Identity deception detection: requirements and a model

Estee van der Walt* and Jan Eloff

Department of Computer Science, University of Pretoria, Pretoria, South Africa

*Corresponding author: Estee van der Walt can be contacted at: estee.vanderwalt@gmail.com

Abstract

Purpose
This paper aims to describe requirements for a model that can assist in identity deception detection (IDD) on social media platforms (SMPs). The model that was discovered demonstrates the usefulness of the requirements. The aim of the model is to identify humans lying about their identity on SMPs.

Design/Methodology/Approach
The requirements of a model for IDD will be determined through a literature study combined with a study that identifies currently available identity related metadata on SMPs. This metadata refers to the attributes that describe a user account on an SMP. The aim is to restrict IDD to be only based on these types of attributes, as opposed to or combined with the contents of a single or multiple communications.

Findings
Data science experiments were conducted and in particular supervised machine learning models were discovered that indeed detects identity deception on SMPs with an area under the receiver operator characteristics curve (ROC-AUC) of 75.5%.

Originality/value
SMPs allow any user to easily communicate with their friends or the general public at large. People can now be targeted at great scale, most often for malicious purposes. The reality is that many of these cyber-attacks involve some form of identity deception, where the attackers lie about who they are. Much focus to date has been on the identification of non-human deceptive accounts. This paper focuses on deceptive human accounts that target vulnerable individuals on SMPs.

Keywords

cyber-security, identity deception, fake identities, social media, big data, Twitter

Type

Research paper

1. Introduction

Social media platforms (SMPs) are used for various purposes in the daily lives of individuals and companies alike. These purposes include amongst others online social networking, blogging, wikis, media sharing, online reviews, news groups, microblogging, and geo-location services (McCay-Peet and Quan-Haase, 2017). In 2017, Facebook announced reaching their 2 billionth active online identity (Constine, 2017). An online identity is an individual or company who logged into the Facebook platform in the last 30 days (Facebook, 2018). Facebook continues to report a 10% year on year growth in the total number of active users (Facebook, 2018). This growth has not only connected many individuals with each other but also permitted many capabilities proposing to improve society at large. These capabilities include, amongst others, finding long lost family members given up at birth (Samuels, 2018) and tracking natural disasters (Chun et al., 2014).

The growth in the number of online identities on SMPs and the resulting voluminous data has however made it very difficult, if not impossible, to know who to trust on SMPs (Warner-Søderholm et al., 2018). Furthermore, humans are gullible and do not, for example, have the ability to discern truth from lies (Sandy et al., 2017). Twitter's loss in revenue (Kastrenakes, 2018) prior to 2018 has, for example, been attributed to the abusive behaviour amongst its online users (Jhaver et al., 2018). Lately, Facebook has been scrutinised for similar deceitful activities. Consider for example the online activities in February 2018 where 13 Russians were charged by the United States Justice Department for subverting the 2016 political campaign (Apuzzo and LaFraniere, 2018). They created social media accounts as if they were American citizens with the assumed intention to create discord in the democracy system through the content they posted. In another example from 2017, women were groomed via Facebook, and then raped and killed (de Villiers, 2017) in South Africa. The victims were lured through a fake profile and killed by the very same person they trusted. Within the cyber-security world, these types of activities are commonly known as impersonation or identity deception (Donath, 1999).

This paper focuses on countering the act of identity deception as executed by humans. The fact that non-human online identities, also known as bots, can deceive will be out of scope for this research. It is also acknowledged that humans can lie to protect themselves without malicious intent. The focus of this research will be on malicious humans only. For this paper in particular, an attempt is made to determine requirements for a model suitable for identity deception detection as well as attributes of online identities on SMPs that have the potential to assist in the automatic detection of identity deception. The contributions of the research results reported on in this paper are summarized as follows:

- To identify the attributes freely available on SMPs that can play a role in detecting identity deception through a literature review.
- To define the requirements for approaching online identity deception.
- To show how these requirements could potentially be implemented through an experimental supervised machine learning prototype.

Section 2 of this paper identifies existing identity related attributes found on SMPs. The section furthermore discusses how these attributes have been applied in related work on identity deception detection. This discussion leads into a definition for the requirements, such as to use content from humans only, expected of a prototype aiming to assist in the automatic detection of identity deception on SMPs by humans in section 3. Section 4 shows how these requirements could potentially be implemented through a high-level design of a prototype. Sections 5 discusses the experiment and section 6 gives further insight into the results of the IDD experimental prototype. Section 7 concludes and poses further considerations to implement the requirements expected of a model that detects identity deception by humans.

2. Background and related work

Many examples of cyber threats that have materialised in real-life incidents can be found on SMPs. A literature study, as illustrated in Table 1, revealed threats like identity theft, trolling, flaming, identity deception, cyber stalking, cyber bullying, grooming, and phishing. Take for example the case where a 23-year-old British woman was jailed for grooming a 13-year-old boy via Facebook and later physically abusing the boy (Association, 2017). In another example, two teenagers were arrested for the cyber bullying of another teenager which potentially resulted in her death (Dearen, 2018). These examples illustrate that cyber threats delivered through SMPs can target humans in particular, whereas in the past, cyber threats were more likely to be directed at hardware devices and infrastructure (Chandramouli, 2011). Those past attacks required great skill from malicious individuals, whereas cyber criminals now use SMPs to exploit the vulnerabilities of the typical user. For example, it is nowadays possible to bully another individual anonymously and at very low risk to the attacker (Peddinti et al., 2017).

In these cases of deception, the attackers lie by changing various of their social media account attributes that defines their identities to hide who they are. When looking at how to assist in the automated detection of identity deception on SMPs, it is important to understand not only which attributes exist, but also which attributes can potentially contain false information and therefore have a bigger impact on identity deception. SMP data are mostly known for the content added by its users.

Besides posting content, information about the user's relationships, behaviour, account and profile can be found on SMPs as illustrated in Table 2. SMP users are required to open an account with the SMP before they can start posting content (Facebook, 2017). During this registration process they are requested to give information like their name (Facebook, 2017), location (Twitter, 2018), and even birth date in some cases (LinkedIn, 2017). This additional data is also generally referred to as metadata or attributes (Sloan et al., 2015). These attributes not only identify the user but also serves to distinguish them from another user. Take Twitter for example. In Twitter, the name of the user and the location are examples of attributes describing the user. It is noticeable that the same attributes are found across the different SMPs. This indicates that a proposal towards detecting identity deception could somehow also apply to other SMPs.

Table 1 : Cyber threats found on SMPs against humans

Cyber threats	(Jakobsson, 2018)	(Gharibi and shaabi, 2012)	(Perez, 2011)	(Willard, 2007)	(Fire *et al.*, 2014a)	(Kirichenko *et al.*, 2017)	(Patel *et al.*, 2017)	(Trivedi *et al.*, 2016)	(Pradhan *et al.*, 2016)	(Broome *et al.*, 2018)	(Acar, 2016)
Identity theft Trolling		x			x		x	x	x		
(defamation) Flaming		x									
(a short-lived argument)				x							
Identity deception				x	x	x					
Cyber stalking Cyber		x		x			x		x		x
bullying Grooming		x			x		x				x
(extremism, paedophilia, etc.)					x				x	x	x
Phishing	x		x		x	x	x	x			

Table 2. Data available on SMPs about a user

Attributes/meta data	(Goel *et al.*, 2013)	(Kim *et al.*, 2010)	(Wang *et al.*, 2006)	(Clarke, 1994)
Describing the user profile	Location		Personal information	Appearance, name, code, who you are, physical
Describing the account	Email		Biometrical information	What you have
Behaviour	Mutual interests	Belong to the same group	Biographical information	Social behaviour
Relationships	Friends/ Followers			What you do
Content	Topics			Knowledge

Past related work proposed various identity attributes, and also combined some identity attributes to engineer new features to detect identity deception (Van Der Walt and Eloff, 2018a). Feature engineering is the process of using domain knowledge to construct new pieces of information (Domingos, 2012). These features can be constructed from the content, also known as linguistic features (Scott and Matwin, 1999). In this case, the attributes available in SMPs are used to create new information about the identity of a user. Lee et al. (Lee et al., 2010), Ribeiro et al. (Ribeiro et al., 2018), and Thomas et al. (Thomas et al., 2013) used linguistic features extracted from various SMPs to detect identity deception. Examples of such linguistic features are: the collection of specific words (Ribeiro et al., 2018), repetitions of content (Lee et al., 2010), and sharing the same naming structure. Chiang and Grant (Chiang and Grant, 2018) extracted the intent from a sentence. They then used the sequence of intents, for example a greeting following by a question, to find users with multiple online deceptive profiles. Similarly, Halawi et al. (Halawi et al., 2018) used ontologies to find similar online profiles. An ontology can be used to group conversations about similar topics. Non-verbal attributes like the date the account was opened (Tsikerdekis, 2017), the type of SMP (Thomas et al., 2013), and profile update time (Gurajala et al., 2016) were useful where the information provided for an account is scarce. Network features, like accounts in the same domain (Thomas et al., 2013), friends (Gurajala et al., 2016), and followers (Gurajala et al., 2016) were used to detect deception. Lastly, identity attributes like gender (Hancock and Toma, 2009), location (Alowibdi et al., 2015), profile image (Hancock and Toma, 2009), age (Tuna et al., 2016), profession (Tuna et al., 2016), name (Peddinti et al., 2017), and email (Xiao et al., 2015) were proposed indicators towards detecting identity deception. Many of the attributes used to detect identity deception, required additional processing to extract knowledge about the identity of a user. For example, the content had to be parsed for specific words to determine sentiment (Ribeiro et al., 2018) and each profile image was manually labelled to determine if that user was an adult or not (Tuna et al., 2016). This additional work required, adds overhead to a model proposing to assist in the automated detection of human identity deception on SMPs.

Cresci et al. (Cresci et al., 2015) and Varol et al. (Varol et al., 2017) used a combination of attributes and features in their research with the aim of reducing the overhead required to develop an identity deception detection model. They showed that the identity and non-verbal attributes were not only easy to mine, but also just as accurate at detecting identity deception for bots, compared to using network, linguistics, or other content related features. Even though Cresci et al. (Cresci et al., 2015) and Varol et. al (Varol et al., 2017) focussed on detecting deceptive non-human accounts on SMPs, these same SMP attributes apply to humans. For this reason, the authors propose to use the identity and non-verbal attributes on SMPs in an experiment to not only assist in the automated detection of human identity deception, but also to understand which attributes are more indicative of such deceptiveness.

Different approaches were also found to be used in the detection of identity deception on SMPs. Concepción-Sánchez et al. (Concepción-Sánchez et al., 2018) proposes fuzzy logic to identify identity deception on SMPs from the content posted by users. Krishnamurthy et al.(Krishnamurthy et al., 2018) followed a deep learning approach which combined audio, video, text and facial micro expressions. These studies both however did not use the attributes of an account but rather its content which could be costly to construe. Cresci et al. (Cresci et al., 2015), on the other hand, proposed machine learning algorithms like decision tree, random forest, support vector machines (SVMs), adaptive boosting, k-nearest neighbours and logistic regression for their research experiments. Gupta et al. (Gupta et al., 2013) in turn suggested Naïve Bayes and decision trees to detect bots successfully. Xiao et al. (Xiao et al., 2015) proposed logistic regression, random forests, and SVMs to detect deceptive accounts. Although this research focussed on detecting bots, they achieved success with detecting deceptive identities using the attributes freely available on SMPs alone. Supervised machine learning also requires a labelled dataset (Galán-GarcÍa et al., 2016), which was available for the research at hand. For these reasons, this paper will use supervised machine learning as a method to develop a model that can assist to detect identity deception by humans on SMPs.

3. Establishing the requirements

The following are the requirements for discovering machine learning models for automated identity deception detection. These requirements are based on the literature discussed in the previous section as well as the research results presented in this paper:

- Use a big dataset that consists of a large volume of identity related heterogeneous data (Van der Walt and Eloff, 2015) - SMPs, being a big data platform, are a good source of such data. Other examples of sources of such datasets could be police records (Wang et al., 2004) or known chat logs (Ebrahimi et al., 2016).
- Use attributes freely available on an SMP (Twitter, 2018) (Facebook, 2017) (LinkedIn, 2017) – To support automated identity deception detection, the attributes should be readily available for easy and fast identification. In some

of these cases time is of the essence and could prevent further or more severe actions of the deceptive human (Dearen, 2018).

- Ignore non-human accounts in the SMP data (Cresci et al., 2015) – The researchers believe that the detection of bots should be handled separately from the detection of humans lying about their identity as they have different goals (Van der Walt and Eloff, 2018b).
- Ignore content posted by users on an SMP (Varol et al., 2017) (Cresci et al., 2015) – It has been showed by related research that identity deception can be detected just as accurately without the additional overhead required to process and extract features from the content posted by users.
- The attributes used for the model, should describe the identity of the user (Meligy et al., 2017) – By using attributes that describe an identity, the risk of drawing conclusions from attributes unrelated to identity deception will be minimized. This is also known as the risk that exists in inferring that correlation implies causation (Domingos, 2012).
- The data should contain both examples of deceptive and trustworthy people (Kuhn et al., 2016) - Supervised machine learning requires a labelled dataset.
- Develop a machine learning model (Cresci et al., 2015) – When dealing with a labelled dataset, supervised machine learning has been shown as a method to detect identity deception (van der Walt and Eloff, 2017).
- Compare the results from various machine learning models (Varol et al., 2017) – The results should be reproducible to deduce its relevance in aiding in the detection of human identity deception.
- Automate the detection due to SMPs' big data nature (Chaffey, 2018) – To assist in the automated detection of identity deception and dealing with the voluminous datasets, automation is required.

The next section provides a high-level design for the prototype.

4. High-level design of a prototype for the automated assistance of identity deception detection on SMPs

The proposed prototype aims to *validate* initial ideas (Drasch et al., 2015), *implement* requirements, allow for *experimentation* (Fallman, 2003), and furthers *collaboration* with the industry and/or fellow academics (Gundecha and Liu, 2012, Fallman, 2003).

There are three components of the prototype namely: feature extraction, model construction, and results, as proposed by the work of Bhat et al. (Bhat et al., 2018) to resolve a user's online identity:

- Feature extraction – For the prototype, freely available SMP attributes are available for feature extraction. The attributes should describe the identity of the user and not include any content they posted. The data should also contain examples of both deceptive and trustworthy accounts. To adhere to these requirements, this component retrieves the data from Twitter, cleans the data from any non-human accounts, labels the data for supervised machine learning, and finally prepares the data for supervised machine learning.

- Model construction – Supervised machine learning is required to construct and evaluate models assisting in the detection of human identity deception on SMPs. This component allows for experimentation by using the prepared data to train various supervised machine learning algorithms using different parameters, such as resampling (Domingos, 2012), and hyperparameters (Dickerson et al., 2014).
- Results – Due to the nature of the data, more specifically its volume and heterogeneity, the process if identity deception detection should be automated. This component allows for unassisted identity deception detection and uses the most accurate machine learning model discovered during experimentation. The results will highlight those individuals most likely being deceptive and warranting further in-depth investigation.

For this research, the proposed prototype was built using infrastructure provided by the Future SOC Lab in Potsdam, Germany (FSOC, 2018). The Twitter data was mined using Apache Flume (Apache, 2018), HDFS (Apache Software Foundation, 2014), and finally stored in a SAP HANA (SAP, 2017) in-memory database consisting of 2TB of RAM and 8TB of storage. Machine learning models were built using the Caret package in R (Kuhn et al., 2016). The prototype components, their functions, and how each component addresses the requirements expected of a prototype assisting in the automated detection of human identity deception on SMPs are illustrated in Figure 1. The next section shows some results delivered by the running prototype.

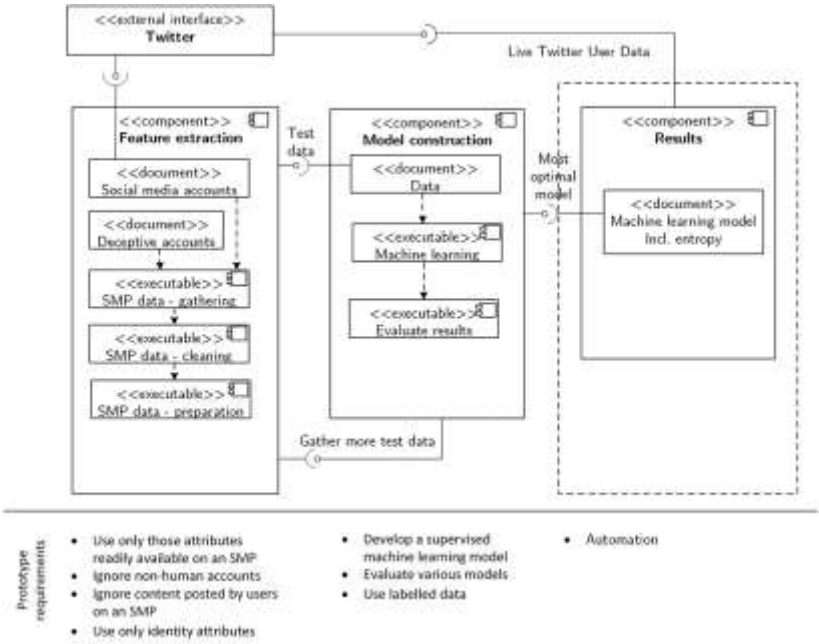

Figure 1: The components of the proposed prototype

5. Experimental results

Identity attributes from Twitter accounts were mined, using a Java API (Yamamoto, 2018) together with Apache Flume (Apache, 2018) during 2016. Apache Flume was able to import volumes of data whilst ignoring non-English-speaking accounts. A total of 606 914 240 tweets were gathered for this research from Twitter. This equates to 200GB of data consisting out of the account profiles of 223 796 Twitter users. 53 091 of the Twitter accounts were discarded at this point, using rules from the research of Cresci et al. (Cresci et al., 2015) that identifies non-human or bot accounts. The contents of the tweets were ignored as the prototype requires only those attributes describing the identity of the user. An additional 15 000 deceptive human accounts were generated using two random human data generator APIs (Armstrong and Hunt, 2017) (Keen, 2017). These fabricated examples of deceptive accounts each had one or more identity attributes not representative of the truth. For example, the location would be a place different from their indicated GPS location. The Pearson's chi-square test of independence (Kothari, 2004) proofed that the fabricated examples were representative of the population. It is expected that the original gathered corpus also includes deceptive accounts. However, deceptive accounts are in the minority and should therefore not have an effect on the results (Halevy et al., 2014).

The experiment used the aforementioned prepared identity data. The results from this experiment, using supervised machine learning and 10-fold cross validation (Peddinti et al., 2017) (Fire et al., 2014b), is shown in Table 3. Given ROC-AUC, which measures the Receiver-Operator Area Under the Curve performance of a machine learning model (Davis and Goadrich, 2006), it is shown that at best, the random forest and nnet (neural net) algorithms detected identity deception by humans with a score of 75.5% and 73.4% (100% being the best, 0 being the worst) respectively. Figure 2 displays the corresponding true positive rate (detecting the deception correctly) vs the false positive rate (indicating a non-deceptive account as deceptive).

Table 3. Summary of experimental results (%)

Algorithm	Accuracy	F1-score	ROC-AUC
Random forest	80.0	33.1	75.5
Adaboost	78.8	28.7	70.2
nnet	73.0	28.2	73.4
bayesglm	65.4	15.8	52.7
J48	77.9	29.6	71.7
knn	71.3	24.6	68.4
rpart	66.3	23.6	63.1
svmRadial	15.3	16.0	53.7

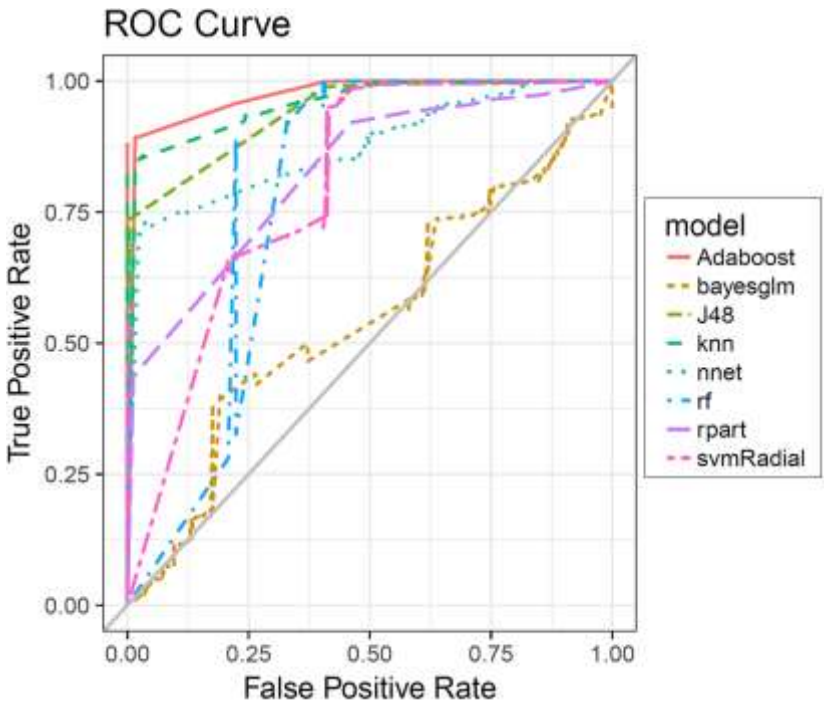

Figure 2 : Experimental results (ROC curve)

The experimental results are discussed in more detail next.

6. Discussion of results

For this paper, the Accuracy, F1-score, and ROC-AUC (area under the receiver operator curve) was considered to assist in the evaluation of the models. The F1-score and AUC metrics are often used in research detecting spam and bot accounts to determine the effectiveness of the machine learning models (Ferrara et al., 2016) (Fire et al., 2014b) (Xiao et al., 2015). Although 10-fold cross validation was used to train the models, it is known that the F1-score suffers (Menardi and Torelli, 2014) (Jeni et al., 2013) in skewed distributions. More recently the PR-AUC (Precision-Recall Curve) has been recommended as an alternative to ROC-AUC (Saito and Rehmsmeier, 2017) (Davis and Goadrich, 2006).

The following recommendations are proposed to address the potential misleading model performances and to further verify the results from the prototype:

- Ensure that the labelled dataset is equally distributed. Techniques like over- and under-sampling can be used (Dal Pozzolo et al., 2013).
- Additional metrics can be used to measure the success of the models. These metrics can include, amongst others, Kappa (Powers, 2011) and PR-AUC (Saito and Rehmsmeier, 2017).

- Experiment with additional features to increase the accuracy of the prototype. For example, by combining SMP attributes like whether the gender on the profile image matches the gender of the SMP user, further lies can potentially be identified. These attributes should still be freely available on SMPs.
- Improve the completeness of attributes on SMPs as many identity attributes were found to be incomplete i.e. not completed by the users at the time of creating the user account. If some of these attributes, like location and profile image were made compulsory by the SMP provider, identity deception detection accuracy could potentially increase.
- Additional validation could be performed by SMPs upon user registration to ensure the veracity of SMP attributes. By, for example, getting someone else to validate that the profile image is representative of that user, could prevent potential identity deception.

7. Conclusion and future work

Requirements that can be useful for the discovery of identity deception detection models are described in this paper. The paper shows that there are multiple attributes, or metadata entities, currently available on SMPs, that could be employed to detect identity deception by humans. It is also shown that data belonging to these attributes could be easily gathered enabling automated human identity deception detection. An experimental prototype shows how supervised machine learning models can be discovered that best assist in the task of automatically detecting identity deception. Future work will focus on increasing the accuracy of the machine learning models. One way of achieving this will be to introduce engineered features such as "age-determined-from-profile-image".

References

Acar, K. V. 2016. Sexual Extortion of Children in Cyberspace. *International Journal of Cyber Criminology,* 10**,** 110-126.

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

Apache. 2018. Flume. Available: https://flume.apache.org/ [Accessed 16 May 2018].

Apache Software Foundation 2014. The Hadoop Distributed File System: Architecture and Design. v2.4.1 ed.

Apuzzo, M. & Lafraniere, S. 2018. 13 Russians Indicted as Mueller Reveals Effort to Aid Trump Campaign. *The New York Times*, 16 Feb 2018.

Armstrong, K. & Hunt, A. 2017. Random User Generator. Available: https://randomuser.me/ [Accessed 8 January 2018].

Association, B. P. 2017. Woman jailed for abusing boy after grooming him over Facebook. *Daily Mail Online* [Online]. Available: http://www.dailymail.co.uk/wires/pa/article-5206541/Woman-jailed-abusing-boy-grooming-Facebook.html [Accessed 22 December 2017].

Bhat, S. I., Arif, T. & Malik, M. B. 2018. A Framework for User Identity Resolutions across Social Networks.

Broome, L. J., Izura, C. & Lorenzo-Dus, N. 2018. A systematic review of fantasy driven vs. contact driven internet-initiated sexual offences: Discrete or overlapping typologies? *Child abuse & neglect*, 79**,** 434-444.

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 June 2018].

Chandramouli, R. 2011. Emerging social media threats: Technology and policy perspectives. Cybersecurity Summit (WCS), June 01-02, 2011 2011 London, United Kingdom. IEEE, 1-4.

Chiang, E. & Grant, T. 2018. Deceptive Identity Performance: Offender Moves and Multiple Identities in Online Child Abuse Conversations. *Applied Linguistics*.

Chun, Y., Hwang, H. S. & Kim, C. S. 2014. Development of a Disaster Information Extraction System based on Social Network Services. *International Journal of Multimedia and Ubiquitous Engineering*, 9**,** pp.255-264.

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

Concepción-Sánchez, J. Á., Molina-Gil, J., Caballero-Gil, P. & Santos-González, I. Fuzzy Logic System for Identity Theft Detection in Social Networks. 2018 4th International Conference on Big Data Innovations and Applications (Innovate-Data), 2018. IEEE, 65-70.

Constine, J. 2017. Facebook now has 2 billion monthly users and responsibility. *TechCruch.com*, 27.

Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A. & Tesconi, M. 2015. Fame for sale: efficient detection of fake Twitter followers. *Decision Support Systems*, 80**,** 56-71.

Dal Pozzolo, A., Caelen, O., Waterschoot, S. & Bontempi, G. 2013. Racing for unbalanced methods selection. International Conference on Intelligent Data Engineering and Automated Learning, 2013. Springer, 24-31.

Davis, J. & Goadrich, M. 2006. The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning, 2006. ACM, 233-240.

De Villiers, J. 2017. 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 [Accessed 4 November 2017].

Dearen, J. 2018. Pre-Teens Arrested for Cyberbullying Before Girl's Suicide. *US News & World Report*.

Dickerson, J. P., Kagan, V. & Subrahmanian, V. 2014. Using sentiment to detect bots on Twitter: Are humans more opinionated than bots? Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on, 2014. IEEE, 620-627.

Domingos, P. 2012. A few useful things to know about machine learning. *Communications of the ACM*, 55**,** 78-87.

Donath, J. S. 1999. Identity and deception in the virtual community. *Communities in cyberspace*, 1996**,** 29-59.

Drasch, B., Huber, J., Panz, S. & Probst, F. 2015. Detecting Online Firestorms in Social Media. Proceedings of the 36th International Conference on Information Systems

(ICIS), December 2015 2015 Fort Worth, USA.

Ebrahimi, M., Suen, C. Y., Ormandjieva, O. & Krzyzak, A. 2016. Recognizing Predatory Chat Documents using Semi-supervised Anomaly Detection. *Electronic Imaging*, 2016**,** 1-9.

Facebook. 2017. The Facebook Graph API. Available: https://developers.facebook.com/docs/graph-api/overview [Accessed 8 January 2018].

Facebook 2018. Facebook Reports Third Quarter 2018 Results.
Fallman, D. 2003. Design-oriented human-computer interaction. Proceedings of the SIGCHI conference on Human factors in computing systems, 2003. ACM, 225-232.
Ferrara, E., Wang, W.-Q., Varol, O., Flammini, A. & Galstyan, A. 2016. Predicting online extremism, content adopters, and interaction reciprocity. International Conference on Social Informatics, 2016. Springer, 22-39.
Fire, M., Goldschmidt, R. & Elovici, Y. 2014a. Online social networks: threats and solutions. *IEEE Communications Surveys \& Tutorials*, 16**,** 2019-2036.
Fire, M., Kagan, D., Elyashar, A. & Elovici, Y. 2014b. Friend or foe? Fake profile identification in online social networks. *Social Network Analysis and Mining*, 4**,** 1-23.
Fsoc. 2018. The HPI Future SOC lab. Available: https://hpi.de/en/research/future-soc-lab.html [Accessed 8 June 2018].
Galán-García, P., De La Puerta, J. G., Gómez, C. L., Santos, I. & Bringas, P. G. 2016. Supervised machine learning for the detection of troll profiles in Twitter social network: Application to a real case of cyberbullying. *Logic Journal of IGPL*, 24**,** 42-53.
Gharibi, W. & Shaabi, M. 2012. Cyber threats in social networking websites. *CoRR*, abs/1202.2420.
Goel, A., Sharma, A., Wang, D. & Yin, Z. 2013. Discovering Similar Users on Twitter. *11th Workshop on Mining and Learning with Graphs.*
Gundecha, P. & Liu, H. 2012. Mining social media: a brief introduction. *Tutorials in Operations Research*, 1.
Gupta, A., Lamba, H., Kumaraguru, P. & Joshi, A. 2013. Faking Sandy: characterizing and identifying fake images on Twitter during hurricane Sandy. Proceedings of the 22nd international conference on World Wide Web, 2013. ACM, 729-736.
Gurajala, S., White, J. S., Hudson, B., Voter, B. R. & Matthews, J. N. 2016. Profile characteristics of fake Twitter accounts. *Big Data \& Society*, 3**,** 1-13.
Halawi, B., Mourad, A., Otrok, H. & Damiani, E. 2018. Few are as Good as Many: An Ontology-Based Tweet Spam Detection Approach. *IEEE Access*, 6**,** 63890-63904.
Halevy, R., Shalvi, S. & Verschuere, B. 2014. Being honest about dishonesty: Correlating self- reports and actual lying. *Human Communication Research*, 40**,** 54-72.
Hancock, J. T. & Toma, C. L. 2009. Putting your best face forward: The accuracy of online dating photographs. *Journal of Communication*, 59**,** 367-386.
Jakobsson, M. 2018. Two-factor inauthentication–the rise in SMS phishing attacks. *Computer Fraud & Security*, 2018**,** 6-8.
Jeni, L. A., Cohn, J. F. & De La Torre, F. 2013. Facing imbalanced data--Recommendations for the use of performance metrics. Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII), 2013. IEEE, 245-251.
Jhaver, S., Ghoshal, S., Bruckman, A. & Gilbert, E. 2018. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 25**,** 12.
Kastrenakes, J. 2018. Twitter's user numbers are growing again. *The Verge*.
Keen, B. 2017. Generate Data. Available: http://www.generatedata.com/ [Accessed 8 January 2018].
Kim, D., Jo, Y., Moon, I.-C. & Oh, A. 2010. Analysis of Twitter lists as a potential source for discovering latent characteristics of users. ACM CHI Workshop on Microblogging, 2010.
Kirichenko, L., Radivilova, T. & Carlsson, A. 2017. Detecting cyber threats through social network analysis: short survey. *SocioEconomic Challenges*, 1**,** 20-34.
Kothari, C. R. 2004. *Research methodology: Methods and techniques*, New Age International.
Krishnamurthy, G., Majumder, N., Poria, S. & Cambria, E. 2018. A deep learning approach for multimodal deception detection. *arXiv preprint arXiv:1803.00344*.

Kuhn , M., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T. & Mayer, Z. 2016. caret: Classification and regression training. 6.0-73 ed.

Lee, K., Caverlee, J. & Webb, S. 2010. The social honeypot project: protecting online communities from spammers. Proceedings of the 19th international conference on World wide web, 2010. ACM, 1139-1140.

Linkedin. 2017. LinkedIn Developers. Available: https://developer.linkedin.com/ [Accessed 8 January 2018].

Mccay-Peet, L. & Quan-Haase, A. 2017. What is Social Media and What Questions Can Social Media Research Help Us Answer? *The SAGE Handbook of Social Media Research Methods***,** 13.

Meligy, A. M., Ibrahim, H. M. & Torky, M. F. 2017. Identity Verification Mechanism for Detecting Fake Profiles in Online Social Networks. *International Journal Computer Network and Information Security,* 1**,** 31-39.

Menardi, G. & Torelli, N. 2014. Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery***,** 1-31.

Patel, R., Bhagat, R., Modi, P. & Joshi, H. 2017. Privacy and Security Issues in Social Online Networks. National Conference on Latest Trends in Networking and Cyber Security (IJIRST), 2017. 130-134.

Peddinti, S. T., Ross, K. W. & Cappos, J. 2017. Mining Anonymity: Identifying Sensitive Accounts on Twitter. *International AAAI Conference on Web and Social Media*. Montreal, Canada.

Perez, S. 2011. Top 8 Web 2.0 security threats. *Read Write Enterprise. Retrieved March,* 9.

Powers, D. M. 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.

Pradhan, P., Misra, S. & Koirala, T. 2016. A Survey on Data Security in Social Networking Sites. *International Journal of Computer Applications,* 155.

Ribeiro, M. H., Calais, P. H., Santos, Y. A., Almeida, V. A. & Meira Jr, W. 2018. Characterizing and Detecting Hateful Users on Twitter. Twelfth International AAAI Conference on Web and Social Media, 2018 Palo Alto, California, USA. AAAI Press, 676-679.

Saito, T. & Rehmsmeier, M. 2017. Precrec: fast and accurate precision–recall and ROC curve calculations in R. *Bioinformatics,* 33**,** 145-147.

Samuels, J. 2018. Search and Reunification. *Adoption in the Digital Age.* Springer.

Sandy, C., Rusconi, P. & Li, S. 2017. Can Humans Detect the Authenticity of Social Media Accounts? *3rd IEEE International Conference on Cybernetics (CYBCONF).* Exeter, UK.

Sap 2017. SAP HANA.

Scott, S. & Matwin, S. Feature engineering for text classification. ICML, 1999. Citeseer, 379-388.

Sloan, L., Morgan, J., Burnap, P. & Williams, M. 2015. Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data. *PloS one,* 10**,** 1-20.

Thomas, K., Mccoy, D., Grier, C., Kolcz, A. & Paxson, V. 2013. Trafficking Fraudulent Accounts: The Role of the Underground Market in Twitter Spam and Abuse. USENIX Security, 2013. Citeseer, 195-210.

Trivedi, S. D., Kathad, C., Bhalodiya, T. & Pandya, T. 2016. ANALYTICAL STUDY OF CYBER THREATS IN SOCIAL NETWORKING. International Conference on Computer Science Networks and Information Technology, Jan 2016 2016 Pattaya.

Tsikerdekis, M. 2017. Identity Deception Prevention Using Common Contribution Network Data. *IEEE Transactions on Information Forensics and Security,* 12**,** 188-199.

Tuna, T., Akbas, E., Aksoy, A., Canbaz, M. A., Karabiyik, U., Gonen, B. & Aygun, R. 2016. User characterization for online social networks. *Social Network Analysis and Mining,* 6**,** 104-131.

Twitter. 2018. Twitter API. Available: https://dev.twitter.com/overview/api [Accessed 8 January 2018].

Van Der Walt, E. & Eloff, J. 2018. Are attributes on Social Media Platforms usable for assisting in the automatic detection of Identity Deception? International Symposium on Human Aspects of Information Security & Assurance (HAISA), 2018a Dundee, Scotland.

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

Van Der Walt, E. & Eloff, J. H. P. 2017. Creating an environment for detecting Identity Deception. *5th HPI Symposium on Operating the Cloud*. Potsdam, Germany.

Van Der Walt, E. & Eloff, J. H. P. 2018b. Using Machine Learning to Detect Fake Identities - Bots versus Humans. *IEEE Access*, 6**,** 6540 - 6549.

Varol, O., Ferrara, E., Davis, C. A., Menczer, F. & Flammini, A. 2017. Online human-bot interactions: Detection, estimation, and characterization. Eleventh International AAAI Conference on Web and Social Media, 2017 Montreal, Canada. 280-289.

Wang, G., Chen, H. & Atabakhsh, H. 2004. Criminal identity deception and deception detection in law enforcement. *Group Decision and Negotiation*, 13**,** 111-127.

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

Warner-Søderholm, G., Bertsch, A., Sawe, E., Lee, D., Wolfe, T., Meyer, J., Engel, J. & Fatilua, U. N. 2018. Who trusts social media? *Computers in Human Behavior*, 81**,** 303-315.

Willard, N. E. 2007. *Cyberbullying and cyberthreats: Responding to the challenge of online social aggression, threats, and distress*, Research Press.

Xiao, C., Freeman, D. M. & Hwa, T. 2015. Detecting clusters of fake accounts in online social networks. The 8th ACM Workshop on Artificial Intelligence and Security, 12 October 2015 2015 Denver, USA. ACM, 91-101.

Yamamoto, Y. 2018. Twitter4J. Available: http://twitter4j.org/en/ [Accessed 16 May 2018].