Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/cose

Computers & Security

Cyber-security: Identity deception detection on social media platforms

Estee van der Walt[a,*], J.H.P. Eloff[a], Jacomine Grobler[b]

[a] Department of Computer Science, Information Technology Building - Level 4, University of Pretoria, Lynnwood Road, Pretoria, South Africa
[b] Department of Industrial and Systems Engineering, Engineering building 2 Level 3, University of Pretoria, Lynnwood Road Pretoria, South Africa

ARTICLE INFO

Article history:
Received 17 January 2018
Revised 11 April 2018
Accepted 31 May 2018
Available online 13 June 2018

Keywords:
Cyber-security
Identity deception
Social media
Big data
Bots
Grooming

ABSTRACT

Social media platforms allow billions of individuals to share their thoughts, likes and dislikes in real-time, without any censorship. This freedom, however, comes at a cyber-security risk. Cyber threats are more difficult to detect in a cyber world where anonymity and false identities are ever-present. The speed at which these deceptive identities evolve calls for solutions to detect identity deception. Cyber-security threats caused by humans on social media platforms are widespread and warrant attention. This research posits a solution towards the intelligent detection of deceptive identities contrived by human individuals on social media platforms (SMPs). Firstly, this research evaluates machine learning models by using attributes such as the “profile image” found on SMPs. To improve on the results delivered by these models, past research findings from the field of psychology, such as that humans lie about their gender, are used. Newly engineered features such as “gender-derived-from-the-profile-image” are evaluated to grasp whether these features detect deception with greater accuracy. Furthermore, research results from detecting non-human (also known as bot) accounts are also leveraged to improve on the initial results. These machine learning results are lastly applied to a proposed model for the intelligent detection and interpretation of identity deception on SMPs. This paper shows that the cyber-security threat of identity deception can potentially be minimized, should the vulnerability in the current way of setting up user accounts on SMPs be re-engineered in the future.

1. Introduction

Nowadays, the cyber-security media are very concerned about people being exposed to all sorts of abuse on Social Media Platforms (SMPs). The malicious intent of humans deceiving other humans constitutes a cyber threat that is one of the most difficult to contend. More importantly, these cyber threats are aggravated by the sheer number of vulnerabilities present in SMPs, the number of available and different types of SMPs (Chaffey, 2016), the poor design and construction of SMPs (Haimson and Hoffmann, 2016), the large volumes of unstructured content (Assunção et al., 2015), and the opportunities that SMPs provide to humans acting in malicious ways (Fire et al., 2014). These factors all contribute to SMPs being extremely vulnerable to cyber threats caused by malicious users. Furthermore, as a result of these cyber threats and SMP vulnerabilities we witness an alarming increase in the prevalence of cyber bullying (Smit, 2015), identity theft

[*] Corresponding author.
E-mail addresses: u14447984@up.ac.za (E. van der Walt), eloff@cs.up.ac.za (J.H.P. Eloff), jacomine.grobler@up.ac.za (J. Grobler).

https://doi.org/10.1016/j.cose.2018.05.015

(Kabay et al., 2014), identity impersonation (Galán-García et al., 2016), dissemination of pornography (Benevenuto et al., 2010), fraud (Gurajala et al., 2015), and the like. As an example, consider a recent cyber-security case reported on in South Africa where, as in any other country in the world, an alarming increase has been noted in cyber threats related to abuse against women (Bliss, 2017). The cyber threat in this case manifested itself in the form of identity impersonation by two malicious users who exploited the "ease-of-opening-a-deceptive-account" vulnerability on Facebook and were arrested for luring, raping, and killing women (de Villiers, 2017). This and other cyber-security cases (Peterson, 2016); (Digital, 2016) point to a common thread in exploiting SMP vulnerabilities, namely the ease of creating fake or deceptive identities (Tsikerdekis and Zeadally, 2014).

In the case of identity deception, a deceptive account is either created with malicious intent or to preserve anonymity. This paper is concerned with the detection of deceptive accounts created with malicious intent, as these pose a cyber threat to other humans at large. A deceptive account with malicious intent could for example be used to defame someone's character (Galán-García et al., 2016) or conduct online bullying (Smit, 2015). These deceptive accounts are generated by humans or bots (Chu et al., 2010). Much research (Oentaryo et al., 2016); (Dickerson et al., 2014); (Cresci et al., 2015) has been done to detect bot accounts that require no human involvement for the actions they perform. These deceptive bot accounts are known to target groups, as opposed to specific individuals (Oentaryo et al., 2016). However, to date, very little research has focused on detecting deceptive human accounts on SMPs.

The research reported on in this paper is a first attempt at minimising the cyber risk of identity deception as exploited by malicious users on SMPs through the intelligent detection of deceptive identities. The aims of the research reported on in this paper are summarised as follows:

- To identify and describe the different types of information available on SMPs – also referred to as attributes (e.g. the date on which the account is created) – that can potentially be used to detect user identity deception.
- To experiment with SMP attributes to detect, in an intelligent way, user identity deception by using various machine learning models.
- To enhance the SMP attributes for improving identity deception detection by using engineered features derived from two fields: Firstly, the field of psychology where we focus specifically on discovering why people lie, and secondly, from the field of detecting bot accounts where we intend to determine how these solutions can be leveraged for detecting human identity deception.
- To propose a model that intelligently detects and interprets the perceived deceptiveness of a SMP user, given the results from the aforementioned experiments.

This will be the first time that features derived from the field of psychology will be applied towards the detection of human identity deception on SMPs.

The rest of the paper is organised as follows: Section 2 describes background and related work, Section 3 describes how the SMP data was mined and prepared for experimentation towards detecting human identity deception, Section 4 presents the results from the aforementioned experiments, Section 5 describes the proposed identity deception detection model, and finally, Section 6 concludes with the overall research findings and scope for potential future work.

2. Background and related work

Cyber threats are widespread, with SMPs being an enabler for cyber attacks (Khandpur et al., 2017). SMPs are vulnerable to cyber threats as they breed trust between individuals without any authority validating or verifying the participants. Cyber crime can potentially have severe consequences. For example, a 14-year-old boy from the UK was groomed online and later killed (Camber, 2014), having been promised "great wealth".

DePaulo et al. (1996) profess that humans are known to lie, for instance about their gender or their age (Drouin et al., 2016). When humans lie about attributes that distinguish them from other humans, it is known as identity deception (Wang et al., 2006).

Much of the past research on identity deception among humans has been psychological in nature and opposing views have been proposed on why humans lie. For example, Halevy et al. (2014) believe that most humans are honest most of the time, whereas DePaulo et al. (1996) are adamant that most humans lie daily, to varying degrees, but mostly in small quantities. Ferrara et al. (2016) believe that the act of deception is deliberate and intended to further a specific goal, such as to recruit other humans for terrorism. Rong et al. (2016) show that incentives such as reward schemes can result in lies being more prevalent. Whilst researchers will continue to debate about when and why humans lie, consensus remains that the act of lying is present. This paper proposes to identify those humans who deceive others on SMPs – with malicious intent – as posited by Ferrara et al. (2016).

For this research past research in psychology was considered to identify those attributes about which humans are most likely to lie; more specifically attributes pertaining to their identity. The hope is that human nature prevails on SMPs and that humans continue to lie, regardless of the medium of communication. Table 1 summarises the conclusions from past research by showing the various identity attributes humans lie about. Evidently, humans lie most often about their image, name, location, age and gender.

From a psychopathological perspective, Stanton et al. (2016) explored whether personality can explain deception such as changing your name or image online. They found that feelings of inadequacy and self-dissatisfaction often lead to deception. Caspi and Gorsky (2006) explored the emotions experienced during deception by using input from different demographics like location, age, gender, marital status, and occupation. They found that identity roleplay and privacy concerns were the main reason for humans being deceptive.

From an online perspective, Hancock (2007) depicted identity-based and message-based deception as two main types of digital deception. He presented a detailed review on

Table 1 – Identity attributes lied about, based on psychological research.

Researcher(s)	Image	Name	Location	Ethnicity	Age	Gender	Marital status	Occupation	Qualification	Appearance
(Stanton et al., 2016)	x	x								
(Jupe et al., 2016)			x					x	x	
(Hancock, 2007)	x	x		x		x				
(Caspi and Gorsky, 2006)			x		x	x	x	x		
(Utz, 2005)		x			x	x				x
(Toma et al., 2008)	x		x		x	x				
(Hancock and Toma, 2009)	x									
(Wang et al., 2006)		x	x		x					

why and how humans lie and concluded that deception on online platforms could be more difficult to detect than face-to-face deception. Utz (2005) defined the most common types of deception to be gender switching, identity concealment, and attractiveness deception. He also showed that these deceptive actions could be ascribed to different motivations.

Online dating deception has been the focus of attention of various researchers. For example, Toma et al. (2008) investigated whether humans present themselves truthfully in their online dating profiles. They found that people deliberately deceive and concluded that deception on certain identity attributes such as image, location, age and gender, are more prevalent. (Hancock and Toma, 2009) did similar research on online dating deception but focused on the images presented on these online dating profiles alone. They found that although users often present deceptive pictures, they try to remain authentic as far as possible. For example, users tend to present an image of their younger self.

Besides online deception, identity deception also occurs in other areas such as job interviews and criminology. Jupe et al. (2016) investigated whether verifiable detail provided during a job interview could successfully distinguish humans telling the truth from those who lie. Wang et al. (2006) considered past criminal records and compared the data provided by the criminals with the true data. The knowledge they gained presented a framework to indicate the identity attributes about which these criminals were most likely to lie. It was found that criminals most frequently lied about their name.

Similar identity attributes as those depicted in Table 1 and identified in past psychological research are found to be lied about on SMPs. Appendix A contains a comparison of the attributes identified on the top six SMPs of 2016 (Chaffey, 2016). These attributes can be grouped according to those describing

- the user's account profile, for example his/her profile image;
- information about the account, for example its opening date;
- the behaviour of the user, for example the time at which he/she posted a message on the SMP;
- the user's relationships, for example his/her friends; and
- the content of the user's posts, for example tweets on Twitter.

For the research in hand, the attributes found in the person's Twitter account were used to detect human identity deception. Twitter is the only platform where no consent from the account holder is required to gather data. By contrast, on Facebook, permission is required from each person before their data can be gathered. (This permission can come in the form of an accepted friend request or the person having opted to become part of a group.) Due to the accessibility of its data, Twitter has been used in many research projects across various disciplines, including for identity deception detection (Gurajala et al., 2016) (Alowibdi et al., 2015). However, since a significant overlap occurs between the attributes of various SMPs (such as the profile image and name), research such as the current, which makes use of data gathered from Twitter, can potentially apply also to another SMP.

Past researchers used the attributes found on SMPs to build new features with which identity deception can be detected. Promising work by Alowibdi et al. (2015) examines identity deception features by detecting inconsistencies in the gender and the expected background colour chosen for the human's account. Alowibdi et al. (2015) also found statistical inconsistencies in geo-location update times that were useful for the detection of deceptive accounts. Tuna et al. (2016) focus on deriving features such as gender and location from the language and the local text used in the content respectively. Other features, such as whether a profile image represents the user truthfully (Hancock and Toma, 2009), similarity of attributes such as name, given the Levenshtein difference (Li and Wang, 2015), and the emotional state of a user given the content they post (Bogdanova et al., 2014), have also been posited as being potentially useful for detecting identity deception.

In addition, researchers proposed various techniques to detect identity deception on SMPs. These techniques included filtering (Thomas et al., 2011), rules (Fire et al., 2014), supervised machine learning (Cresci et al., 2015), semi-supervised machine learning (Ebrahimi et al., 2016), reinforcement learning (Venkatesan et al., 2017), and unsupervised machine learning (Gu et al., 2008). For the purposes of this research, we focused on supervised machine learning, as the problem at hand is one of classifying whether human accounts are to be classified as "deceptive" or "not" . Classification problems are typically solved with supervised machine learning (Ma et al., 2014), which is similar to research done in the past to detect bots on SMPs (Cresci et al., 2015); (Oentaryo et al., 2016); (Dickerson et al., 2014). They presented supervised machine learning to solve the problem in classifying an account as "bot" or "not", which has synergies with the current research. Therefore, past research in the detection of bots was used to identify appropriate supervised machine learning algorithms for the research under consideration. In machine learning,

Table 2 – Supervised machine learning models.

ML algorithm name	Related research	ML algorithm description
Adaboost	(Fire et al., 2014); (Bellinger et al., 2012)	Adaptive boosting combines the weak results from various decision trees to create a boosted classifier.
bayesglm	(Sedhai and Sun, 2017); (Choudhary and Jain, 2017)	The Bayesian-generalised linear algorithm uses simple logistic regression.
J48	(Choudhary and Jain, 2017); (Galán-García et al., 2016)	J48 is a version of a decision tree algorithm.
kknn	(Ebrahimi et al., 2016); (Al-garadi et al., 2016)	K nearest neighbours use clustering to group and predict similar classifications.
nnet	(Rubin, 2017); (Tuteja, 2016)	A neural network simulates the neurons in a brain to classify.
rf	(Galán-García et al., 2016); (Tsikerdekis, 2017)	Random forests build a number of decision trees to find the best accuracy with one such tree.
rpart	(Dal Pozzolo et al., 2013); (Genuer et al., 2010)	The recursive partitioning tree is a very basic representation of a decision tree.
svmLinear	(Tsikerdekis, 2017); (Peddinti et al., 2017)	SVMs use a high-dimensional feature space to classify.

no universal algorithm is expected to outperform the rest (Wolpert and Macready, 1997), also referred to as “no free lunch” theorem. Hence, the need exists to test various algorithms on the same problem. Eight machine learning models, found in bot detection research and shown in Table 2, were applied in this research to detect deceptive humans.

However, a problem with the results from supervised machine learning models is that these results are seldom interpretable and intuitive (Lipton, 2016). The *Financial Times* (Anonymous, 2017) recently reported on the criticality of being able to explain decisions made from artificial intelligence (AI), as garbage inputs could result in garbage outputs without forewarning. The correct interpretation is for example critical when the detection of deceptive users could have criminal consequences (Burrell, 2016) or a serious impact on people’s lives (Ribeiro et al., 2016). The interpretation or explanation of results will become law in the European Union during 2018 and ensure that every European citizen has the right to an explanation where decisions were based on data and algorithms (Goodman and Flaxman, 2016).

The research reported on in this paper introduces the notion of entropy as a potential solution to explain the results from supervised machine learning. In general, entropy refers to uncertainty. Claude Shannon already introduced information entropy in 1948 (Shannon, 2001), where initially it was applied to information compression by determining the quantity of information which can be discarded during transmission before a message becomes irretrievable. This same concept has since been applied to the input of machine learning models (Gurajala et al., 2015), where entropy indicates how much information is gained or lost when input is added or discarded to train machine learning models. The entropy result is helpful to determine the importance of each input for the expected outcome and accuracy of a machine learning model.

3. Preparing data for the experiments

3.1. *Gathering the data*

Twitter data was mined to create an initial corpus of social media accounts. The sheer volume of data on Twitter makes mining all account data since the SMP’s inception in 2006 unfeasible and impractical for the research at hand. For the purposes of the corpus we chose to limit the data to a demographic known to be the target of deceptive users. Minors are susceptible to cyber bullying (Galán-García et al., 2016), extremist recruitment (Klausen, 2015), and grooming (Kierkegaard, 2008), among others. Therefore, the corpus was limited to accounts that used the words “school” and “homework”, as these are words used widely by minors (Schwartz et al., 2013). The friends and followers of these accounts were also mined, as it is known that friends usually have similar friends (Cook, 2014) – in this case, more minors.

223 796 Twitter accounts were gathered over a six-month period starting in June 2016. This data, which is publicly available, showed that the accounts were created between 2006 and 2017 and were still actively contributing to Twitter.

3.2. *Cleaning the data*

Since the current research is focused on addressing deception by human users, an attempt was made to rid the corpus of non-human accounts included in the initial gathered corpus. To this end, research work presented by Cresci et al. (2015), which has shown good results in identifying bots or non-humans, was applied to clean the data. Cresci et al. identified three sets of rules to distinguish humans from bots. Their top three rules were applied to clean the mined corpus as these rules had an accuracy of over 75%. Their rules required that the account must have

- at least 30 followers;
- at least 50 tweets; and
- replied to at least one direct tweet from another user.

Known celebrities (Twitter, 2017) were also removed from the corpus, as these accounts have been validated by Twitter as being trustworthy. Of the original corpus of 223,796 Twitter accounts, 69,279 were deemed to be known non-human or verified celebrity accounts. Although some bot accounts might still remain in the cleaned corpus, the researchers believe that the rules applied would have removed most bots with an accuracy of over 75% (Cresci et al., 2015).

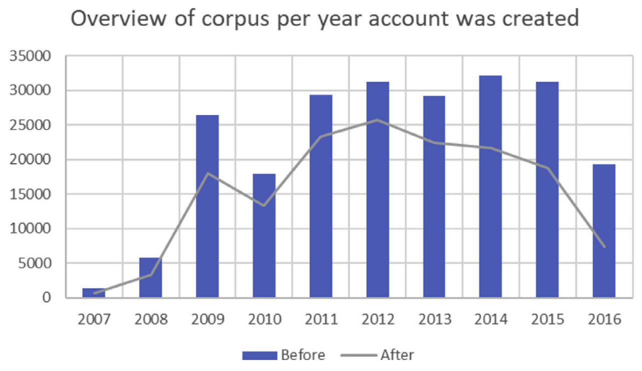


Fig. 1 – Data before and after cleaning.

Fig. 1 shows the breakdown of the corpus before and after cleaning. The year indicates when the account was created (i.e. the age of the account). It is noteworthy that the cleaning had a greater effect on accounts created after 2013 than earlier accounts, as the emergence of non-human or bot accounts on SMPs is a relatively recent phenomenon (Oentaryo et al., 2016). After cleaning, 154,517 accounts remained in the corpus.

Finally, certain attributes were also removed to reduce variance and bias. Variance reflects the tendency to learn random things unrelated to the problem (Yarkoni and Westfall, 2018). An example would be if the corpus includes data unrelated to the problem that is being solved, for example the ID assigned to an account by Twitter. Bias is the tendency to learn the same wrong thing (Yarkoni and Westfall, 2018), for example the background image attribute (empty for the gathered corpus and therefore assumed to be always empty) – which is not the case. The following attributes were removed and the reason for removal is shown in brackets:

- Where the attributes were unique to a specific account. – for example, the ID, name, and account description (variance).
- All remaining zero variance attributes were removed, in other words data with a remarkably high ratio of uniqueness – for example, longitude, latitude, and location (variance).
- Data that was mostly empty in the corpus – for example, background image and background colour (bias).

3.3. *Injecting deceptive accounts*

Supervised machine learning requires a labelled dataset for training purposes. For this reason, known deceptive accounts were injected into the original corpus. There is however a challenge in finding examples of deceptive human accounts. Zafarani and Liu (2015) suggested manual crowd-sourcing mechanisms, like Amazon's Mechanical Turk (Amazon, 2017), to label accounts and classify a predicted outcome. Furthermore, there are known means for user groups to oust malicious accounts, like accounts linked to terrorism (Ferrara et al., 2016). Lastly, Peddinti et al. (2017) used labelled datasets from their own previous research work to identify sensitive accounts referring to, for example, topics about pregnancy and paedophilia. None of these options were viable solutions to be used in the current research, due to the absence of known deceptive human accounts and expertise required to manually identify deceptive human accounts correctly as a group. Due to these challenges (if not at all impossible to collect confirmed deceptive accounts in the real world on the Twitter SMP), the choice was made to fabricate deceptive accounts.

The injected deceptive accounts were created using two random human data generator APIs from the internet (Armstrong and Hunt, 2017); (Keen, 2017). Further manual intervention was required to complete the remaining attributes which the APIs could not do. An example of manual data injection was the number of friends and followers of an account. Values were chosen such that they were similar to what was observed within the bounds of the current mined corpus. The deceptive accounts were classified as "deceptive" and the original corpus as "trustworthy".

In an absence of deceptive human accounts and for the sake of the validity of the research, it was decided to ensure that the fabricated deceptive accounts are as far as possible aligned with the data contained in the original corpus. This was done to make the research results as realistic as possible. Most importantly, the following two statistical tests were employed to validate that the injected deceptive accounts were still representative of the original mined corpus:

- *Wilcoxon Signed Rank test, also known as the Mann-Whitney-U test* (Kothari, 2004). This test compares the sum of ranks, or indirectly the medians, of two sets of distributions. If the means are similar, the data can be assumed to be from the same population. This test does not require data to be normally distributed or sample sizes to be the same (Mann and Whitney, 1947). For example, evaluate the distribution of one attribute, like "number_of_friends" for both the "deceptive" and "trustworthy" corpus. If both distributions are found to be similar, they are believed to represent similar data. If all attributes individually pass the Wilcoxon test, both datasets can be said to be from the same population, which is in this case Twitter.
- *Pearson's chi-square test of independence* (Kothari, 2004). This test assumes that subjects in a single population are classified similarly. This test will show when attributes in the population are correlated, and thus from the same population. This test works well when samples were generated at random; the attributes were categorical; and the resultant categories were greater than 5 (McDonald, 2009). Per example, evaluate the correlation between one attribute, such as "number_of_friends" for both the "deceptive" and "trustworthy" corpus. If the attribute is found to be highly correlated, it can be said that attribute contain similar data. If all attributes succeed in the Pearson's Chi square test, it can be said that both datasets are from the same population, in this case Twitter.

For this research the injected deceptive accounts were not only representative of data found in Twitter, but also had to be actually deceptive. For this reason, past research from the field of psychology, highlighting identity attributes humans are known to lie about were analysed. It was found that humans lie most often about their image, name, location, age, and gender. Therefore, we confirmed that each injected deceptive account was deceptive in respect of each of these attributes by applying rules to test for deceptiveness. An example was to ensure that the name used for an account and its pseudonym was never the same. Another was that the age detected in the user's image was different from their actual age. By ensuring that all attributes, as per psychological identity deception research results, were deceptive, each account was created to be as deceptive as possible – even though humans might lie only about some of these attributes in the real world.

Over 15,000 fabricated deceptive accounts were injected, which constituted almost 10% of the corpus. Halevy et al. (2014) found that 5% of people tell 40% of all lies. With the introduction of 10% fabricated deceptive accounts most lies should be catered for.

3.4. *Preparing data for machine learning*

Before any machine learning models can be trained towards identity deception, the data must be in the correct format. Most machine learning models expect data to be discretised, centred, and scaled (Kuhn et al., 2016). Discretisation implies that numerical data is converted to categorical data. An example is if the number of friends is grouped into bins of 500. The result would be accounts falling into the ranges of 0–500, 501–1000, and so on. All nominal values are then centred. For centring, the sample mean is subtracted. For example, if the mean of the "number_of_friends" is 1,500 for the total corpus, this value will be subtracted from each account for their respective "number_of_friends". Lastly, these centred values are divided by the standard deviation. This ensures that all input is similarly scaled and will not introduce bias if the values for other inputs are higher. An example is where the number of friends is on a different scale initially from the number of tweets or posts. The proposed scaling method ensures that machine learning models will treat both inputs as equally important.

4. Detecting identity deception

As described earlier, a corpus was created by gathering data from Twitter, cleaning it, and injecting deceptive accounts. The next step involved using the corpus as input to train machine learning models in detecting identity deception. For this, two experiments were defined:

- Experiment 1: Only data from the corpus was used to detect identity deception. This data was based on the original attributes as found in Twitter, for example their "number_of_friends", also denoted as FRIENDS_COUNT in Twitter.
- Experiment 2: The original attributes used in Experiment 1 were then extended upon with new engineered features. These features were engineered from psychological principles that identify deception and included previously engineered features that were applied to detect non-human or bot accounts. An example of such a feature is "gender". This feature uses the original SMP attributes, namely the name and profile image of the account. The "gender" feature shows the correlation between the gender derived from each individual attribute. The intention of the second experiment was to evaluate whether these features could possibly improve the accuracy of identity deception detection by humans on SMPs (results of Experiment 1).

Table 3 – Machine learning results for Experiment 1.

ML model	Accuracy (%)	F1 Score (%)	PR-AUC (%)	Cost
svmLinear	16.01	16.28	8.47	68.257
rf	79.94	32.92	29.25	101.842
J48	70.07	28.08	21.79	110.317
bayesglm	66.75	15.12	9.84	4.047
kknn	71.17	23.91	13.30	62.317
Adaboost	77.89	29.51	32.27	891.145
rpart	66.03	23.65	13.55	3.759
nnet	63.34	25.74	32.05	38.897

4.1. *Results of detecting identity deception*

4.1.1. *Experiment 1*

After removing attributes from the corpus to avoid bias and variance, only a few viable SMP attributes remained given the original attributes described in appendix A. Regardless, the previously identified supervised machine learning algorithms were trained to determine whether these attributes could detect human identity deception. These machine learning algorithms were trained using 3-repeat, 10-fold, cross validation (Anguita et al., 2012) and the default parameters as defined by the caret package in R (Kuhn et al., 2016). The results from these machine learning models, built from the attributes in Twitter only, are shown in Table 3. The accuracy measure can be deceptive in skewed datasets such as the current corpus. The F1 score was selected as the indicative metric of performance as it includes precision (the number of the predicted deceptive accounts that were actually deceptive) and recall (the number of actual deceptive accounts which were identified correctly). The PR-AUC metric shows the area under the precision recall curve. Lastly, cost was added as an additional metric to describe how many seconds were required to execute each machine learning model.

The results in Table 3 show that the random forest (rf) model detected identity deception with a F1 score of 32.92%. The Adaboost model performed next best, with an F1 score of 29.51%. Thus, the rf model could be preferred to the Adaboost model, as the former can be trained much faster and shows better results. Table 3 also shows that the attributes used to train the machine learning models did not predict human identity deception successfully. The machine learning models performed worse than selecting the prediction at random (which has a 50% chance of success).

Much could nonetheless be learnt from these results – more specifically, which attributes were more important than others. This was achieved by investigating the entropy of each attribute used to train the machine learning model. Entropy (Shannon, 2001) indicates how much information is gained – in this case a better detection of identity deception – by introducing that specific attribute. If p is the probability of A, given n attributes; then entropy can be calculated as follows:

$$H(A) = -\sum_{i=1}^{n} p_i \log_2 pi$$

Table 4 – Entropy results for Experiment 1.

Attributes	rf entropy result	J48 entropy result	Adaboost entropy result
FOLLOWERS_COUNT	100.00	0.00	0.00
FRIENDS_COUNT	80.88	72.30	72.30
LISTED_COUNT	0.00	28.61	28.61
STATUS_COUNT	79.11	44.61	44.61
TIMEZONE	41.43	100.00	100.00

(Shannon, 2001)

Table 4 shows the entropy results of the top three machine learning models based on the F1 scores. The entropy values are depicted as a value between 0 and 100, with 100 indicating that the model was highly dependent on the attribute and 0 meaning that the attribute had very little influence on the outcome.

The entropy results in this experiment showed that all attributes, except for the LISTED_COUNT attribute, played some part in the predictions at some point or another. According to Cresci et al. (2015), these attributes were also known to successfully detect bot accounts; hence attributes previously used in the detection of bots can be applied equally well to detect human deception. The bot attributes and new features engineered from psychology will be introduced during Experiment 2 in the hope of improving the results achieved during Experiment 1.

4.1.2. *Experiment 2*

For the second experiment, the original corpus data (based on original Twitter attributes) was extended with new engineered features. The entropy results from Experiment 1 already suggested that past research in bot detection could be used to detect human identity deception (Van der Walt and Eloff, 2018). In addition, the field of psychology was consulted to engineer features that could indicate human identity deception. These features were engineered using the attributes of Twitter as presented in Appendix A. Table 5 lists all newly engineered features, whether originating from past research into bot detection or from the field of psychology. It also shows which Twitter attributes were used and/or combined to construct the feature.

Next follows a brief description of each attribute:

- ACCOUNT_AGE_IN_MONTHS – This feature shows the number of months since the account was opened.
- AGE – For this feature, the discrepancy between the age of the account and the age of the user was calculated. If AGE contributes to deception, the calculated ages for deceptive accounts will be different from those of in the trustworthy corpus. The age of the user was determined using the Google Vision API (Google, 2017). This API can extract faces and their age from any given image by using Google's own proprietary machine learning models for those accounts that have images.
- GENDER – Google Vision API (Google, 2017) was used to determine whether the face shown on a profile image was male or female. Combining this knowledge with a name database (of male and female names) creates a feature that can compare the gender of an image to the gender of a name.
- DISTANCE_LOCATION – In Twitter, the geo-tag of the last tweet is stored for geo-enabled users. This feature uses the geo-location lookup API to retrieve the location given in the status of the user. This location is compared with the geo-tag. The Haversine distance (Van Liere, 2010) (Shumaker and Sinnott, 1984) between the two determines the feature value.
- DISTANCE_TZ - Twitter also stores the time zone of the user. This time zone is captured during registration and can be changed by users at any time. The feature employs the geo-location lookup API to retrieve the location given in the status of the user and then compares it with the stated time zone of the user. The Haversine distance (Van Liere, 2010) (Shumaker and Sinnott, 1984) between the two determines the feature.
- DUP_PROFILE – This feature shows whether the current account has a similar profile description as another.
- FRIENDS_VS_FOLLOWERS – The ratio of friends vs followers is determined for this feature.
- FOLLOWERS_COUNT – In Twitter, the number of followers is recorded for each account. This value was discretised for the current experiment.
- FRIENDS_COUNT – The number of friends – also recorded for each account in Twitter – was discretised for the experiment in hand.
- GEO_ENABLED – Twitter stores whether an account is enabled to store its location in terms of longitude and latitude.
- HAS_IMAGE – This feature is constructed as a binary indicator and shows whether a profile image has been defined for an account or whether the account is still using the default Twitter image as its profile.
- HAS_NAME – This feature is constructed as a binary indicator showing whether the name could be found in a name database.
- HAS_PROFILE – This feature is constructed as a binary indicator showing whether the account has a description or not.
- NAME - The Levenshtein distance (Li and Wang, 2015) (Levenshtein, 1966) between the screen-name and registered username.
- LISTED_COUNT – In Twitter the number of public lists the account belongs to is recorded. This value is discretized for the experiment at hand.
- PROFILE_HAS_URL – This feature is constructed as a binary indicator showing whether the account's description contains an URL or not.
- TWEET_COUNT – The number of tweets posted by the account are discretized for the experiment at hand.
- NAME_LENGTH – The number of characters contained in the screen name or pseudonym of the account.

Together, these features were used to train the same 8 supervised machine leaning models used for experiment 1. The results from these machine learning models are shown in Table 6. The Adaboost model achieved an F1 score of 84.65% and random forest (rf) achieved an F1 score of 86.24%. These

Table 5 – Features engineered for Experiment 2.

Feature	Origin	Constructed from these Twitter attributes
ACCOUNT_AGE_IN_MONTHS	Bot	created_at
AGE	Psychology	created_at, profile_image
GENDER	Psychology	name, profile_image
DISTANCE_LOCATION	Psychology	location, latitude, longitude
DISTANCE_TZ	Psychology	location, time_zone
DUP_PROFILE	Bot	description
FRIENDS_VS_FOLLOWERS	Bot	friends_count, followers_counts
FOLLOWERS_COUNT	Bot	followers_count
FRIENDS_COUNT	Bot	friends_count
GEO_ENABLED	Bot	geo_enabled
HAS_IMAGE	Bot	profile_image
HAS_NAME	Bot	Name
HAS_PROFILE	Bot	description
NAME	Psychology	name, screen_name
LISTED_COUNT	Bot	listed_count
PROFILE_HAS_URL	Bot	description
TWEET_COUNT	Bot	status_count
NAME_LENGTH	Bot	screen_name

Table 6 – Machine learning results for Experiment 2.

ML Model	Accuracy (%)	F1 Score (%)	PR-AUC (%)	Cost
svmLinear	92.20	66.29	76.80	45.198
rf	97.49	86.24	93.00	157.801
J48	95.79	79.05	64.94	178.649
bayesglm	92.07	65.85	77.01	6.932
kknn	94.18	72.93	81.32	80.933
Adaboost	97.03	84.65	93.70	2127.87
rpart	87.32	55.83	38.29	5.091
nnet	95.21	76.94	87.76	62.796

results were considerably better than those achieved using Twitter attributes alone (Experiment 1).

Table 7 shows the entropy results of the top three machine learning models. These results show which features contributed most significantly in the final machine learning model. AGE, NAME, HAS_NAME, HAS_PROFILE, DUP_PROFILE, NAME_LENGTH, and DISTANCE_TZ showed most promise as features indicative of human identity deception. These features coincide with what we know from psychology in that humans lie about the image, name, location, and age. It seems that gender was a poor indicator of deceptiveness. This could be due to the fact that the images in Twitter are in general misrepresentative of the users, and the trained models were thus unable to use this specific feature to identify human identity deception.

5. Identity Deception Detection Model (IDDM) for SMPs

IDDM provides automated assistance for the detection of identity deception on SMPs. Based on the experimental results discussed in the previous sections IDDM is structured to consist of the following 2 sub-models:

Table 7 – Entropy results for Experiment 2.

Features	rf entropy result	J48 entropy result	Adaboost entropy result
ACCOUNT_AGE_IN_MONTHS	15.23	37.66	37.66
AGE	**100.00**	**100.00**	**100.00**
GENDER	17.80	12.29	12.29
DISTANCE_LOCATION	0.13	0.95	0.95
DISTANCE_TZ	**18.61**	**53.13**	**53.13**
DUP_PROFILE	**28.24**	**64.34**	**64.34**
FRIENDS_VS_FOLLOWERS	0.33	0.90	0.90
FOLLOWERS_COUNT	8.41	8.50	8.50
FRIENDS_COUNT	6.12	17.10	17.10
GEO_ENABLED	4.81	13.80	13.80
HAS_IMAGE	0.00	0.72	0.72
HAS_NAME	**57.83**	**79.91**	**79.91**
HAS_PROFILE	**26.16**	**61.59**	**61.59**
NAME	**59.27**	**81.55**	**81.55**
LISTED_COUNT	0.63	0.00	0.00
PROFILE_HAS_URL	4.21	9.86	9.86
TWEET_COUNT	7.92	8.86	8.86
NAME_LENGTH	**25.59**	**27.74**	**27.74**

(1) Identity Deception Detection Machine Learning Model (IDDMLM). The IDDMLM employs machine learning to identify appropriate attributes and features of identity related information on SMPs. IDDMLM calculates accuracy and entropy information of these attributes and features. IDDMLM determines if an identity is deceptive or not. Because of its machine learning nature, it provides little interpretation as to why an identity is perceived as deceptive or not.

(2) Identity Deception Detection Score Model (IDDSM). The IDDSM uses the outputs of the IDDMLM. These outputs include the accuracy and entropy related information about attributes and features. The entropy information is used by IDDSM to determine the importance of attributes and features. This information is then used as weighted variables in a linear formula which determines if an identity is deceptive or not. Furthermore, the model then provides an interpretation as to why an identity is deceptive or not.

5.1. The IDDMLM sub-model

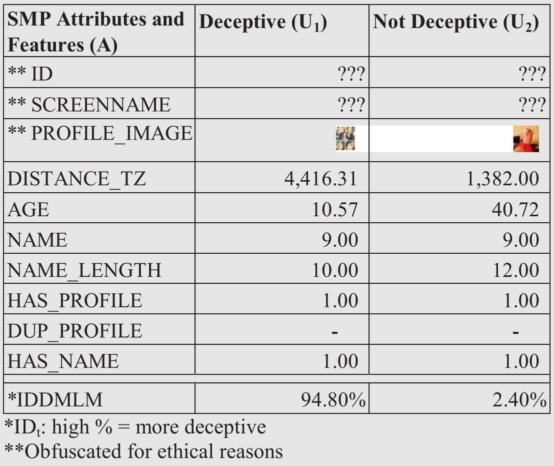
IDDMLM uses the Random Forest algorithm, which is a collection of randomized decision trees (Biau, 2012). IDDMLM is represented as follows:

```
Let SMP={SMP_i: SMP_i is a Social Media Platform}
  Let A={$a_1, a_2,\dots,a_n$} be subset of SMP_i attributes,
  with $n <$ the number of attributes in SMP_i
  Where:
    A is periodically created,
    $A = A^1 \cup A^2$,
    $A^1 = \{a_1, a_2,\dots,a_n\}$ is a random extracted training data set with
     number of deceptive examples = number of not deceptive
     examples,
    $A^2 = \{a_1, a_2,\dots,a_n\}$ is a random extracted test data set with $\Sigma$
     deceptive examples = $\Sigma$ not deceptive examples,
    Note: It is typical for A to be created and thereafter split into
  training and test data where $A^1$ contains 75% of A and the remaining
  25% belongs to $A^2$ (Menardi and Torelli, 2014).
  Let $F = \{f_1, f_2,\dots,f_m\}$ be a set of features,
  $m$ = number of engineered features
  Where:
   $f_i \in A \vee f_i = f(a_j,\dots,a_k)$
    Where:
     $j \geq 1$,
     $k \leq n$.
  Let RF = {$h(x|\Theta_1), h(x|\Theta_2),\dots, h(x|\Theta_t)$} (Breiman, 2001)
  Where:
    RF = Random Forest algorithm,
    $t$ = number of decision trees,
    $h(x|\Theta_i)$ = a single decision tree
    Where:
     $\Theta_i \subseteq ((F \mid A^1) \cup A^1)$,
     $x$ = the values of $A^1 \vee F$ given $\Theta_i$,
     $1 \leq i \leq t$.
```

Note:

For the final classification each decision tree $h(x|\Theta_i)$ casts a vote for the most popular output, given input x. The class with the most votes win. There is no indication of which $h(x|\Theta_i)$'s votes won and also votes differ for each input given. This issue is known as the machine learning interpretability problem (Ribeiro et al., 2016).

Table 8 – Results from the IDDMLM model.



SMP Attributes and Features (A)	Deceptive (U_1)	Not Deceptive (U_2)
** ID	???	???
** SCREENNAME	???	???
** PROFILE_IMAGE		
DISTANCE_TZ	4,416.31	1,382.00
AGE	10.57	40.72
NAME	9.00	9.00
NAME_LENGTH	10.00	12.00
HAS_PROFILE	1.00	1.00
DUP_PROFILE	-	-
HAS_NAME	1.00	1.00
*IDDMLM	94.80%	2.40%

*ID_t: high % = more deceptive
**Obfuscated for ethical reasons

*ID_t: high % = more deceptive
**Obfuscated for ethical reasons

```
Let $RF_{Results} = \{(f1_i, e_i)$: calculated for
    $\forall (a_i \vee f_i) \in \Theta\}$                (Biau, 2012)
    Where:
        $RF_{Results}$ = Results of Random Forest,
        $f1_i$ = an F1 value,           (Jeni et al., 2013)
        $e_i$ = an Entropy value.       (Rényi, 1961)
Let $A^3 = \{a_i \vee f_i$: selected based on optimum values out of the set
    generated by $f(f_i, e_i)\}$                (Breiman, 2001)
Where:
    $1 < i \leq n$,
    $A^3 \subseteq A$.
Let $M_i$ = final Identity Deception Score (IDS) for $U_p$
Where:
    $U_p$ is a user of SMP_i,
    $M_i = RF_p = \{h(x_p|\Theta_1), h(x_p|\Theta_2), \dots, h(x_p|\Theta_t)\}$,
    $x_p$ = values of $\{a_i \vee f_i\} \in A^3$ for $U_p$.
```

5.2. The IDDSM sub-model

The IDDSM sub-model uses the output of the IDDMLM sub-model. IDDSM includes interpretation as to why the identity of a SMP user is perceived as deceptive or not.

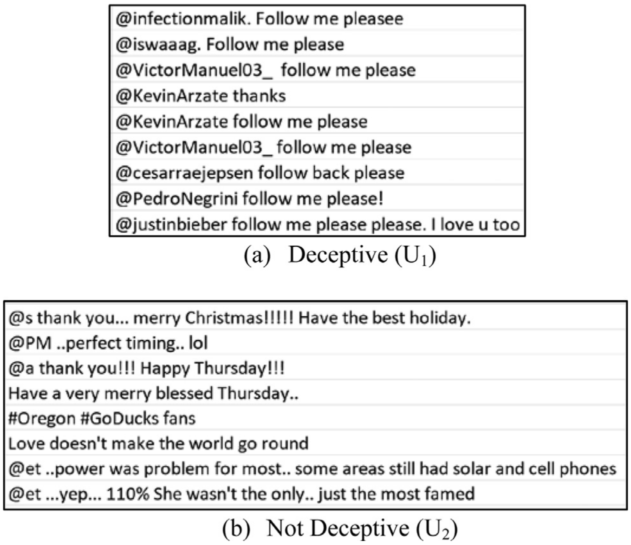
The IDDSM is represented as follows:

```
Let $S_i$ be the Identity Deception Score (IDS) for $U_p$
Then
 $S_i \Rightarrow \sum_{i=1}^{m} f(w, x_p)_i$
Where:
    $m$ = number of elements $\{a_i \vee f_i\}$ in $A^3$,
    $f(w,x_p) = w|x_p|$,
    $w \in [0,100]$,
    $x_p$ = values of $\{a_i \vee f_i\} \in A^3$ for $U_p$,
    $w = e_i \in A^3$
```

If $S_i \sim M_i$ then w, together with x_p can be used to interpret the results of M_i for U_p.

5.3. Illustrating the working of IDDM

Table 8 shows IDDMLM results for two identities, with some features obfuscated due to privacy and ethical reasons. These identities were taken from the original corpus of Twitter data



@infectionmalik. Follow me pleasee
@iswaaag. Follow me please
@VictorManuel03_ follow me please
@KevinArzate thanks
@KevinArzate follow me please
@VictorManuel03_ follow me please
@cesarraejepsen follow back please
@PedroNegrini follow me please!
@justinbieber follow me please please. I love u too

(a) Deceptive (U_1)

@s thank you... merry Christmas!!!!! Have the best holiday.
@PM ..perfect timing.. lol
@a thank you!!! Happy Thursday!!!
Have a very merry blessed Thursday..
#Oregon #GoDucks fans
Love doesn't make the world go round
@et ..power was problem for most.. some areas still had solar and cell phones
@et ...yep... 110% She wasn't the only.. just the most famed

(b) Not Deceptive (U_2)

Fig. 2 – Tweets for individual users (U).

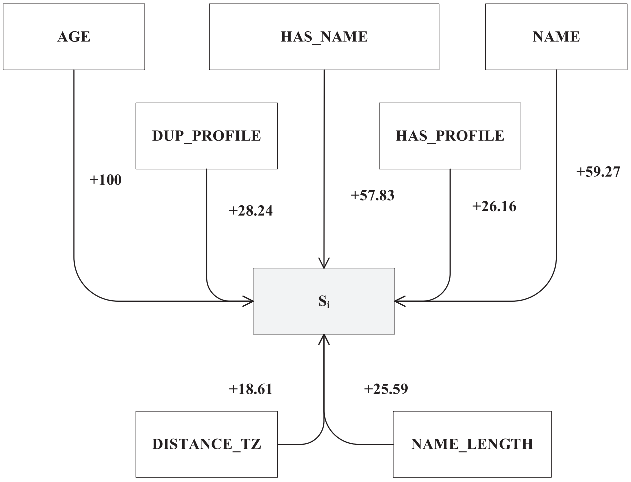


Fig. 3 – The IDDSM sub-model.

used for experiments one and two, as earlier discussed in this paper. One of the identities was determined to be trustworthy and the other one not.

To validate the IDDMLM results, shown in Table 8, a subset of tweets for each individual are presented for clarity. This is shown in Fig. 2.

Given these tweets shown in Fig. 2, it is clear why the first individual could be perceived as trolling the profiles of celebrities and being deceptive, and the second not. Although the conclusion is still perhaps subjective, the IDDMLM model was able to identify potential identity deceptiveness.

The IDDSM proposes to explain the decisions given by IDDMLM. Fig. 3 shows the entropy values determined by RF_{Result} in the form of A^3. The entropy results are indicated by values between 0 and 100 with the latter being most influential. A^3 was subsequently used in the IDDSM towards identity deception detection.

Using the same examples (Table 8) as for the IDDMLM sub-model, Table 9 presents the results according to the IDDSM sub-model where entropy was added to highlight (with color) those features most indicative of the deceptiveness.

Table 9 – IDDSM results.

SMP Attributes and Features (A^3)	Deceptive (U_1)	Not Deceptive (U_2)
AGE	🟥 40	🟥 10
DISTANCE_TZ	🟥 4,416	🟥 1,382
NAME	9	9
NAME_LENGTH	🟥 12	🟥 10
HAS_PROFILE	1	1
DUP_PROFILE	-	-
HAS_NAME	1	1

The IDDSM results therefore adds an interpretation feature to our IDDM model.

5.4. *Comparing IDDM with other models*

The IDDM model differs from other proposed machine learning interpretation models in that the results require no further machine learning or repetitive processing. The LIME model (Ribeiro et al., 2016), for example, generates additional local values for each user, similar but not equal to the original. All values for a particular user, are then trained, ignoring the values of other users, with a linear supervised machine learning algorithm. If the original cost of one iteration of a supervised machine learning model is represented as O(1), then the cost of the LIME model would be O(1+*n*), for n users as a separate machine learning model is trained additionally for each user. The result is a local linear explanation specific to each user. Baehrens et al. (2010) followed a similar local approximation approach at the same cost as LIME, but by generating new local values using a different method. Besides linear interpretation models, others propose to use game theory (Beillevaire, 2016). With game theory, the Shapely value (Shapley, 1953) shows promise by calculating all potential outcomes using different combinations of inputs. This, however, becomes computationally expensive when many inputs are used. The cost can be represented as O (k!) with *k* the number of inputs. There are other machine learning interpretation models that are dependent on the machine learning model used. An example for such a model is the 'treeinterpreter' (Saabas, 2018) which uses the knowledge gathered from all trees in a random forest to interpret the final result for a user. In this scenario the cost can be represented as O(t) where t is the number of random forest trees generated by the model during training.

The IDDM presented in this paper, on the other hand, proposes to only use the results from the original trained supervised machine learning model. No further computations are required. The cost of the IDDM thus remains O (1) which is imperative for SMPs dealing with large volumes of data.

6. Conclusion and future work

Cyber security in general can benefit from the research work presented in this paper which deals with the development of intelligent identity deception detection by means of machine learning models. An Identity Deception Detection Model

(IDDM) was proposed to not only detect, but also interpret perceived deceptiveness. The model consists of two sub-models.

The first sub-model, IDDMLM, used input from prior experiments to present a machine learning model that detects identity deception on SMPs with an F1 score of 86.24%. This result was achieved using engineered features identified for the detection of bots, as well as new features based on insights from the field of Psychology. The entropy values extracted from the results furthermore shows the contribution of each feature which was then applied by the second sub-model. The IDDSM interpreted the results from IDDMLM by means of a simple weighted linear formula, given the known entropy of the features involved. The IDDSM results highlighted which features a specific user was found to be most likely deceptive about. This is invaluable in use cases where that particular user should be investigated further. It was also shown how the cost of IDDSM remains O(1). This low-cost interpretative model is valuable in scenarios where near real-time results are required in big data environments.

Given what was learned from this research, the following recommendations can be made to address the vulnerabilities of SMP platforms so as to improve the environment for the intelligent detection of deceptive identities:

- The engineered features as identified in this paper (such as name_length) can be added as default attributes in the creation of user accounts on SMPs. This modification will save on computational workload when intelligent models, such as IDDM, are employed to detect human identity deception.
- The sub-set of existing attributes that are compulsory for users to complete during the creation of an account (e.g. the user's location) should be expanded. This information need not be public knowledge, so that the person's privacy is protected.
- Validation mechanisms can be added during account creation or updates to ensure the authenticity of the information that is provided. Profile images can, for example, be validated to ensure they contain the image of a person.

Future research work proposes to identify more features valuable towards the detection of identity deception detection on SMPs including the refinement of the IDDM.

Acknowledgments

We would like to thank the HPI Future SOC research lab in Potsdam, Germany for its assistance and support during the research. The lab provided the research environment to not only mine and store the Twitter data, but also to execute the experiments.

Appendix A

Table A.1 – Social media attributes as identified in for the top 6 SMPs of 2016.

Attribute group	General attribute description	Social media attributes					
		(Facebook, 2017)	(Google+, 2017)	(LinkedIn, 2017)	(Twitter, 2017)	(Pinterest, 2017)	(Instagram, 2017)
Profile information	The ID of the person's user account	ID	ID	ID	ID	ID	ID
	Their first name	first_name	name.givenName	first-name		first_name	
	Their last name	last_name	name.familyName	last-name		last_name	
	Middle name	middle_name	name.middleName				
	Full name		name.formatted		name		full_name
	Name to display	Name	displayName	formatted-name	screen_name	username	Username
	Age	age_range	ageRange				
	Birth date	Birthday	Birthday	date-of-birth			
	Profile picture	Cover	Image	picture-url	profile_image	image	profile_picture
	Background picture				background_image		
	Email/Phone	Email/phone	emails[]	email-address	email/phone	email	email/phone
	Gender	Gender	Gender				
	Relationship status	Relationship_status	relationship Status				
	Language	Languages	Language	Languages	lang		
	Location	Location	placesLived[].primary	Location	location		
	Geolocation				geo-enabled, latitude, longitude		
	Timezone	Timezone			time_zone		
	UTC offset				utc_offset		
	Bio field	about	about Me	Summary	description	bio	Bio
Account information	Authenticity of account	is_verified	Verified		verified		
	Updated time	updated_time			created_at		
					protected		
Behaviour	List of devices	Devices					
	Likes	Likes			statuses_count		media.likes
Relationships	Friends	Friends		num-connections	followers_count		
	Groups	Groups		Following	friends_count		
	Listed				listed_count		
	Family	Family		Job-bookmarks			
Content	Recommendations			num-recommenders	user_mentions	Counts	Counts
	Content specific fields	Albums	Curls[]	Position	tweets	Boards	Media
		Feeds	Organizations[]	Skills	tweets.created_at	Pins	
		Events	Bragging Rights	Certifications			media.created_time
		Photos	Occupation	Educations			media.location
		Videos	Skills	Courses			
				Volunteer			
				Publications			
				Interests			
				honors-awards			

Note: Sources are shown in square brackets under the relevant SMP.

REFERENCES

Al-Garadi MA, Varathan KD, Ravana SD. Cybercrime detection in online communications: the experimental case of cyberbullying detection in the Twitter network. Comput Hum Behav 2016;63:433–43.

Alowibdi JS, Buy UA, Philip SY, Ghani S, Mokbel M. Deception detection in Twitter. Soc Netw Anal Min 2015;5:1–16.

AMAZON. Mechanical Turk. Available, https://www.mturk.com/; 2017 [Accessed 08 01 18].

Anguita D, Ghelardoni L, Ghio A, Oneto L, Ridella S. The 'K'in K-fold cross validation. Proceedings of the European symposium on artificial neural networks, computational intelligence and machine learning; 2012. p. 441–6.

Anonymous. Ceding powers of decision to AI presents a paradox. Financial Times 2017. [Online]. Available: https://www.ft.com/content/63542534-ebf6-11e7-bd17-521324c81e23. [Accessed 4 01 18].

Armstrong K, Hunt A. Random User Generator. Available:, https://randomuser.me/; 2017 [Accessed 08 01 18].

Assunção MD, Calheiros RN, Bianchi S, Netto MA, Buyya R. Big data computing and clouds: trends and future directions. J Parallel Distrib Comput 2015;79:3–15.

Baehrens D, Schroeter T, Harmeling S, Kawanabe M, Hansen K, Mãžller K-R. How to explain individual classification decisions. J Mach Learn Res 2010;11:1803–31.

Beillevaire, M. 2017. *Inside the Black Box: How to Explain Individual Predictions of a Machine Learning Model*. Computer Science and Engineering Masters, KTH Royal Institute of Technology.

Bellinger C, Sharma S, Japkowicz N. One-class versus binary classification: which and when?. Proceedings of the 2012 11th International Conference on Machine learning and applications (ICMLA). IEEE; 2012. p. 102–6.

Benevenuto F, Magno G, Rodrigues T, Almeida V. Detecting spammers on twitter. Proceedings of the collaboration, electronic messaging, anti-abuse and spam conference (CEAS); 2010. p. 12.

Biau G. Analysis of a random forests model. J Mach Learn Res 2012;13:1063–95.

Bliss L. The law, social media and the victimisation of women. Proceedings of the *Socio-Legal Studies Association Annual Conference*. Newcastle University, 2017.

Bogdanova D, Rosso P, Solorio T. Exploring high-level features for detecting cyberpedophilia. Comput Speech Lang 2014;28:108–20.

Breiman L. Random forests. Mach Learn 2001;45:5–32.

Burrell J. How the machine 'thinks': understanding opacity in machine learning algorithms. Big Data Soc 2016;3.

Camber R. Oil millionaire's son, 14, 'found murdered in flat 30 miles from home after being groomed through computer games': Man, 18, charged after stabbing. Available:, http://www.dailymail.co.uk/news/article-2562890/BREAKING-NEWS-Son-14-oil-millionaire-murdered-man-groomed-computer-games.html; 2014 [Accessed 08 01 18].

Caspi A, Gorsky P. Online deception: prevalence, motivation, and emotion. Cyber Psychol Behav 2006;9:54–9.

Chaffey, D. 2018. *Global social media research summary* [Online]. Smart Insights. Available: https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/ [Accessed 23 Jun 2018].

Choudhary N, Jain AK. Advanced informatics for computing research. Towards filtering of SMS spam messages using machine learning based technique. Singapore: Springer; 2017.

Chu Z, Gianvecchio S, Wang H, Jajodia S. Who is tweeting on Twitter: human, bot, or cyborg?. Proceedings of the 26th annual computer security applications conference. ACM; 2010. p. 21–30.

Cook DM. Birds of a feather deceive together: the chicanery of multiplied metadata. J Inf Warfare 2014;13:85–96.

Cresci S, Di Pietro R, Petrocchi M, Spognardi A, Tesconi M. Fame for sale: efficient detection of fake Twitter followers. Decis Supp Syst 2015;80:56–71.

Dal Pozzolo A, Caelen O, Waterschoot S, Bontempi G. Racing for unbalanced methods selection. Proceedings of the international conference on intelligent data engineering and automated learning. Springer; 2013. p. 24–31.

De Villiers J. Suspects use fake Facebook profile to lure women, rape and kill them. *News24* [Online]. Available, https://www.news24.com/SouthAfrica/News/suspects-use-fake-facebook-profile-to-lure-women-rape-and-kill-them-20171104; 2017 [Accessed 04 11 17].

Depaulo BM, Kashy DA, Kirkendol SE, Wyer MM, Epstein JA. Lying in everyday life. J Pers Soc Psychol 1996;70:979.

Dickerson JP, Kagan V, Subrahmanian V. Using sentiment to detect bots on Twitter: are humans more opinionated than bots?. Proceedings of the 2014 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM). IEEE; 2014. p. 620–7.

Digital T. Mbete doesn't recognise these tweets - speaker spoofed on Twitter. Times Live 2016. [Online]. Available: http://www.timeslive.co.za/politics/2016/04/13/Mbete-doesnt-recognise-these-tweets-speaker-spoofed-on-Twitter. [Accessed 13 04 16].

Drouin M, Miller D, Wehle SM, HERNANDEZ E. Why do people lie online? "Because everyone lies on the internet". Comput Hum Behav 2016;64:134–42.

Ebrahimi M, Suen CY, Ormandjieva O, KRZYZAK A. Recognizing predatory chat documents using semi-supervised anomaly detection. Electron Imag 2016;2016:1–9.

FACEBOOK. The Facebook Graph API. Available:, https://developers.facebook.com/docs/graph-api/overview; 2017 [Accessed 08 01 18].

Ferrara E, Wang W-Q, Varol O, Flammini A, Galstyan A. Predicting online extremism, content adopters, and interaction reciprocity. Proceedings of the international conference on social informatics. Springer; 2016. p. 22–39.

Fire M, Kagan D, Elyashar A, Elovici Y. Friend or foe? Fake profile identification in online social networks. Soc Netw Anal Min 2014;4:1–23.

Galán-García P, De La Puerta JG, Gómez CL, Santos I, Bringas PG. Supervised machine learning for the detection of troll profiles in twittersocial network: Application to a real case of cyberbullying. Log J IGPL 2016;24:42–53.

Genuer R, Poggi J-M, Tuleau-Malot C. Variable selection using random forests. Pattern Recognit Lett 2010;31:2225–36.

Goodman B, Flaxman S. European Union regulations on algorithmic decision-making and a "right to explanation". ICML workshop on human interpretability in machine learning; 2016 NewYork, USA.

GOOGLE. Google Vision API. Available:, https://cloud.google.com/vision/; 2017 [Accessed 08 01 18].

GOOGLE+. Google+ API. Available:, https://developers.google.com/+/web/api/rest/; 2017 [Accessed 08 01 18].

Gu G, Perdisci R, Zhang J, Lee WB. Clustering analysis of network traffic for protocol-and structure-independent botnet detection. Proceedings of the USENIX security symposium; 2008. p. 139–54.

Gurajala S, White JS, Hudson B, MATTHEWS JN. Fake Twitter accounts: profile characteristics obtained using an activity-based pattern detection approach. Proceedings of the 2015 international conference on social media & society. ACM; 2015. p. 9.

Gurajala S, White JS, Hudson B, Voter BR, Matthews JN. Profile characteristics of fake Twitter accounts. Big Data Soc 2016;3.

Haimson OL, Hoffmann AL. Constructing and enforcing "authentic" identity online: Facebook, real names, and non-normative identities. First Monday 2016;21 [Online].

Halevy R, Shalvi S, Verschuere B. Being honest about dishonesty: correlating self-reports and actual lying. Hum Commun Res 2014;40:54–72.

Hancock JT. Digital Deception. Oxford University Press; 2007.

Hancock JT, Toma CL. Putting your best face forward: the accuracy of online dating photographs. J Commun 2009;59:367–86.

INSTAGRAM. Instagram API. Available:, https://www.instagram.com/developer/; 2017 [Accessed 08 01 18].

Jeni LA, Cohn JF, De La Torre F. Facing imbalanced data–Recommendations for the use of performance metrics. Proceedings of the 2013 Humaine association conference on affective computing and intelligent interaction (ACII). IEEE; 2013. p. 245–51.

Jupe LM, Vrij A, Nahari G, Leal S, Mann SA. The lies we live: using the verifiability approach to detect lying about occupation. J Artic Support Null Hypothesis 2016;13:1–13.

Kabay M, Salveggio E, Guess R, Rosco RD. Anonymity and identity in cyberspace. Comput Secur Handb Sixth Ed 2014;70:1–70 37.

Keen B. Generate Data. Available:, http://www.generatedata.com/; 2017 [Accessed 08 01 18].

Khandpur RP, Ji T, Jan S, Wang G, Lu C-T, Ramakrishnan N. Crowdsourcing Cybersecurity: Cyber Attack Detection using Social Media. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. Singapore; 2017. p. 1049–57.

Kierkegaard S. Cybering, online grooming and ageplay. Comput Law Secur Rev 2008;24:41–55.

Klausen J. Tweeting the jihad: social media networks of Western foreign fighters in Syria and Iraq. Stud Conflict Terrorism 2015;38:1–22.

Kothari CR. Research methodology: methods and techniques. New Age International; 2004.

Kuhn M, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, et al. Caret: classification and regression training. R Packag Version 2016;6:0–73.

Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. Soviet Phys Doklady 1966:707–10.

LI J, Wang AG. A framework of identity resolution: evaluating identity attributes and matching algorithms. Secur Inform 2015;4:1.

LINKEDIN. LinkedIn Developers. Available:, https://developer.linkedin.com/; 2017 [Accessed 08 01 18].

Lipton ZC. The mythos of model interpretability. ICML Workshop on Human Interpretability in Machine Learning; 2016 New York.

Ma C, Zhang HH, Wang X. Machine learning for Big Data analytics in plants. Trends Plant Sci 2014;19:798–808.

Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat 1947:50–60.

Mcdonald JH. Handbook of Biological Statistics. MD: Sparky House Publishing Baltimore; 2009.

Menardi G, Torelli N. Training and assessing classification rules with imbalanced data. Data Min Knowl Discov 2014;28:1–31.

Oentaryo RJ, Murdopo A, Prasetyo PK, Lim E-P. On profiling bots in social media. Proceedings of the international conference on social informatics. Springer; 2016. p. 92–109.

Peddinti ST, Ross KW, Cappos J. Mining Anonymity: Identifying Sensitive Accounts on Twitter. International AAAI Conference on Web and Social Media; 2017 Montreal, Canada.

Peterson, T. 2016. *Rapist who used social media to lure childvictims sentenced to 20 years* [Online]. Available: http://www.news24.com/SouthAfrica/News/rapist-who-used-social-media-to-lure-child-victims-sentenced-to-20-years-20160615 [Accessed].

PINTEREST. Pinterest API. Available:, https://developers.pinterest.com/; 2017 [Accessed 08 01 18].

Rényi A. On measures of entropy and information. Proceedings of the fourth Berkeley symposium on mathematical statistics and probability; 1961. p. 547–61.

Ribeiro MT, Singh S, Guestrin C. Why should i trust you?: explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM; 2016. p. 1135–44.

Rong R, Houser D, Dai AY. Money or friends: social identity and deception in networks. Eur Econ Rev 2016;90:56–66.

Rubin VL. Deception Detection and Rumor Debunking for Social Media. In: Sloan L, Quan-Haase A, editors. The SAGE Handbook of Social Media Research Methods. Sage, UK; 2017.

Saabas, A. 2018. *Package for interpreting scikit-learn's decision tree and random forest predictions.* [Online]. Available: https://pypi.org/project/treeinterpreter/ [Accessed 23 Jun 2018].

Schwartz HA, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones SM, Agrawal M, et al. Personality, gender, and age in the language of social media: the open-vocabulary approach. PloS One 2013;8:e73791.

Sedhai S, Sun A. Semi-Supervised Spam Detection in Twitter Stream. IEEE Transactions on Computational Social Systems 2017;5:169–75.

Shannon CE. A mathematical theory of communication. ACM SIGMOBILE Mob Comput Commun Rev 2001;5:3–55.

Shapley LS. A value for *n*-person games. Contrib Theory Games 1953;2:307–17.

Shumaker B, Sinnott R. Astronomical computing: 1. Computing under the open sky. 2. Virtues of the haversine. Sky Telesc 1984;68:158–9.

Smit D. Cyberbullying in South African and American schools: a legal comparative study. South Afr J Educ 2015;35:01–11.

Stanton K, Ellickson-Larew S, Watson D. Development and validation of a measure of online deception and intimacy. Personal Individ Differ 2016;88:187–96.

Thomas K, Grier C, Song D, Paxson V. Suspended accounts in retrospect: an analysis of twitter spam. Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference. ACM; 2011. p. 243–58.

Toma CL, Hancock JT, Ellison NB. Separating fact from fiction: an examination of deceptive self-presentation in online dating profiles. Personal Soc Psychol Bull 2008;34:1023–36.

Tsikerdekis M. Identity deception prevention using common contribution network data. IEEE Trans Inf Forensics Secur 2017;12:188–99.

Tsikerdekis M, Zeadally S. Multiple account identity deception detection in social media using nonverbal behavior. IEEE Trans Inf Forensics Secur 2014;9:1311–21.

Tuna T, Akbas E, Aksoy A, Canbaz MA, Karabiyik U, Gonen B, et al. User characterization for online social networks. Soc Netw Anal Min 2016;6:104.

Tuteja SK. A survey on classification algorithms for email spam filtering. Int J Eng Sci 2016:5937.

TWITTER. Twitter API. Available:, https://dev.twitter.com/overview/api; 2017 [Accessed 08 01 18].

Utz S. Types of deception and underlying motivation: What people think. Soc Sci Comput Rev 2005;23:49–56.

Van Der Walt E, Eloff JHP. Using machine learning to detect fake identities - Bots versus Humans. IEEE Access 2018;6:6540–9.

Van Liere D. How far does a tweet travel?: Information brokers in the twitterverse. Proceedings of the international workshop on modeling social media. ACM; 2010. p. 6.

Venkatesan S, Albanese M, Shah A, Ganesan R, Jajodia S. Detecting Stealthy Botnets in a Resource-Constrained Environment using Reinforcement Learning. Workshop on Moving Target Defense. Dallas, Texas: ACM; 2017. p. 75–85.

Wang GA, Chen H, Xu JJ, Atabakhsh H. Automatically detecting criminal identity deception: an adaptive detection algorithm. *IEEE Trans Syst Man Cybern Part A: Syst Hum* 2006;36:988–99.

Wolpert DH, Macready WG. No free lunch theorems for optimization. IEEE Trans Evol Comput 1997;1:67–82.

Yarkoni T, Westfall J. Choosing prediction over explanation in psychology: Lessons from machine learning. Perspectives on Psychological Science: J Assoc Psychol Sci 2016;12:1100–22.

Zafarani R, Liu H. Evaluation without ground truth in social media research. Commun ACM 2015;58:54–60.

Estée van der Walt was born in Johannesburg, South Africa in 1977. She received B.Sc. and M.Sc. degrees in Computer Science from the University of Johannesburg, South Africa, in 1997 and 2001, respectively. She is currently completing her Ph.D. degree in Information Technology at the University of Pretoria, South Africa. Her research interests include cyber-security and identity deception on social media platforms. She is particularly interested in the protection of humans on these big data platforms. She makes use of machine learning and data mining techniques to not only understand this type of deception but also find a solution to detect and warn authorities about identity deception. She was the recipient of the “Best Poster Award" at the 3rd International Conference on Information Systems Security and Privacy that was hosted in Porto, Portugal during February 2017.

Jan Eloff graduated in 1985 with a Ph.D. in Computer Science. Up to June 2015 he was appointed as the Research Director for SAP Research in Africa and is currently appointed as Deputy Dean Research & Postgraduate studies: Faculty of Engineering, Built Environment and IT (EBIT), and as a full professor in computer science at the University of Pretoria, South Africa. From 2007 he is an associate-editor of the Computers & Security journal and an editorial member for the International Computer Fraud & Security bulletin published by Elsevier. He is an internationally recognised researcher and has published 113 peer reviewed papers with 3537 citations.

Jacomine Grobler is a senior lecturer in the Department of Industrial and Systems Engineering at the University of Pretoria in South Africa. Her main fields of expertise are multi-method optimization algorithms, swarm intelligence, multi-objective optimization, supply chain optimization and big data science. She completed her Ph.D. in 2014 and was recently awarded the 2015 JD Roberts emerging researcher award for her contribution to the development of mathematical models and optimization algorithms. She also received the 2017 South African Institute for Industrial Engineering Most Outstanding Young Industrial Engineering Researcher Award. She regularly reviews papers for leading international journals and have presented various invited lectures, for example, a tutorial at the Seventh International Conference on Swarm Intelligence in Bali, Indonesia.