITY DECEPTION FOR SMPS

The simplest suggestion for creating an IDI is to aggregate the DSs of all attributes and divide the result by the number of identity attributes. This is depicted by the following formula where *n* is the number of identity attributes being used:

$$IDI = (\sum_{k=1}^{n} DS\,(k))/n \quad (1)$$

Twitter user *x*, for example, achieves a DS per identity attribute as depicted in Table 2.

Table 2: DS per identity attribute for user x.

Identity Attribute	DS
Number of devices	1
Friends vs followers	0.5
Type of images	0.2
Average tweet time	1
Distance deception	0
Sentiment of the user	0.8

In terms of the suggested formula, the IDI for user *x* is calculated as: (1+0.5+0.2+1+0+0.8)/6 = 0.58

Calculating the IDI per user could however lead to an incorrect assumption that users with a higher IDI are more deceptive. The following flaws were identified in this first proposed approach towards calculating an IDI:

- In SMPs there are many examples of users who are harmless, but nevertheless deceptive. An example in point would be celebrities who want to remain anonymous.
- All DSs do not carry the same weight. One attribute could provide a higher indication of deception than another. The formula above suggests otherwise.
- Additional identity attributes from other SMPs than Twitter could enhance the IDI.

Future research will aim to address these issues as they fall outside of scope of the current paper.

5 CONCLUSIONS

Numerous identity attributes exist to describe the identity of persons on SMPs. These identity attributes are either provided by the persons themselves, i.e. their user name, or captured by the SMP, i.e. their geo-location.

Nonetheless, different strategies can be applied to deceive and hide a person's identity to allow him/her to do harm. This paper identified various identity attributes that are all perceived to be potentially vulnerable to deception.

The experiment discussed in this paper evaluated each of the identity attributes to understand their contribution towards deception detection. The evaluation was made by means of a DS. A first step towards establishing a so-called IDI on SMPs was also defined. It was a simple approach to start with and has potential for improvement.

Future research will focus on applying DSs more cleverly to improve the identity deception indicator. The IDI should take note of the fact that deceptive Twitter users can be totally harmless and that certain attributes are more important than others. The aim is to identify – with good accuracy – clusters of harmful deceptive users. It is envisaged that these clusters of users can be passed on to law enforcement agencies for further analysis and for the early detection of potentially harmful behaviour on SMPs.

REFERENCES

Al-Garadi, M. A., Varathan, K. D. & Ravana, S. D. 2016. Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. *Computers in Human Behavior*, 63**,** 433-443.

Alowibdi, J. S., Buy, U. A., Philip, S. Y., Ghani, S. & Mokbel, M. 2015. Deception detection in Twitter. *Social Network Analysis and Mining*, 5**,** 1-16.

Back, M. D., Stopfer, J. M., Vazire, S., Gaddis, S., Schmukle, S. C., Egloff, B. & Gosling, S. D. 2010. Facebook profiles reflect actual personality, not self-idealization. *Psychological science*.

Bond, C. F. & Depaulo, B. M. 2006. Accuracy of deception judgments. *Personality and social psychology Review*, 10**,** 214-234.

Buller, D. B. & Burgoon, J. K. 1996. Interpersonal deception theory. *Communication theory*, 6**,** 203-242.

Clarke, R. 1994. Human identification in information systems: Management challenges and public policy issues. *Information Technology & People*, 7**,** 6-37.

Conroy, N. J., Rubin, V. L. & Chen, Y. 2015. Automatic Deception Detection: Methods for Finding Fake.

Cook, D. M., Waugh, B., Abdipanah, M., Hashemi, O. & Abdul Rahman, S. 2014. Twitter Deception and Influence: Issues of Identity, Slacktivism, and Puppetry. *Journal of Information Warfare*, 13**,** 58-71.

Dedkova, L. 2015. Stranger Is Not Always Danger: The Myth and Reality of Meetings with Online Strangers. *LIVING IN THE DIGITAL AGE***,** 78.

Depaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck,

L., Charlton, K. & Cooper, H. 2003. Cues to deception. *Psychological bulletin,* 129**,** 74.

Drasch, B., Huber, J., Panz, S. & Probst, F. 2015. Detecting Online Firestorms in Social Media.

Duranti, L. & Rogers, C. 2016. Trust in Records and Data Online. *Integrity in Government through Records Management. Essays in Honour of Anne Thurston***,** 203-214.

Ghiassi, M., Skinner, J. & Zimbra, D. 2013. Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with Applications,* 40**,** 6266-6282.

Guillory, J. & Hancock, J. T. 2012. The effect of Linkedin on deception in resumes. *Cyberpsychology, Behavior, and Social Networking,* 15**,** 135-140.

Haque, M. A. & Rahman, T. 2014. Sentiment analysis by using fuzzy logic.

Kashy, D. A. & Depaulo, B. M. 1996. Who lies? *Journal of Personality and Social Psychology,* 70**,** 1037.

Levine, T. R., Ali, M. V., Dean, M., Abdulla, R. A. & Garcia-Ruano, K. 2016. Toward a Pan-cultural Typology of Deception Motives. *Journal of Intercultural Communication Research,* 45**,** 1-12.

Li, J. & Wang, A. G. 2015. A framework of identity resolution: evaluating identity attributes and matching algorithms. *Security Informatics,* 4**,** 1.

Liu, H., Han, J. & Motoda, H. 2014. Uncovering deception in social media. *Social Network Analysis and Mining,* 4**,** 1-2.

Mohammad, S. 2016. NRC Emotion Lexicon.

Navidi, W. C. 2006. *Statistics for engineers and scientists,* McGraw-Hill New York.

Oxford 2012. The English Oxford Dictionary. Third Edition, March 2012 ed.: Oxford University Press.

Peterson, T. 2016. Rapist who used social media to lure child victims sentenced to 20 years. *News24,* 15 Jun 2016.

Quercia, D., Kosinski, M., Stillwell, D. & Crowcroft, J. Our Twitter profiles, our selves: Predicting personality with Twitter. Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on, 2011. IEEE, 180-185.

Radziwill, N. M. & Benton, M. C. 2016. Bot or Not? Deciphering Time Maps for Tweet Interarrivals. *arXiv preprint arXiv:1605.06555*.

Robinson, D. 2016. Two people write Trump's tweets. He writes the angrier ones. *Washington Post,* 12 Aug 2016.

Schulz, A., Bergen, E., Schuhmann, P., Hoyer, J. & Santtila, P. 2015. Online Sexual Solicitation of Minors How Often and between Whom Does It Occur? *Journal of Research in Crime and Delinquency***,** 0022427815599426.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D. & Seligman, M. E. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one,* 8**,** e73791.

Sharma, S. 2013. Black Twitter? Racial hashtags, networks and contagion. *New Formations,* 78**,** 46-64.

Sloan, L., Morgan, J., Burnap, P. & Williams, M. 2015. Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data. *PloS one,* 10**,** e0115545.

Squicciarini, A. & Griffin, C. 2014. Why and how to deceive: game results with sociological evidence. *Social Network Analysis and Mining,* 4**,** 1-13.

Van Der Walt, E. & Eloff, J. H. P. 2015. Protecting minors on social media platforms - A Big Data Science experiment *HPI Cloud Symposium "Operating the Cloud"*. Potsdam, Germany: Hasso Plattner Institut.

Van Liere, D. How far does a tweet travel?: Information brokers in the twitterverse. Proceedings of the International Workshop on Modeling Social Media, 2010. ACM, 6.

Von Solms, R. & Van Niekerk, J. 2013. From information security to cyber security. *Computers & Security,* 38**,** 97-102.

Wang, G. A., Chen, H., Xu, J. J. & Atabakhsh, H. 2006. Automatically detecting criminal identity deception: an adaptive detection algorithm. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on,* 36**,** 988-999.

Wick, M. 2016. Geonames.