A Big Data Science experiment – Identity Deception Detection

Estée van der Walt
Department of Computer Science
Cyber-security & Big Data Science Research Group
University of Pretoria, South Africa
estee.vanderwalt@gmail.com

J.H.P. Eloff
Department of Computer Science,
Cyber-security & Big Data Science Research Group
University of Pretoria, South Africa
eloff@cs.up.ac.za

Abstract – Identity Deception Detection is a problem on social media platforms today. Not only is there challenges towards determining the authenticity of people, but also with analyzing the data that forms part of the communications. These data are of heterogeneous type and include photos, videos and sound. Furthermore, most social media platforms are operating in an uncontrolled environment. Any person can contribute content and take part. Even though age restrictions do exist there are no enforcement of these laws and honesty of the public is expected. This is dangerous for minors specifically as they are either unaware of the dangers or not mature enough to be responsible for their actions online. Online predators are aware of this fact and targeting this group specifically. This paper presents work-in-progress towards developing an intelligent Identity Deception Indicator (IDI). It is envisaged that this work could eventually assists authorities in doing large-scale observation on publicly available social media platforms, such as Twitter. Of particular interest are those personas whose behavior and online content does not fit with the age group they are conversing with.

Keywords **-** Identity deception, Big Data, Data Science, Cyber-security, Social media

I. INTRODUCTION

Deception is defined as a "deliberate attempt, without forewarning, to create in another a belief which the communicator considers to be untrue" [1]. According to a paper by Michail Tsikerdekis and Sherali Zeadally [2] identity deception can be divided into three different types: hiding your real identity, using an identity of another real person and counterfeiting an identity. Of particular interest to this study is the case of counterfeiting an identity. It is easy for a predator to counterfeit an identity and to go unnoticed in a big data environment, such as social media platforms [3].

There is a need for new innovative solutions that can minimize the risk of identity deception. Innovations are needed on both a large scale as well as on a small scale. Large-scale innovations will be of benefit to societies in the large e.g. designing intelligent identity deception tools that can aid the prosecution of harmful people. Small-scale innovations will be of benefit to the individual e.g. designing tools that can provide safe browsing on the Internet. The research at hand focuses on "large-scale" innovations with regard to identity deception.

The remainder of this paper is structured as follows:
Section II describes the convergence of cyber-security, big data, data science and human factors specifically from the viewpoint of protecting minors via deception detection on social media platforms. Section III continues with work-in-progress, in the form of an experiment showing how to possibly work towards the construction of an intelligent Identity Deception Indicator (IDI). Section IV concludes the discussion with an indication towards expanding the experiment and future research.

II. BACKGROUND

Big Data combined with machine learning and predictive modeling allows for extrapolating value out of the large volumes of data. Literature refers to the three characteristics (volume, velocity and variety) of big data as the 3V's [4]. Many other characteristics of big data have been proposed like 'Value' [5], 'Viability' [6], 'Validity' [7] and 'Veracity' [8]. Because data from social media platforms originates from uncontrolled environments it makes minors especially vulnerable to harmful behavior of bad people e.g. pedophiles. Social media platforms provide the ideal platform for an attack mainly because of its big data nature and the complexity of its non-textual data. It is easy for a predator to go unnoticed in this environment [9].

Figure 1 illustrates: firstly, the convergence of the particular fields that are of interest to this study, and secondly it shows that there is an emanating risk of identity deception as a result of the convergence.

The protection of minors has always had the privilege of much focus but various laws are either in draft or have been enacted to this effect like CDA (Communications of Decency Act), COPA (Child Online Protection Act), CIPA

978-1-4673-9795-7/15 $31.00 © 2015 IEEE
DOI 10.1109/CSCI.2015.169

(Children Internet Protection Act), DOPA (Deleting Online Predators Act) and COPPA (Children Online Privacy Protection Act) [10]. This however did not stop minors in themselves from using social media platforms and thereby being exposed to the threats such as cyber bullying and pedophilia. The problem of protecting minors are even more complicated by the fact that many minors lie about their age whilst communicating online [9] [11] [12]. Deception is common in everyday life [13] and lying on Online Social Networks (OSNs) are quite common [14] [15]. Deception is always used to accomplish goals [16] and in the case of minors sometimes to a detrimental effect.

Identity deception is usually very difficult on a large scale for law enforcement agencies as they rely on machines to find exact matches [17]. Large volumes (big data) and the reliance on historic data however pose a problem as the crime is usually only detected too late, after it already occurred.

In summary, the following concluding remarks are of importance for the research at hand:

- All the different characteristics of big data are important to understand and to take cognizance of when searching for solutions to detect identity deception on social media platforms.
- From a cyber-security perspective the goal is to minimize the risk of minors being exposed to harmful behavior on social media platforms. Designing appropriate countermeasures, such as the Identity Deception Indicator proposed in this paper, should reduce this risk.
- Considering online activities of minors and perpetrators on social media platforms there is a need for determining the authenticity of an identity.

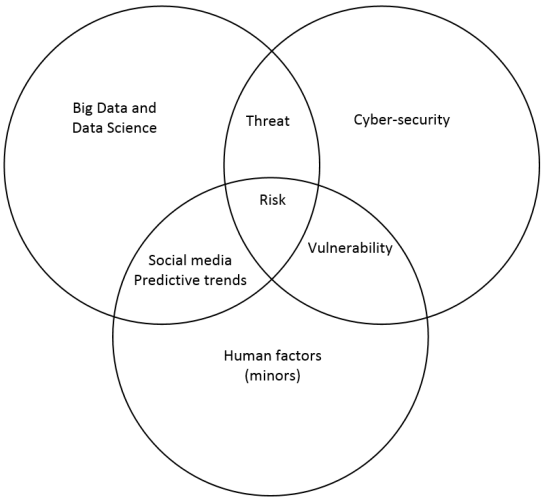

Fig. 1. The convergence of Big Data and Data Science, Cyber-security and Human factors

III. BIG DATA SCIENCE - EXPERIMENT

A. Process

The Big Data Science experiment, as presented in this paper, follows and extends the work presented in a previous paper [18]. This process is depicted in Fig. 2. Experimental work [18] completed prior to the paper at hand, identified Twitter as the social media platform to be used. Two sets of Twitter-data were retrieved. Set-1 is an initial set of tweets where the hash tags 'school' or 'homework' is used. These hash tags were chosen as they were deemed to be the words mostly applicable to minors according to a study from Schwartz et al [19]. Set-2 is a set of the last 200 tweets of Set-1 including the last 200 tweets of their followers and friends. Combined these two sets creates a big data set of a network of social media messages between "potential" minors.

The discussion that follows focuses only on two of the tasks as shown in figure 2, namely: Understanding the data gathered and Enrichment of variables.

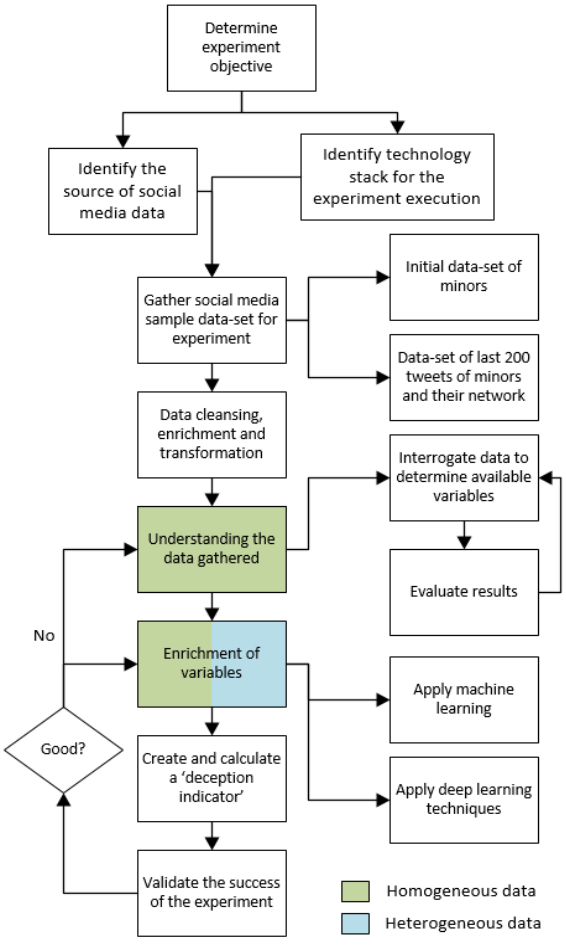

Fig. 2. Big Data Science experiment

B. Understanding the data gathered

In [18] it was postulated that tweets themselves could carry some indication of age. Further perusal of the data indicated that heterogeneous data, like profile images, could also hold valuable information. The fact that the default profile image or a picture of an inanimate object was used with a Twitter account could for example indicate that a person is trying to hide something. As an extension of the work presented in [18], Table I includes some of the new variables identified. These are indicated with a 'Yes' in the 'New' column.

TABLE I. ADDITIONAL VARIABLES INTRODUCED

Variable	Description	New
R	average number of re-tweets per hour	No
T	total number of tweets	No
W_O	the hash tags most used in the last 200 tweets of the original tweet user	No
TZ	the top time zones of all users	No
A_T	the tweets with the words 'age, 'yr' or 'year' in the actual tweet	Yes
I_{OP}	the users who changed their profile images	Yes

Some results for the 'new variables' are discussed in Table II.

TABLE II. ADDITIONAL VARIABLE RESULTS

Variable	Description
A_T	27,580 out of 265,535 tweets had the words 'age','yr' or 'year' in their tweet text
I_{OP}	73 of 2,812 accounts still have the default profile images

a. Initial insights from observing the new variables

Based on the results from further data interrogation additional interesting information presented it-self. This indicates that these variables are valuable for the experiment at hand. Examples of these insights are:

- More tweets have a time zone allocated than a location. Therefore, time zone could rather be used in investigations surrounding locality of individuals.
- Heterogeneous data, like images, could be an important variable in identity deception and requires further analysis.
- There are 5% of users who did not update their default profile image. Could this perhaps indicate that they are hiding something?

b. Using the initial insights from the observed data for defining an Identity Deception Indicator (IDI_T)

Gleaned from the results of the data interrogation the following represents example IDI_i's that can potentially contribute towards the construction of an IDI_T per online user on social media platforms.

$$IDI_1 = ((W_O) / (hashtags\ (A_T))) * w_1$$

The hash tags of all users who indicated their age (A_T) to be minor are compared to the hash tags (W_O) we already know came from minors to determine deception.

$$IDI_2 = ((tweet\ time \vdash TZ_l) / actual\ tweet\ time) * w_2$$

The appropriate tweet time is determined from the time zone and compared against the actual tweet time. When these differ it could indicate that the user lied about their location.

$$IDI_T = \sum_{n}^{i=1} IDIi$$

Where;

- IDI_i denotes a contributing *IDI*-component; $1 < i < n$
- IDI_T denotes the final *IDI*
- w_i denotes a weighted value

It is envisaged that IDI_T will be the combination of various individually calculated IDI_i's. Each IDI_i can be assigned a different weight reflecting its importance towards contributing to the calculation of IDI_T.

C. Enrichment of variables

It should be noted that the IDI_i's described in the previous section are based on direct data observations. It is expected that expanding the experiment to include, amongst other things, machine learning and heterogeneous input data can generate a more interesting set of IDI_i's for IDI_T. New interesting variables can be introduced with more insight into identity of an online persona resulting in a comprehensive and useful IDI_T.

One such algorithm that was experimented with was the Apriori algorithm. According to the PAL SAP HANA library document, "…Apriori is a classic predictive analysis algorithm for finding association rules used in association analysis. Association analysis uncovers the hidden patterns, correlations or casual structures among a set of items or objects…" [20]. For the experiment it is important to understand which hash tags were used most in combination with other hash tags. This could help identifying a training set of hash tags associated with minors and in doing so allows for the classification of other users as being minors or not.

The result of running the Apriori algorithm over the set of twitter accounts and their last 200 tweets, which contained the hash tags 'school' and 'homework' are shown in Table III. The initial results showed hash tags that were most likely (with a confidence level between 0 and 1) to be found with another hash tag. It can for example be said, with a confidence level of 99%, that a user who used the hash tag '#Read' will next use the hash tag '#Book'. We could use

this similar principle with minors to understand what they are most likely talking about or what they will tag next. The set of hash tags identified will be used as input to train a machine learning algorithm to classify tweets as originating from minors or not. The classification of a tweet as not originating from a minor, in this example, should raise concerns for investigation towards protecting minors on social media platforms.

TABLE III. HASH TAG ASSOCIATION

PRERULE	POSTRULE	SUPPORT	CONFIDENCE	LIFT
Tecno	Tecnología	0.004909180166912126	1	71.295
ATX	#967kissfminvades...	0.0048390490216705 24	1	80.55932203389831
#ShuffleGame1D	EMABiggestFans1D	0.002103934357248054	1	347.7804878048781
#Read	#Book	0.01402622904832036	0.990099009...	70.58910891089108
#UST	#UAAPCDC2015	0.002174065502489655 6	0.96875	84.2280868902439
PushAwardsJaDines	Histo3	0.002033803212006452	0.935483870...	317.5967741935484
AskReddit	Reddit	0.002875376954905673 4	0.931818181...	233.1016746411483
news	generalnews	0.004558524440704117	0.928571428...	129.80882352941177
Tecnología&#Tecno	Tech	0.006101409636019356 5	0.925531914...	114.75790934320075
#Tecno	Tecnología&Tech	0.006101409636019356 5	0.925531914...	114.75790934320075
#Tecno	Tech	0.006101409636019356 5	0.925531914...	114.75790934320075
construction&nyc	newyork	0.002664983519180868	0.904761904...	189.72058823529414
nyc	newyork	0.003927344133529700 5	0.888888888...	186.3921568627451
construction&newy...	nyc	0.002664983519180868	0.863636363...	195.46969696969694

IV. CONCLUSION

Minors are at risk and needs to be protected. A viable option from the research presented in this paper is to work towards an early warning mechanism in the form of an Identity Deception Indicator (*IDI*). Based on the experiment presented, many useful variables were identified towards the creation of an intelligent *IDI*. It appears that much more enrichment and data cleanup is however required to work with a dataset fit for analysis. It was for example found that the word 'Budget Cut' was most prevalent in tweets from the sample set. As this is unrelated to 'school' and 'homework', it is indicative of some advertisement tweets still contained in the sample set. This could skew results and is unimportant to the experiment as we are only interested in protecting minors from actual people. The data cleanup will be addressed next and thereafter further work will be done towards applying machine learning and deep learning for variable enrichment. Additional machine learning techniques will be investigated to understand their usefulness in creating an *IDI*. It is envisaged that the experiment will culminate into a more mature *IDI* for social media related big data sets, which are complex in nature.

ACKNOWLEDGMENT

The support of "The HPI Future SOC"-research Lab in Potsdam (Germany) is acknowledged for making available powerful infrastructure to conduct the research activities as presented in this paper.

REFERENCES

[1] A. Vrij, "Guidelines to catch a liar," 2004.
[2] M. Tsikerdekis and S. Zeadally, "Detecting and Preventing Online Identity Deception in Social Networking Services," Internet Computing, IEEE, vol. 19, pp. 41-49, 2015.
[3] D. Bogdanova, P. Rosso, and T. Solorio, "Exploring high-level features for detecting cyberpedophilia," Computer Speech & Language, vol. 28, pp. 108-120, 2014.
[4] D. Laney, "3D Data Management: Controlling data volume, velocity and variety," ed: Meta Group, 2001.
[5] L. Cai and Y. Zhu, "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era," Data Science Journal, vol. 14, p. 2, 2015.
[6] M. D. Assunção, R. N. Calheiros, S. Bianchi, M. A. Netto, and R. Buyya, "Big Data computing and clouds: Trends and future directions," Journal of Parallel and Distributed Computing, vol. 79, pp. 3-15, 2015.
[7] M. Ali-ud-din Khan, M. F. Uddin, and N. Gupta, "Seven V's of Big Data understanding Big Data to extract value," in American Society for Engineering Education (ASEE Zone 1), 2014 Zone 1 Conference of the, 2014, pp. 1-5.
[8] R. Han, Z. Jia, W. Gao, X. Tian, and L. Wang, "Benchmarking Big Data Systems: State-of-the-Art and Future Directions," arXiv preprint arXiv:1506.01494, 2015.
[9] S. Kierkegaard, "Cybering, online grooming and ageplay," Computer Law & Security Review, vol. 24, pp. 41-55, 2008.
[10] I. Liccardi, M. Bulger, H. Abelson, D. J. Weitzner, and W. Mackay, "Can apps play by the COPPA Rules?," in Privacy, Security and Trust (PST), 2014 Twelfth Annual International Conference on, 2014, pp. 1-9.
[11] C. Peersman, W. Daelemans, and L. Van Vaerenbergh, "Predicting age and gender in online social networks," in Proceedings of the 3rd international workshop on Search and mining user-generated contents, 2011, pp. 37-44.
[12] G. S. O'Keeffe and K. Clarke-Pearson, "The impact of social media on children, adolescents, and families," Pediatrics, vol. 127, pp. 800-804, 2011.
[13] J. T. Hancock, J. Thom-Santelli, and T. Ritchie, "Deception and design: The impact of communication technology on lying behavior," in Proceedings of the SIGCHI conference on Human factors in computing systems, 2004, pp. 129-134.
[14] J. S. Alowibdi, U. A. Buy, S. Y. Philip, S. Ghani, and M. Mokbel, "Deception detection in Twitter," Social Network Analysis and Mining, vol. 5, pp. 1-16, 2015.
[15] J. S. Alowibdi, U. Buy, P. S. Yu, and L. Stenneth, "Detecting Deception in Online Social Networks," in Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on, 2014, pp. 383-390.
[16] J. Guillory and J. T. Hancock, "The effect of Linkedin on deception in resumes," Cyberpsychology, Behavior, and Social Networking, vol. 15, pp. 135-140, 2012.
[17] G. A. Wang, H. Chen, J. J. Xu, and H. Atabakhsh, "Automatically detecting criminal identity deception: an adaptive detection algorithm," Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, vol. 36, pp. 988-999, 2006.
[18] E. Van der Walt and J. H. P. Eloff, "Protecting minors on social media platforms - A Big Data Science experiment " presented at the HPI Cloud Symposium "Operating the Cloud", Potsdam, Germany, 2015.
[19] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, et al., "Personality, gender, and age in the language of social media: The open-vocabulary approach," PloS one, vol. 8, p. e73791, 2013.
[20] SAP, "SAP HANA predictive analytics library," SAP HANA Platform SPS 10, vol. Document Revision 1.0, 21 Aug 2015 2015.