IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 5, 2017, accepted January 12, 2018, date of publication January 23, 2018, date of current version March 9, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2796018

Using Machine Learning to Detect Fake Identities: Bots vs Humans

ESTÉE VAN DER WALT AND JAN ELOFF
Department of Computer Science, University of Pretoria, Pretoria 0002, South Africa
Corresponding author: Estée Van Der Walt (estee.vanderwalt@gmail.com)

ABSTRACT There are a growing number of people who hold accounts on social media platforms (SMPs) but hide their identity for malicious purposes. Unfortunately, very little research has been done to date to detect fake identities created by humans, especially so on SMPs. In contrast, many examples exist of cases where fake accounts created by bots or computers have been detected successfully using machine learning models. In the case of bots these machine learning models were dependent on employing engineered features, such as the ‘‘friend-to-followers ratio.’’ These features were engineered from attributes, such as ‘‘friend-count’’ and ‘‘follower-count,’’ which are directly available in the account profiles on SMPs. The research discussed in this paper applies these same engineered features to a set of fake human accounts in the hope of advancing the successful detection of fake identities created by humans on SMPs.

INDEX TERMS Big data, bots, data science, fake accounts, fake identities, identity deception, social media, veracity.

I. INTRODUCTION

Identity deception on big data platforms (like social media) is an increasing problem, due to the continued growth and exponential evolvement of these platforms. Social media is one of the preferred means of communication [1] and has become a target for spammers and scammers alike [2]. Cyberthreats like spamming, which involves the sending of unsolicited emails, are common in email applications. These same threats - and more - now emerge on social media platforms (SMPs), although in different manifestations.

Much can be learned about people’s behaviour and needs through analysing their interactions with one another. Habits and topics of conversations can be evaluated to deliver a better service or product to customers and ultimately to people at large [1], [3]. The same information can however also be used against people, very often in a deceptive way. For example, a cluster of people may influence an opinion [4] when the other participants in the conversation are unaware that the ‘‘people’’ in the cluster are not real.

Since the detection of fake social engagement is quite challenging [5], this vulnerability is greatly abused [6]. We believe that these fake accounts can be attributed to, among others, the following factors:

- *The privacy policies of SMPs not expecting persons to reveal their true identity* [7]. The authenticity of people is constantly being questioned [1], and this can detrimentally affect [2] those who are falsely accused or misled. An example is the case of cyberbullying [8] where children are bullied online through the spreading of false rumours.
- *Malicious individuals and groups on SMPs striving to spread chaos and pandemonium.* A recent example was the spreading of fake news about Hurricane Sandy in the US [9]. False news about the hurricane went viral and became a main source of information for those affected by the storm.
- *The gamification of sites, with more ‘‘likes’’ or ‘‘followers’’ inadvertently meaning greater popularity and higher social ratings* [2]. This trend drives people to find new means to artificially or manually [2] stay ahead of their competitors. By analogy, the most popular candidate in a political election usually receives most of the votes [10].
- *The ease with which false accounts and actions can be obtained.* An example is false accounts being bought online at a marketplace [11] at minimal cost, or delivered through crowdsourcing services [12]. It is even possible to buy Twitter followers and Facebook ‘‘likes’’ online [5].

Fake accounts can be either human-generated, computer-generated (also referred to as ‘‘bots’’), or cyborgs [13]. A cyborg is a half-human, half-bot account [13]. Such an account is manually opened by a human, but from then onwards the actions are automated by a bot. Variations exist between bots and human accounts. For example, bots are known as ''Sybil'' accounts when the accounts are fake [1], [12] and not stolen from legitimate users [14]. On the other hand, fake human accounts are known as ''trolls'' when their purpose is to defame the character of another person [8]. Regardless of the origin of the account, the malicious intent of these fake identities is as follows:

- To change the actions of an individual or group - examples are online extremism; terrorist propaganda; and radicalisation campaigns [15].
- To change perceptions of an individual or group - examples are changing the creditworthiness of an account [16]; spreading rumours and false news [9], [17]; defaming someone's character [8], [17]; polarising opinions [4], [8]; influencing popularity [11], [18]; and skewing perceptions [10].
- To hide the malicious activity of an individual or group - examples are identity impersonation [8]; identity theft [11]; cyberbullying [8]; dissemination of pornography [17]; and fraud [11].
- To spread malware - examples are the creation of false communications to steal credentials [19]; or misdirecting users to fake web sites [20].

Past research has done much to detect fake identities generated by bots. Machine learning [16], [21] has been used to not only detect bots on SMPs but also identify the intent of the bot [21]. Fake identities can possibly be detected by various approaches. These can include, amongst others, the detection of fake content linked to the account [10], investigating the account profile itself [1], or using non-verbal indicators, for example the time between opening an account and the first entry posted [22].

The problem is that very little has been done so far to detect actual human identities that are fake. The field of psychology has provided suggestions as to what constitutes a fake identity [23]–[25]. Humans are just as responsible for the malicious intents found on SMPs and they therefore warrant the same attention. The difference, according to the authors, is that fake bot accounts target groups at large, whereas fake human accounts rather tend to target specific individuals. This could lead to severe consequences for the targeted individual.

In a previous study [26] we investigated the use of social media attributes found in Twitter, with the aim of detecting instances of identity deception by humans on SMPs. We found that standard attributes alone, such as the number of friend and followers that are available through application programming interfaces (APIs) and describing accounts in SMPs [27]–[29] like Twitter, were not sufficient to successfully detect fake identities created by humans.

In this paper we evaluate whether readily available and engineered features that are used for the successful detection, using machine learning models, of fake identities created by bots or computers can be used to detect fake identities created by humans. This is done in the hope that similar features can serve as a catalyst for uncovering identity deception by humans on SMPs.

II. RELATED WORK

Seeing that very little has been done so far to detect actual fake human identities on SMPs, we looked towards past research addressing similar problems. Spam behaviour found in emails and SMS, for example, shows similar malicious intent with fake accounts spreading false rumours [30]. Spamming occurs when electronic media such as emails, SMSs and SMPs are used to send unsolicited content to an individual or group [31]. Besides spam, fake identities are also present on SMPs in the form of bots.

Previous research towards understanding and identifying spam behaviour presented techniques like filtering [32], rules [30], and machine learning [16] to detect fake identities. The same techniques, and more, have been applied to SMPs to detect fake bot accounts:

- *Filtering* is mostly reactive: only when a new threat is identified and verified will that sender be added to a blacklist. Similar methods of dealing with spam have been proposed on Twitter to blacklist known malicious URL content and to quarantine known bots [6]. Spam filtering, however, becomes very difficult when spammers use dynamically adaptive and automated strategies to circumvent the proposed methods. This is even more true for SMPs. Humans easily adapt themselves to avoid detection and, in the case of blacklisting, they simply create a new account and fake identity [16] as soon as the current detected account is blacklisted.
- Besides filtering techniques, *rules* have been established to identify fake accounts during detection. Examples of such rules are based on words (such as 'win') that are known to belong to spam within the messages [30], [33]. If a message contains such a word or number of words, it is regarded as spam. These same rules have been applied to SMPs with success [17]. The problem, however, is that new words are created constantly, and abbreviated words are common on SMPs, such as 'lol' meaning 'laugh out loud'. This is problematic in the sense that detection rules are usually outdated. More adaptive rules were proposed on SMPs by means of pattern matching [1]. For example, if an account has been tweeting about three or more trending topics, or if an account took part in trending topics but is less than a day old, it can be classified as fake [18]. On Facebook, Fire *et al.* [34] scored friends for deceptiveness by using rules based on similar relationships, tagging, and chat history with others. These rules have success in detecting bot accounts but fail to detect actual fake human accounts. Human behaviour is deemed to be more random [35] than that of bot accounts [14] and thus hard to represent by means of rules.
- *Supervised machine learning* models have been proposed to detect fake accounts. For email spam detection [36], supervised classification machine models like support vector machines (SVMs), decisions trees, Naïve Bayes and neural networks were proposed by Tuteja [36]. Features were engineered, as input for the

models, based on the header and content of the body of the email [36]. For SMS spam detection [37], ten features, inter alia SMS length, were engineered by Choudhary and Jain [37]. These features predicted SMS spam with great success by using supervised machine learning models like random forest, decision trees, J48, logistic regression, and Naïve Bayes. Cresci *et al.* [16] proposed a supervised machine learning model based on those attributes describing the identity of an account only, to detect bots on SMPs. Gupta *et al.* [9] in turn suggested that behaviour, such as the frequency of messages and time of day, provides enough information to detect bots successfully through supervised machine learning models. Supervised machine learning models require a label included in the corpus to predict the expected outcome [38].

- Various *semi-supervised machine learning* models have been proposed. Amongst others, Ebrahimi *et al.* [39] compared a one-class support vector machine model to a Naïve Bayes machine learning model and showed how the one-class SVM outperforms the binary classification model when one of the classes is the minority. The norm is to train a one-class SVM on the minority class [39], [40]. In SMPs it is not practical to mine the minority class consisting of fake accounts [39] or be certain that an account is indeed deceptive [41]. Semi-supervised machine learning models require a clear boundary between classes [42].
- *Unsupervised machine learning* was successfully applied by Gu *el al.* [43], Wu *et al.* [44], and Yahyazadeh and Abadi [45]. Their research showed how clustering, which is a common unsupervised machine learning method, can be used to detect bots. With unsupervised machine learning, the data is unlabelled, and data are grouped based on similarity [38]. Clustering works well to detects bots as these bots usually share similar characteristics and has the same purpose. Not the same can necessarily be said of fake human accounts.
- Venkatesan *et al.* [46] presented a *reinforcement* proof-of-concept model that rewards itself for detecting bots successfully. Spam in SMPs was detecting by Arif *et al.* [47] whereby the importance of features was used to build a better performing set of rules iteratively. Reinforcement machine learning models require feedback from the environment to adjust and improve. This is not readily available in SMPs.

Given these techniques proposed by previous work, the research at hand will focus on supervised machine learning. The reasons being that supervised machine learning are well suited for classification problems [38], is preferred above unsupervised machine learning techniques for bot detection [4], and have shown good results in past research work detecting bots [16]. We also believe that human deceptive accounts are not as common as bots and therefore less likely to be clustered appropriately through unsupervised machine learning approaches.

TABLE 1. Twitter attributes used in a previous study [26].

Attribute	Description
NAME	The name of the account holder
SCREENNAME	The pseudonym for the account
CREATED	The date the account was created
FOLLOWERS_COUNT	The number of followers for the account
FRIENDS_COUNT	The number of friends for the account
LANGUAGE	The language of the account holder
LISTED_COUNT	The number of groups the account belongs to
PROFILE_IMAGE	The profile image of the account
STATUS_COUNT	The number of tweets made by the account
LOCATION	The location of the account holder
TIMEZONE	The time zone of the account holder
UTC_OFFSET	The UTC offset, given the TIMEZONE
LATITIDE	The latitude where the last tweet was made
LONGITUDE	The longitude where the last tweet was made

Supervised machine learning algorithms require a dataset of features with a label classifying each row or outcome. Features are thus the input used by supervised machine learning models to predict an outcome. These features can be the attributes found via APIs that describes a single piece of information about an SMP account, like the number of friends. Features can also be engineered by combining attributes from an SMP account, past engineered features, and/or domain knowledge. An example of an engineered feature is the combination of the number of friends and followers to present their relationship as a ratio for input to a machine learning model.

Features used by machine learning models are mostly referred to as "engineered features" as they are a combination of attributes and engineered features. There are however exceptions. In the previous study [26] by the authors, only the attributes in Twitter were used [27]. These attributes are shown in Table 1. It was possible with these attributes alone to identify fake accounts generated by humans, but the result was worse than getting the prediction right by chance.

With SMPs, the engineered features are divided into three distinct groups: data describing the *identity* of the account, the *relationships* of the account to others, and lastly the *behaviour* or messages of the account. This is different from email- or SMS-engineered features in that they only consider features pertaining to the header and body of the message. To detect fake bot accounts in SMPs, various combinations of the above three groups of engineered features were applied to the machine learning models. Cresci *et al.* [16] proposed a lightweight classification model based on the identity of the account (thus, excluding its relationships and behaviour). They suggested that features about the identity of an account are sufficient to detect bots. Gupta *et al.* [9] in turn suggested that behaviour, such as the frequency of messages and time of day, provides more information relevant to deception than the identity of the account itself. Detecting the behaviour through sentiment was also successful for specific topics of interest, for example elections [10]. Given the positive results presented by Cresci *et al.* [16] we propose to also use a similar light weight classifier that only includes data describing the identity of an account. When humans are being deceptive the intent could have detrimental consequences to the targeted individual. The sooner the deception can be detected, the better.

Various options were evaluated to obtain a dataset of potentially deceptive humans, given past research, for SMPs. Some researchers used data from *available datasets*, like paedophiles [48] and extremism groups [15], to label accordingly. To the best of our knowledge, no labelled dataset exists of humans lying specifically about their identity. Other researchers employed the help from the *Amazon Mechanical Turk* crowdsourcing platform where people from the public are paid to label data manually [49], [50]. Due to the volumes of data in SMPs this was not an option for the research at hand. Twitter also indicate *suspended accounts* [15]. This information can be used to label an account. Unfortunately, Twitter does not give the reason for the account being suspended and will thus include more than accounts only suspended for being untruthful about their identity. Furthermore, Zhu [51] listed various other *semi-supervised machine learning* techniques where one approach is to cluster the data first and use the output to label the data. This approach does require the dataset to be clusterable in accord with unsupervised machine learning. Lastly, deceptive accounts can be *manually injected* into the existing corpus gathered. As none of the previous options were viable for the research at hand and collecting deceptive accounts is not practical in real-world examples to date [39], this was the option taken. The accounts were however generated in an informed way.

Past research in psychology has done much work on understanding why people lie [52]–[54]. We looked towards this research showing that people lie about their age [24], [55], gender [25], [55], image [49], location [23], [56], and their name [49], [57]. Using input from the field of psychology, allows for the creation of a set of 'informed'-deceptive accounts.

Although all previous studies focused on the detection of bots and spam, a few more recent ones have subsequently addressed the detection of deceptive human accounts. Bogdanova *et al.* [48] proposed that the behaviour found in the messages of paedophiles should be used to protect minors. The approach of these researchers relied on sentiment and text analytics to predict deception with an SVM machine learning model. Cyberbullying was addressed by Galán-García *et al.* [8]. They relied on the fact that cyberbullies have a distinctive relationship with the user they target. This knowledge can be used to great effect to identify cyberbullies. Gogoglou *et al.* [58] illustrated how social graph features and SVMs can aid in highlighting those relationships that are susceptible to online grooming. Lastly, Ferrara *et al.* [15] predicted online extremism using the identity of the user and his/her behaviour features through random forest trees and logistic regression models.

Although the studies mentioned were successful, two shortcomings were found:

- The studies aimed at detecting deceptive humans used engineered features that relied on the behaviour of the account. Mining account behaviour, which is part of the account's messages, is computationally costly and time consuming. This is a problem when deception detection should be real-time and not reactive.
- Past research was concerned with accounts being deceptive in general, and not that they were deceptive given their identity. In other words, detecting unsolicited content from accounts constitutes one way to curb deception, but another is to find actual individuals who are lying about their identity.

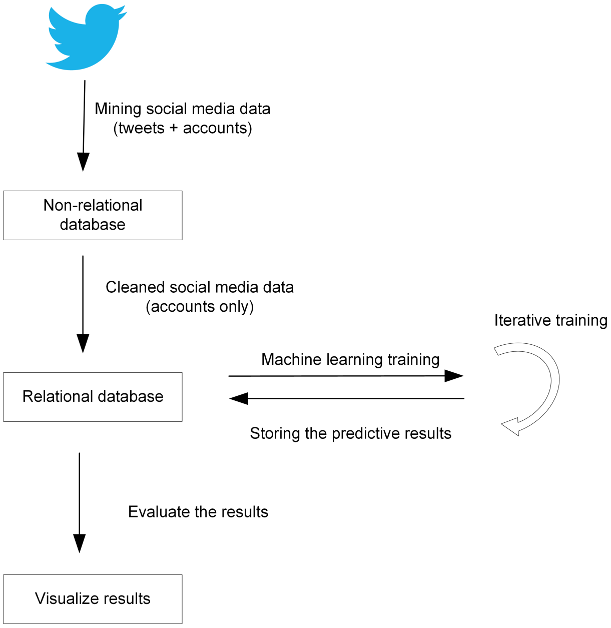

FIGURE 1. **The flow of data to detect identity deception.**

Very little has been done so far to detect actual fake human identities on SMPs, independent of their behaviour. However, bot detection approaches and past research in psychology showed clear promise in detecting fake non-human identities.

We conducted multiple machine learning experiments with existing bot detection approaches to evaluate their efficacy in detecting identity deception committed by humans for malicious purposes. It is hoped that these approaches will address the shortcomings mentioned and serve as a catalyst for uncovering identity deception by humans on SMPs.

III. FINDING DECEPTIVE ACCOUNTS

During the process of detecting identity deception by humans, data is mined, cleaned, stored and applied to supervised machine learning models, and the results are evaluated. This flow of data is illustrated in Fig. 1.

For this research, social media data from Twitter was mined using the twitter4J [27] application programming interface (API) and a non-relational database, Hadoop [59]. Non-relational databases cater for the unstructured nature and vast volumes expected from mining Twitter data [60].

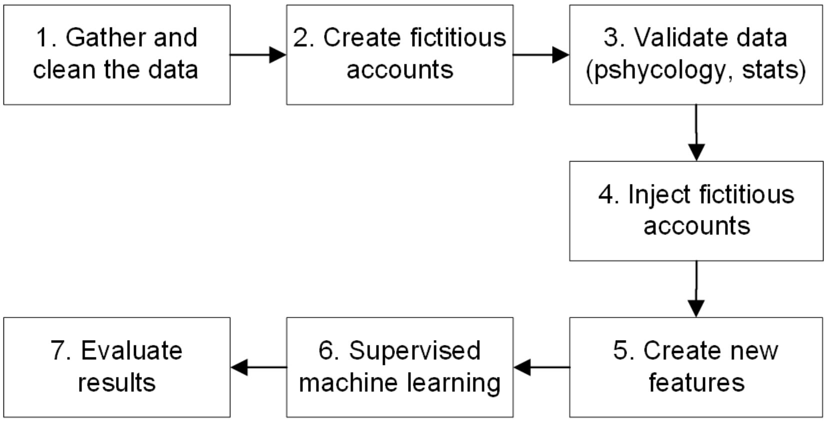
1. Gather and clean the data
2. Create fictitious accounts
3. Validate data (pshycology, stats)
4. Inject fictitious accounts
5. Create new features
6. Supervised machine learning
7. Evaluate results

FIGURE 2. Research steps.

This resulted in a corpus of over 200 million tweets from 223 796 accounts opened between 2006 and 2017. Thereafter, only data related to the identity of the accounts was injected into a relational database, namely SAP HANA [61]. The account data is required for the proposed experiment to determine if past bot detection approaches can apply to humans as well. A relational in-memory database such as SAP HANA, is well suited to process data at speed [61], although a non-relational database would have served the purpose just as well. Supervised machine learning models were trained with the cleaned data and the results were written back to the relational database. The results were lastly visualised and compared to determine whether features engineered to detect bots could be applied to detect fake human identities on SMPs.

Different research steps were executed to discover deceptive accounts. We were specifically interested in human accounts and, more precisely, identity deception in human accounts. To discover such deceptive accounts, we followed the research steps listed next. Fig. 2 presents the steps in diagram format.

1) The corpus was cleaned of bot and cyborg accounts as far as possible. Previous research suggested various simple rules to distinguish humans from bots [7], [16], for example that human accounts will always have a name and image. We applied the rules determined by Cresci *et al.* [16], such as discarding accounts that have 30 or more followers. The remaining corpus consisted of 154 517 accounts. It is expected that the corpus might still contain some bots and even fake human accounts. The few remaining fake accounts should have very little effect on the supervised machine learning models as the majority of the class has been classified correctly.
2) 15 000 fictitious deceptive accounts were created by the authors of the current paper. These deceptive accounts were manually created as if by humans and therefore not by bots. The reason being due to ethical considerations for reporting on sensitive content, such as social media data and the lack of an existing dataset for research. By including examples of deceptive accounts, we avoided the risk of reporting sensitive content by mistake. To ensure that the attributes introduced were indeed deceptive, we looked towards research in psychology showing that people lie on their age [24], [55], gender [25], [55], image [49], location [23], [56], and their name [49], [57] the most. We therefore ensured that the new fictitious accounts were deceptive on all 5 these areas to create accounts as deceptive as possible. The injected accounts were classified as ‘‘fake’’ and the original corpus accounts were classified as ‘‘human’’. Table 2 presents an example of one fictitious deceptive account. This account is perceived as fake due to a number of potential reasons: the name and screenname are unrelated, the image represents a different gender than suggested by the name, the latitude and longitude coincide somewhere over the Arctic ocean, and New York’s UTC offset is actually -4 and not -18 as suggested.

TABLE 2. Example of a fictitious deceptive account.

Attribute	Value
NAME	Marco Duran
SCREENNAME	tinyfrog537
CREATED	3/6/2016
PROFILE_IMAGE	https://randomuser.me/api/portraits/women/99.jpg
LOCATION	Las Palmas De Gran Canaria
LANGUAGE	En
FRIENDS_COUNT	14
FOLLOWERS_COUNT	924
STATUS_COUNT	1 670
LISTED_COUNT	409
TIMEZONE	America/New York
UTC_OFFSET	-18 000
LATITUDE	85.03769
LONGITUDE	-136.27587

3) To ensure that no bias was introduced with the fictitious accounts, this set was compared with the original corpus by means of two statistical tests. The Mann-Whitney-U test proofs that the means of the two sets

TABLE 3. Engineered features previously used to detect fake bot accounts that were added to the corpus.

Short name	Feature description
ACC_AGE_MONTHS	The account age in months
DUP_PROFILE	Whether the account has a duplicate profile
FF_RATIO	Friends-to-followers ratio
HAS_IMAGE	Whether the account has a profile image
HAS_NAME	Whether the account has a profile name
HAS_PROFILE	Whether the account has a profile description
PROFILE_HAS_URL	Whether the profile description contains a URL
USERNAME_LENGTH	The length of the profile username

are similar per attribute [62]. The Chi Square test for independence proofs that the datasets are not correlated and therefor independent [62]. This means that both the deceptive and original corpus must have similar data and show the same distributions.

4) The next step was to inject these fictitious accounts into the original mined corpus.
5) Up to now, the corpus consisted of attributes found in social media only. The corpus was further enriched with engineered features that were taken from past research [16] and were able to successfully detect bot accounts by using data that describe the identity of the account only. These engineered features are shown in Table 3.
6) Supervised machine learning models were trained, using cross validation and resampling, to detect the accounts denoted as fake in the corpus. The machine learning models used were random forest, boosting, and support vector machines as they had been successfully used in past research towards spam [36], [37] and bot detection [16].
   - For the random forest model, the rf library in R software was used. The random forest model creates many variations of trees. The best outcome will be used to predict identity deception. This model works well for bot detection, as rules are easily represented in tree format [2]. An example would be where accounts that have an image or name are considered human, whereas the rest are denoted as bots. Each of these outcomes represents a different section in the tree.
   - For the boosting model, the Adaboost function in R was used. This is a popular model for detecting bots [10], as different features are assigned different weights to predict the outcome. The model makes use of decision trees [2], which are iteratively adjusted with weights. After each iteration, identity deception detection effectiveness is evaluated. This iterative process is continued until the best result towards identity deception detection is achieved.
   - Lastly, for the support vector model, the svmLinear library in R software was used. This algorithm is typically used to model curves on a hyperplane [63], [64]. Trees typically split on single features, whereas SVMs can do so on combinations of features. The SVM algorithm accounts for complex features identifying fake accounts that were missed by trees.

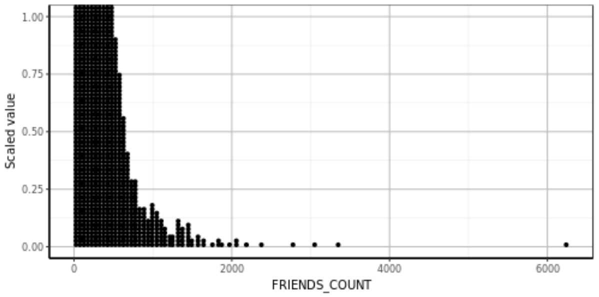

FIGURE 3. Distributions of friends.

7) Once the supervised machine learning models had been trained, their effectiveness was evaluated. We used the following metrics to determine the effectiveness of each model:
   - Accuracy - this determined how many accounts from the total corpus were correctly identified as fake or not.
   - F1 Score - this was a measure of the harmonic mean between precision and recall. "Precision" refers to how successful the model was at detecting identity deception by humans; "recall" means how successful the model was at filtering out the human accounts that were truthful about their identity.
   - Precision-Recall Area under curve (PR-AUC) - this is the statistical value of the area under the precision-recall curve. The PR-AUC measured how successful the current model (in totality) was at predicting identity deception by humans.

IV. RESEARCH RESULTS

The engineered features created during step 5 of the research were explored to understand the corpus and it was noted that most accounts had few friends and followers. The distribution of friends is shown in Fig. 3.

Next, the data exploration looked at the profile descriptions of these accounts. The exploration showed that not all accounts had a profile description and that some profile descriptions were shared among accounts. A few profile descriptions also contained URLs. The results of profiles having an URL as part of their profile is shown in Fig. 4 where 0 = no and 1 = yes. These exploratory results showed that even though we are dealing with human accounts only, they still show characteristics known to bots, such as having a URL in their profile description. This further affirmed that research previously conducted to detect fake bot accounts on SMPs could well be applicable to detect fake human identities too.

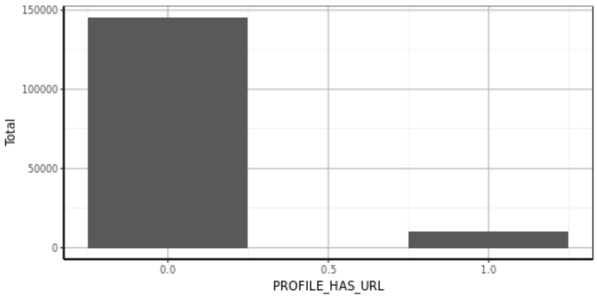

FIGURE 4. Number of accounts with a URL in the profile.

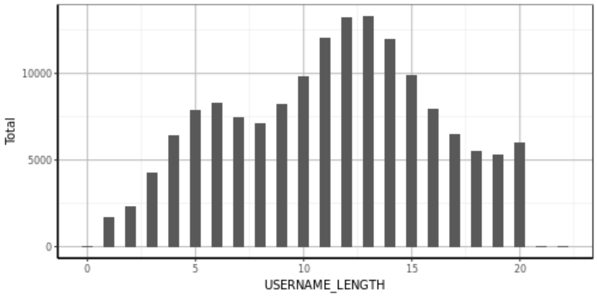

FIGURE 5. Number of accounts per user name length.

TABLE 4. Supervised machine learning results.

Model	Accuracy	F1 score	PR-AUC
svmLinear	68.05%	32.16%	27.76%
rf	87.11%	49.75%	49.90%
Adaboost	85.91%	47.54%	49.53%

Furthermore, a tailed distribution seemed to occur regarding the length of user names chosen for accounts. Any outliers on this distribution could indicate potential deception. This pattern, which is illustrated in Fig. 5, is useful, as supervised machine learning models will be able to detect this type of anomaly.

In summary, it was shown that even though the corpus had mostly been cleansed of bots, the engineered features that were used in past research to detect bot accounts were still present in the corpus of human accounts. Examples of these features were the duplicates found in human profile descriptions, the fact that certain human profile descriptions contained URLs, and lastly, the fact that some human profiles had no description at all. These features are just as prevalent in bot accounts. Therefore, it is assumed that the same engineered features and supervised machine learning models can be applied to the human accounts in the hope of detecting fake identities. The supervised machine learning results are shown in Table 4.

The overall accuracy across all machine learning models was very high, with the highest being 87.11%. These results could incorrectly indicate that the supervised machine learning models are good predictors of identity deception by humans on SMPs. The accuracy measure, however, does not account for wrong predictions and suffers in skewed distributions [65], [66]. The specific corpus was a good example of a skewed distribution because only 15 000 accounts of the total corpus were denoted as fake.

TABLE 5. Entropy results.

Features	svmRadial	rf	Adaboost
ACC_AGE_MONTHS	58.70	34.64	58.70
DUP_PROFILE	89.65	79.02	89.65
FF_RATIO	0.57	1.24	0.57
FOLLOWERS_COUNT	1.76	21.67	1.76
FRIENDS_COUNT	41.79	20.07	41.79
GEO_ENABLED	28.10	21.23	28.10
HAS_IMAGE	0.00	0.00	0.00
HAS_NAME	100.00	100.00	100.00
HAS_PROFILE	9.03	11.93	9.03
LISTED_COUNT	14.05	9.68	14.05
PROFILE_HAS_URL	9.87	20.24	9.87
STATUS_COUNT	43.05	51.89	43.05
USERNAME_LENGTH	58.70	34.64	58.70

Therefore, we looked towards the F1 score and PR-AUC results, which account for getting the predictions wrong. At best, an F1 score of 49.75% was achieved from the random forest (rf) machine learning model and a PR-AUC score of 49.90%. These results are just below what one would expect from getting the prediction right by chance (50%).

Furthermore, entropy indicated which of the features contributed most towards identity deception detection. The entropy results from each supervised machine learning model are shown in Table 5. Values are indicated as having an importance out of 100, in which case 100 means the model is completely dependent on the feature. The entropy results showed that username and profile could be dependent features towards the detection of identity deception. In simple terms, it means that humans lie about their name and the description of themselves on SMPs when they are trying to be deceptive.

These findings are very closely related to what is already known about the social sciences and psychology. From psychology we know that deceptive people lie about their name [49], [57] and age [24], [55]. We also learn that people lie about their image [49], location [23], [56], and gender [25], [55].

This can be used in going forward to engineer more features in the hope of uncovering identity deception by humans. For example, does the gender presented in the profile image match the gender of the name provided for the account?

V. SUMMARY OF THE RESULTS OBTAINED

What we learned and gathered from the experiment discussed in this paper:

- There are many attributes available in SMPs that describe the identity of an SMP account. For example, the name, location, and profile image.

- Human accounts and bot accounts have similar attributes and they share similar characteristics. For example, human accounts have a name and so do accounts generated by bots.
- Features can be engineered from SMP attributes similar to what has been engineered in past research to detect fake accounts generated by bots or computers (for example, whether the account is a duplicate of another).
- Engineered features that have been created to detect fake identities generated by bots can be applied to the existing corpus of human accounts.
- The predictive results from the trained machine learning models only yielded a best F1 score of 49.75%. Given that predicting the correct answer by chance alone would be represented as 50%, this is not optimal.
- Even though only three machine learning models were used in the experiments, these machine learning models have been successfully used in the past towards spam and bot detection. Given the results, these machine learning models are unable to detect fake humans.
- Entropy presents an indication of which engineered features performed well and which not. For example, the fact that an account had a duplicate profile seemed to have made a difference in the accuracy of the predictions.

Based on the predictive results from the machine learning models, it seems that existing features and machine learning models used to detect bot accounts are not suited to detect fake human accounts.

VI. CONCLUSION

The main contribution of this paper is to show that the engineered features that were previously used to detect fake accounts generated by bots are not similarly successful in the detection of fake accounts generated by humans.

This paper reports on a study that focused on detecting fake accounts created by humans, as opposed to those created by bots. We investigated whether the results from past studies to detect bot accounts could be applied successfully to detect fake human accounts. A corpus of human accounts was enriched with engineered features that had previously been used to successfully detect fake accounts created by bots. These features were applied to various supervised machine learning models. The machine learning models were trained to use engineered features without relying on behavioural data. This made it possible for these machine learning models to be trained on very little data, compared to when behavioural data is included.

The findings indicate that engineered features that were previously used to detect fake accounts generated by bots, at best predicted fake accounts generated by humans with an F1 score of 49.75%. This can be attributed to the fact that humans have different characteristics and behaviours than bots which cannot be modelled similarly. Human fake accounts are also not as common as fake accounts generated by bots. Machine learning models might miss these sparse deceptions in the mass.

Future work will investigate the enrichment of the feature set used in the research for this paper by engineering features from the social sciences knowledge domain - especially psychology. The aim will be to enrich the corpus with new features engineered from the same attributes, as used in this study, found on SMPs. It is hoped that these new features will show better results in the detection of identity deception on SMPs.

ACKNOWLEDGMENT

Estée van der Walt would like to thank the HPI Future SOC research lab in Potsdam, Germany for their assistance and support during the research. The HPI Future SOC research lab provided the SAP HANA instance, research environment, and storage used during the experimentations.

REFERENCES

[1] S. Gurajala, J. S. White, B. Hudson, B. R. Voter, and J. N. Matthews, "Profile characteristics of fake Twitter accounts," *Big Data Soc.*, vol. 3, no. 2, p. 2053951716674236, 2016, doi: 10.1177/2053951716674236.

[2] C. Xiao, D. M. Freeman, and T. Hwa, "Detecting clusters of fake accounts in online social networks," in *Proc. 8th ACM Workshop Artif. Intell. Secur.*, 2015, pp. 91–101.

[3] S. Mainwaring, *We First: How Brands and Consumers Use Social Media to Build a Better World.* New York, NY, USA: Macmillan, 2011.

[4] V. S. Subrahmanian *et al.* (2016). "The DARPA Twitter bot challenge." [Online]. Available: https://arxiv.org/abs/1601.05140

[5] Y. Li, O. Martinez, X. Chen, Y. Li, and J. E. Hopcroft, "In a world that counts: Clustering and detecting fake social engagement at scale," in *Proc. 25th Int. Conf. World Wide Web*, 2016, pp. 111–120.

[6] K. Thomas, C. Grier, D. Song, and V. Paxson, "Suspended accounts in retrospect: An analysis of Twitter spam," in *Proc. ACM SIGCOMM Conf. Internet Meas. Conf.*, 2011, pp. 243–258.

[7] T. Tuna *et al.*, "User characterization for online social networks," *Social Netw. Anal. Mining*, vol. 6, no. 1, p. 104, 2016.

[8] P. Galán-García, J. G. De La Puerta, C. L. Gómez, I. Santos, and P. G. Bringas, "Supervised machine learning for the detection of troll profiles in Twitter social network: Application to a real case of cyberbullying," *Logic J. IGPL.* vol. 24, no. 1, pp. 42–53, 2015.

[9] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi, "Faking sandy: Characterizing and identifying fake images on Twitter during hurricane sandy," in *Proc. 22nd Int. Conf. World Wide Web*, 2013, pp. 729–736.

[10] J. P. Dickerson, V. Kagan, and V. S. Subrahmanian, "Using sentiment to detect bots on Twitter: Are humans more opinionated than bots?" in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2014, pp. 620–627.

[11] S. Gurajala, J. S. White, B. Hudson, and J. N. Matthews, "Fake Twitter accounts: Profile characteristics obtained using an activity-based pattern detection approach," in *Proc. Int. Conf. Social Media Soc.*, 2015, p. 9.

[12] B. Viswanath *et al.*, "Towards detecting anomalous user behavior in online social networks," in *Proc. Usenix Secur.*, vol. 14. 2014, pp. 223–238.

[13] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Who is tweeting on Twitter: Human, bot, or cyborg?" in *Proc. 26th Annu. Comput. Secur. Appl. Conf.*, 2010, pp. 21–30.

[14] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna, "Compa: Detecting compromised accounts on social networks," in *Proc. NDSS*, 2013, pp. 1–17.

[15] E. Ferrara, W.-Q. Wang, O. Varol, A. Flammini, and A. Galstyan, "Predicting online extremism, content adopters, and interaction reciprocity," in *Proc. Int. Conf. Social Inform.*, 2016, pp. 22–39.

[16] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, ''Fame for sale: Efficient detection of fake Twitter followers,'' *Decision Support Syst.*, vol. 80, pp. 56–71, Dec. 2015.

[17] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, ''Detecting spammers on Twitter,'' in *Proc. Collaboration, Electron. Messaging, Anti-Abuse Spam Conf. (CEAS)*, vol. 6. 2010, p. 12.

[18] H. Kwak, C. Lee, H. Park, and S. Moon, ''What is Twitter, a social network or a news media?'' in *Proc. 19th Int. Conf. World Wide Web*, 2010, pp. 591–600.

[19] Z. Zhang and B. B. Gupta, ''Social media security and trustworthiness: Overview and new direction,'' *Future Generat. Comput. Syst.*, to be published, doi: 10.1016/j.future.2016.10.007.

[20] A. K. Jain and B. B. Gupta, ''Phishing detection: Analysis of visual similarity based approaches,'' *Secur. Commun. Netw.*, vol. 2017, Jan. 2017, Art. no. 5421046. [Online]. Available: https://doi.org/10.1155/2017/5421046

[21] R. J. Oentaryo, A. Murdopo, P. K. Prasetyo, and E.-P. Lim, ''On profiling bots in social media,'' in *Proc. Int. Conf. Social Inform.*, 2016, pp. 92–109.

[22] M. Tsikerdekis and S. Zeadally, ''Multiple account identity deception detection in social media using nonverbal behavior,'' *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 8, pp. 1311–1321, Aug. 2014.

[23] J. T. Hancock, ''Digital deception,'' in *Oxford Handbook of Internet Psychology*. London, U.K.: Oxford Univ. Press, 2007, pp. 289–301.

[24] G. A. Wang, H. Chen, J. J. Xu, and H. Atabakhsh, ''Automatically detecting criminal identity deception: An adaptive detection algorithm,'' *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 36, no. 5, pp. 988–999, Sep. 2006.

[25] E. Bergen *et al.*, ''The effects of using identity deception and suggesting secrecy on the outcomes of adult-adult and adult-child or-adolescent online sexual interactions,'' *Victims Offenders*, vol. 9, no. 3, pp. 276–298, 2014.

[26] E. Van der Walt and J. H. P. Eloff, ''Protecting minors on social media platforms—A big data science experiment,'' in *Proc. HPI Cloud Symp.*, 2015, pp. 1–78.

[27] Twitter. (2017). *Twitter API*. [Online]. Available: https://dev.twitter.com/overview/api

[28] Facebook. (2017). *The Facebook Graph API*. [Online]. Available: https://developers.facebook.com/docs/graph-api/overview

[29] Instagram. (2017). *Instagram API*. [Online]. Available: https://www.instagram.com/developer/

[30] M. Jiang, P. Cui, and C. Faloutsos, ''Suspicious behavior detection: Current trends and future directions,'' *IEEE Intell. Syst.*, vol. 31, no. 1, pp. 31–39, Jan. 2016.

[31] Wikipedia. (2017). *Spamming*. [Online]. Available: https://en.wikipedia.org/wiki/Spamming

[32] M. S. Alishahi, M. Mejri, and N. Tawbi, ''Clustering spam emails into campaigns,'' in *Proc. Int. Conf. Inf. Syst. Secur. Privacy (ICISSP)*, Feb. 2015, pp. 90–97.

[33] P. Hayati and V. Potdar, ''Toward spam 2.0: An evaluation of Web 2.0 anti-spam methods,'' in *Proc. 7th IEEE Int. Conf. Ind. Inform. (INDIN)*, Jun. 2009, pp. 875–880.

[34] M. Fire, D. Kagan, A. Elyashar, and Y. Elovici, ''Friend or foe? Fake profile identification in online social networks,'' *Social Netw. Anal. Mining*, vol. 4, no. 1, p. 194, 2014.

[35] N. M. Radziwill and M. C. Benton. (2016). ''Bot or not? Deciphering time maps for tweet interarrivals.'' [Online]. Available: https://arxiv.org/abs/1605.06555

[36] S. K. Tuteja, ''A survey on classification algorithms for email spam filtering,'' *Int. J. Eng. Sci.*, vol. 6, no. 5, pp. 5937–5940, 2016.

[37] N. Choudhary and A. K. Jain, ''Towards filtering of SMS spam messages using machine learning based technique,'' in *Advanced Informatics for Computing Research*. Singapore: Springer, 2017, pp. 18–30.

[38] S. Miller and C. Busby-Earle, ''The impact of different botnet flow feature subsets on prediction accuracy using supervised and unsupervised learning methods,'' *J. Internet Technol. Secured Trans.*, vol. 5, no. 2, pp. 474–485, Jun. 2016.

[39] M. Ebrahimi, C. Y. Suen, O. Ormandjieva, and A. Krzyzak, ''Recognizing predatory chat documents using semi-supervised anomaly detection,'' *Electron. Imag.*, vol. 2016, no. 17, pp. 1–9, 2016.

[40] C. Bellinger, S. Sharma, and N. Japkowicz, ''One-class versus binary classification: Which and when?'' in *Proc. 11th Int. Conf. Mach. Learn. Appl. (ICMLA)*, vol. 2. Dec. 2012, pp. 102–106.

[41] L. M. Jupe, A. Vrij, G. Nahari, S. Leal, and S. A. Mann, ''The lies we live: Using the verifiability approach to detect lying about occupation,'' *J. Articles Support Null Hypothesis*, vol. 13, no. 1, pp. 1–13, 2016.

[42] C. Beleites, K. Geiger, M. Kirsch, S. B. Sobottka, G. Schackert, and R. Salzer, ''Raman spectroscopic grading of astrocytoma tissues: Using soft reference information,'' *Anal. Bioanal. Chem.*, vol. 400, no. 9, p. 2801, 2011.

[43] G. Gu, R. Perdisci, J. Zhang, and W. Lee, ''BotMiner: Clustering analysis of network traffic for protocol-and structure-independent botnet detection,'' in *Proc. USENIX Secur. Symp.*, vol. 5. 2008, pp. 139–154.

[44] W. Wu, J. Alvarez, C. Liu, and H.-M. Sun, ''Bot detection using unsupervised machine learning,'' *Microsyst. Technol.*, vol. 24, no. 1, pp. 209–217, 2018.

[45] M. Yahyazadeh and M. Abadi, ''BotOnus: An online unsupervised method for botnet detection,'' *ISC Int. J. Inf. Secur.*, vol. 4, no. 1, pp. 51–62, 2012.

[46] S. Venkatesan, M. Albanese, A. Shah, R. Ganesan, and S. Jajodia, ''Detecting stealthy botnets in a resource-constrained environment using reinforcement learning,'' in *Proc. Workshop Moving Target Defense*, 2017, pp. 75–85.

[47] M. H. Arif, J. Li, M. Iqbal, and K. Liu, ''Sentiment analysis and spam detection in short informal text using learning classifier systems,'' in *Soft Computing*. Berlin, Germany: Springer, 2017, pp. 1–11.

[48] D. Bogdanova, P. Rosso, and T. Solorio, ''Exploring high-level features for detecting cyberpedophilia,'' *Comput. Speech Lang.*, vol. 28, no. 1, pp. 108–120, 2014.

[49] K. Stanton, S. Ellickson-Larew, and D. Watson, ''Development and validation of a measure of online deception and intimacy,'' *Per. Individual Differences*, vol. 88, pp. 187–196, Jan. 2016.

[50] M. A. Al-Garadi, K. D. Varathan, and S. D. Ravana, ''Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network,'' *Comput. Hum. Behav.*, vol. 63, pp. 433–443, Oct. 2016.

[51] X. Zhu, ''Semi-supervised learning literature survey,'' Dept. Comput. Sci., Univ. Wisconsin-Madison, Madison, WI, USA, Tech. Rep. TR 1530, 2005.

[52] M. Drouin, D. Miller, S. M. J. Wehle, and E. Hernandez, ''Why do people lie online? 'Because everyone lies on the Internet,''' *Comput. Hum. Behav.*, vol. 64, pp. 134–142, Nov. 2016.

[53] R. Rong, D. Houser, and A. Y. Dai, ''Money or friends: Social identity and deception in networks,'' *Eur. Econ. Rev.*, vol. 90, pp. 56–66, Nov. 2016.

[54] V. L. Rubin, ''Deception detection and rumor debunking for social media,'' in *The SAGE Handbook of Social Media Research Methods*. Newbury Park, CA, USA: SAGE, 2017, p. 342.

[55] C. L. Toma, J. T. Hancock, and N. B. Ellison, ''Separating fact from fiction: An examination of deceptive self-presentation in online dating profiles,'' *Per. Social Psychol. Bull.*, vol. 34, no. 8, pp. 1023–1036, 2008.

[56] A. Caspi and P. Gorsky, ''Online deception: Prevalence, motivation, and emotion,'' *CyberPsychol. Behav.*, vol. 9, no. 1, pp. 54–59, 2006.

[57] S. Utz, ''Types of deception and underlying motivation: What people think,'' *Social Sci. Comput. Rev.*, vol. 23, no. 1, pp. 49–56, 2005.

[58] A. Gogoglou, Z. Theodosiou, T. Kounoudes, A. Vakali, and Y. Manolopoulos, ''Early malicious activity discovery in microblogs by social bridges detection,'' in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol. (ISSPIT)*, Dec. 2016, pp. 132–137.

[59] A.S. Foundation. (2014). *The Hadoop Distributed File System: Architecture and Design*. [Online]. Available: http://hadoop.apache.org/docs/r2.4.1/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html

[60] R. Kannadasan, R. Shaikh, and P. Parkhi, ''Survey on big data technologies,'' *Int. J. Adv. Eng. Res.*, vol. 3, no. 3, pp. 1–11, 2013.

[61] *SAP Hana*, SAP SE, Walldorf, Germany, 2017.

[62] C. R. Kothari, *Research Methodology: Methods and Techniques*. New Age International, 2004. [Online]. Available: https://books.google.co.uk/books?id=hZ9wSHysQDYC

[63] A. M. Meligy, H. M. Ibrahim, and M. F. Torky, ''Identity verification mechanism for detecting fake profiles in online social networks,'' *Int. J. Comput. Netw. Inf. Secur.*, vol. 9, no. 1, pp. 31–39, 2017.

[64] S. T. Peddinti, K. W. Ross, and J. Cappos. (2017). ''Mining anonymity: Identifying sensitive accounts on Twitter.'' [Online]. Available: https://arxiv.org/abs/1702.00164

[65] G. Menardi and N. Torelli, ''Training and assessing classification rules with imbalanced data,'' *Data Mining Knowl. Discovery*, vol. 28, no. 1, pp. 92–122, 2014.

[66] L. A. Jeni, J. F. Cohn, and F. De La Torre, ''Facing imbalanced data–recommendations for the use of performance metrics,'' in *Proc. Hum. Assoc. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2013, pp. 245–251.

ESTÉE VAN DER WALT was born in Johannesburg, South Africa, in 1977. She received the B.Sc. and M.Sc. degrees in computer science from the University of Johannesburg, South Africa, in 1997 and 2001 respectively. She is currently pursuing the Ph.D. degree in information technology with the University of Pretoria, South Africa.

Her research interests include cybersecurity and identity deception on social media platforms, and in particular, 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 a recipient of the Best Poster Award at the 3rd International Conference on Information Systems Security and Privacy that was hosted in Porto, Portugal, in 2017.

JAN ELOFF received the Ph.D. degree in computer science in 1985. Up to 2015, he was appointed as the Research Director for SAP Research, Africa. He is currently appointed as the Deputy Dean for research and postgraduate studies with the Faculty of Engineering, Built Environment and IT, and as a Full Professor in computer science with the University of Pretoria, South Africa.

Since 2007, he has been an Associate Editor of the *Computers & Security journal* and an Editorial Member of the *International Computer Fraud & Security bulletin* (Elsevier). He is an internationally recognized Researcher and has published 113 peer reviewed papers with 3537 citations.