future internet

MDPI

Review

Vulnerabilities to Online Social Network Identity Deception Detection Research and Recommendations for Mitigation

Max Ismailov [1], Michail Tsikerdekis [1,*] and Sherali Zeadally [2]

[1] Computer Science Department, Western Washington University, Bellingham, WA 98225, USA; ismailm@wwu.edu

[2] College of Communication and Information, University of Kentucky, Lexington, KY 40506, USA; szeadally@uky.edu

* Correspondence: Michael.Tsikerdekis@wwu.edu; Tel.: +1-360-650-3968

Received: 29 July 2020; Accepted: 24 August 2020; Published: 31 August 2020

Abstract: Identity deception in online social networks is a pervasive problem. Ongoing research is developing methods for identity deception detection. However, the real-world efficacy of these methods is currently unknown because they have been evaluated largely through laboratory experiments. We present a review of representative state-of-the-art results on identity deception detection. Based on this analysis, we identify common methodological weaknesses for these approaches, and we propose recommendations that can increase their effectiveness for when they are applied in real-world environments.

Keywords: identity; deception; detection

1. Introduction

The spread of deceptive behaviors online has strongly reshaped the way that people interact. There is an arms race between more advanced deception methods and detection methods that researchers use to identify them as platform administrators are always trying to stay one step ahead of malicious users. Identity deception in particular has been the focus of several recent studies [1–4]. The problem often presents itself in online communities as fake accounts that are generated with ease by attackers that aim to cause disruption (e.g., fake content, Sybil attacks).

Over the years, several methods have presented highly efficient results in detecting identity deception. These include methods that detect malicious accounts using non-verbal behavior [5], clickstream sequences [6], genetic algorithms [7], content similarity [8], and user social behavior [3]. However, a noticeable reduction in fake accounts that we should expect from the implementation of these methods on online platforms has not yet been observed. Although a delay in the adoption of new technologies is not rare, the case is that most platforms have been attempting to address the issue of identity deception with little success. There is a disconnect between the academic literature presenting highly effective solutions and the industry having less than optimal results when it comes to practical implementations.

In this paper, we conduct an analysis of the key causes that could be responsible for the efficiency gap between lab results and real-world implementations by analyzing the research approaches on identity deception detection that have been reported in the literature recently. In Section 3, we present the current practices in identity deception detection research. Then, in Section 4, we identify shortcomings with assumptions and methods that could influence the validity of the results or otherwise lead to circumstances where "security theater" occurs. The term security theater refers

to the false sense of security that papers with high accuracy results may present when it comes to having their methods implemented in real-world social platforms. Finally, in Section 5, we provide recommendations that can address some of the shortcomings we have identified in this domain of research and improve the external validity of identity deception detection models. Our work focuses primarily on identity deception detection methods as they are applied in online social networks. We summarize the main research contributions of this work as follows:

- We highlight representative studies in the domain of identity deception detection on online social networks.
- We present the key mechanisms that are used to generate detection models and evaluate them.
- We identify the key shortcomings that may inhibit the effectiveness of these methods when implemented in real-world online social networks.
- We provide recommendations that can help address these shortcomings and improve the quality of research in this domain.

2. Identity Deception

Identity deception occurs when a malicious actor exploits the inherent assumption that other users have that the person behind an account is who they say they really are, typically on Online Social Networks (OSNs). Identity deception is used to achieve some goal such as financial gain, social influence, or destabilization of an online platform. In turn, these various goals will lead to various kinds of identity deception. For example, some attackers generate fake accounts, while others choose to conduct identity theft [9]. In order to create a more holistic picture of the scope of these practices, we further quantify the harm that they cause.

2.1. Societal Cost of Identity Deception

Since malicious actors engage in their activities on platforms that have a massive user base, an account's malicious activity could potentially reach a large number of people depending on the mode of dissemination. The degree of dissemination can be catalyzed depending on the type of behavior that is taking place. For example, if the malicious account is creating posts that spread misinformation, genuine users may retweet or share this post without realizing that it is spam. A post that contains a malware link and appears as a credible news source could be shared by one genuine user with all of their connections. As these behaviors are carried out on sites that involve substantial social interactions, the harmful impact of these activities is most directly exhibited in various aspects of our society and economy. This is a typical propagation method that malware uses and particularly on Twitter, wherein the method is effective even with a low probability for users to click on malicious links [10].

Since so much of our social interaction and information acquisition take place on social platforms, injecting distrust into this social ecosystem can create a situation where people have a fundamental suspicion of the information that they are receiving. Subverting the trust that an individual has in the information that is being presented to him/her can have far-reaching effects on an individual's relationship with true information and the factors of his/her decision-making process [11,12]. An example of tearing down individual trust can be seen in the 2016 U.S. Presidential Election that saw massive amounts of misleading information being spread by social bots [13,14]. In the months leading up to Election Day, a large quantity of social bots sprung up on Twitter and began repeatedly posting politically-polarizing tweets about candidates in the race. In turn, this had a strong impact on the relationship that people had with their sources of political information. Attackers can easily create new accounts and have the ability to flood popular channels of communication with a particular agenda, which has very clear disruptive effects on democratic discussions.

2.2. Economic Cost of Identity Deception

The maintenance of large-scale OSNs is not a trivial feat. A large workforce of administrators and content reviewers is required to combat the rise of identity deception. These administrators need to evaluate accounts that were flagged by automated detection mechanisms or reported by other users, which is all highly labor intensive for companies running and maintaining these platforms. For example, Facebook disabled 2.19 billion fake accounts in the first quarter of 2019 [15]. Additional costs are incurred by companies for the research and development into better identity deception security mechanisms, as well as dealing with legal challenges that may arise on their platforms due to the exploitation of social features on social platforms by attackers. Finally, companies that use social platforms for promoting products are also expected to incur costs ($1.3 billion a year) due to fake follower accounts [16].

3. State-of-the-Art Results on the Detection of Identity Deception

Since attackers performing identity deception do so in a variety of ways (e.g., identity theft, identity forgery), detection and mitigation mechanisms also vary. This in turn leads to a constantly evolving area of research due to the ever-changing nature of approaches that attackers use. Learning the behaviors of these malicious actors and developing novel mechanisms for detecting their behaviors is akin to building a house on mud. A researcher can spend much time and effort developing a model that exploits a particular pattern of behavior for these attackers and roll out a detection model in the real-world successfully. However, the attackers this model is trying to detect are typically very aware of the fact that their actions are under scrutiny and will modify their patterns of behavior in order to invalidate the classification model. As a result of this constant dynamic challenge, the techniques security practitioners employ to detect identity deception need to be adaptable and able to fit to the ever-changing behaviors of these malicious actors.

3.1. Review Analysis Method

In the analysis that follows, we identify representative papers from the literature that offer different approaches to identity deception detection. The list is intentionally not exhaustive because the scope of this paper is to highlight what we deem to be serious shortcomings for the state of research in identity deception detection.

The literature review was conducted by a faculty member and a graduate student at Western Washington University. The initial article search involved terms such as “identity deception detection” and “malicious account detection” using well-known databases such as IEEE Xplore, ACMDigital Library, and Google Scholar. We selected articles of interest based on their accuracy (i.e., such as the article falling within the scope of our investigation) and impact. The latter means that the article was either published in a venue of high impact, a reputable venue, and claimed to have substantially improved on past identity deception detection models. For example, reference [17] was both published in a reputable venue and has demonstrated a high malicious account detection accuracy of 99%. As such, the selection process led to representative papers with diverse approaches that have been or are likely to be cited frequently by other papers. In the end, we focused on and analyzed the methodological approaches described in 20 papers. We used threats to internal validity, external validity, and construct validity as the primary criteria for evaluating the 20 selected articles. These threats have been repeatedly identified as a serious source of concern for studies primarily in the medical field [18,19], but also in the field of cybersecurity [20]. Typical threats to external validity involve faulty sampling practices (e.g., not having a random or representative sample), inability to control online environmental conditions (e.g., different online social network designs have different features), and the interaction of history with a study (e.g., the implementation of a detection system can affect behavior and vice versa, and as such, data history becomes important). Additionally, construct validity, which relates to external validity, focuses on how operational definitions of variables are

being conceptualized. In other words, are the features selected for identity deception detection models representative of the definitions the researchers have set for them? A typical example where construct validity may be threatened is when a study uses proxy variables due to the inability to measure an effect accurately.

We proceed by discussing the techniques used, the criteria for evaluation used by these techniques, and the goals of these detection methods. Figure 1 depicts a high-level diagram for these items.

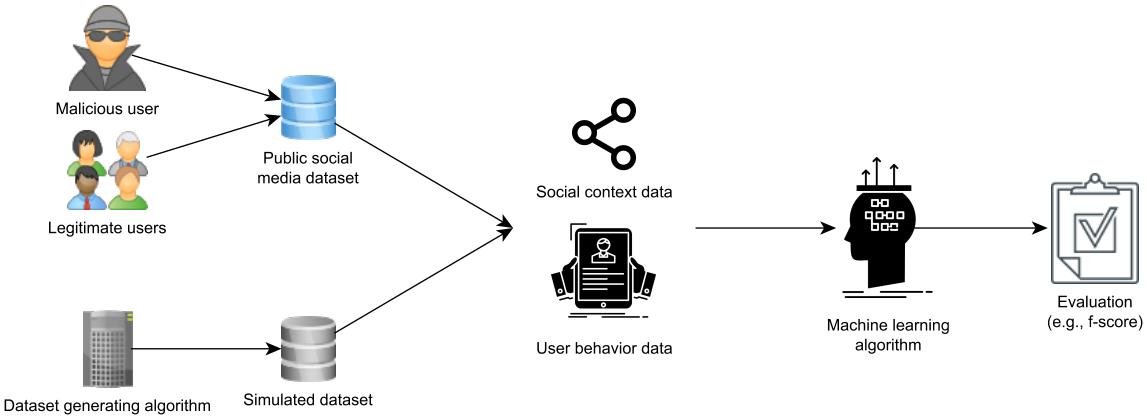

Figure 1. Typical identity deception detection experimental workflow. Data are selected from public sources or are simulated. Features are selected based on social and/or user behavior data. A model is constructed based on a machine learning algorithm, and it is further evaluated using metrics such as the f-score.

3.2. Detection Features

Techniques that identity deception detection models use depend on the social platform and the objective (e.g., detecting explicitly bots). For example, a malicious bot that is designed to make friends with as many accounts as possible will require a different detection approach than a bot that is designed to repeatedly post links to malicious websites. As such, our analysis identified two broad categories that we can divide models into: social context models and user behavior models. The two models accomplish the same task (i.e., detecting malicious accounts), but they use different sets of features. Many models also use some combination of both social features, as well as user behavior features. By features, we refer to the elements (or variables) in data that are used to differentiate between different data points (e.g., the number of posts can differentiate a malicious from a non-malicious user).

3.2.1. Social Context Models

Social context models achieve detection of malicious actors by examining features related to an account's social presence. This includes features such as relationships to other accounts, similarities to other users' behaviors, and a variety of graph-based features [2–4,21–23]. These types of features are valuable when classifying various types of malicious behaviors such as social bots and Sybil networks [24,25]. The approach allows for a high-level view of how users are interacting. There are relatively common patterns that benign users follow with regards to their social interactions on OSNs. They are typically the friends with an equal number of people who are friends with them (reciprocity of friendship), and the friends that they do have will likely be related to one another in a cluster. Users tend to further have common secondary interests (e.g., following the same categories of interest in a social network) [26]. On the other hand, a malicious account will typically not obey this sort of predictable behavior. For example, a spam bot will follow or send friend requests to a large number of unrelated users. It is through this juxtaposition of normal social behaviors against irregular social behaviors that allows models to determine malicious actors from a birds-eye view of the OSN on which the model is operating. Based on this focus on the greater social context of an account, graph-based

features are typically used in detection models. These include features such as the different numbers of connections that a user has, the degree to which they connect different clusters of users in the network (sometimes indicated by betweenness centrality), and other social network metrics that measure the level of a user's influence in the network [25,26].

3.2.2. User Behavior Models

User behavior models primarily focus on features that relate to an individual user's behavior, such as frequency of activities (e.g., number of tweets or posts per time interval), patterns of activity, and clickstream sequences [3,6–8,21,22,27]. This set of features can help identify malicious actors who may not interact with other users frequently, but still engage in unwanted online behaviors such as repeatedly posting spam or malware in the form of misleading URLs. With this perspective, a detection model can focus on features related to an individual's behavior rather than their behavior in a social context. These features are also essential to detect malicious users whose area of influence lies outside of the social interactions with other users. An effective example of this phenomenon is the predictability of a user's behavior as an indicator of being a malicious bot. A normal person who is using some social platform will likely have some degree of variation in the type, frequency, and duration of activities in which he/she engages. The sequence and the amount and time of activity would not be identical on a daily basis, but the same cannot be said for malicious bots. These bots are often designed to repeatedly perform a certain task to achieve a certain goal. They engage in predictable behaviors at relatively regular intervals. This predictability of malicious bot accounts provides a strong metric through which these accounts can be readily detected.

3.3. Detection Methods

Although heuristic methods exist for detecting malicious accounts, they are fairly rudimentary and can be easily bypassed by an experienced adversary (e.g., requiring an account to have at least one post). As such, all identity deception detection methods use some machine learning algorithm. These may use either social or user behavior features, and the effectiveness of these may vary depending on the choice of algorithm (e.g., regression, random forest, or neural networks). For example, clustering algorithms have been utilized in many studies that have used different features and approaches for pre-processing the data [2,6,21].

Supervised vs. unsupervised machine learning: Machine learning models are also used for anomaly and intrusion detection in order to solve some of the shortcomings of supervised machine learning models [28]. Supervised classifier models require large datasets that contain examples of each class that they are classifying (e.g., malicious or legitimate accounts). However, because the most advanced adversaries typically use novel techniques (i.e., behaviors), gathering this representative dataset becomes very difficult. The challenge with these types of models is that we expect to find deviations from normal behavior without having the necessary data to actually "know" what these irregularities are [29]. Furthermore, because of the high consequences of allowing attackers to remain undetected in OSNs, gathering data representative of the attackers becomes more challenging. As such, anomaly detection models (e.g., unsupervised models) are instead used. These are machine learning models that have been trained on normal data (i.e., legitimate user behavior). Any deviations from that behavior are classified as anomalies and probable malicious behaviors that need to be further examined by an administrator. As such, the bigger challenge associated with crafting these models is to have a homogeneous legitimate user behavior on the OSNs and identify features that can discern this behavior with high accuracy.

Computational overhead: The choice of machine learning algorithm will also influence the computational power needed for effective detection. Data, as well as the frequency of invoking and updating a machine learning model all result in computational needs for identity deception detection. Much of the research on new machine learning models has focused on leveraging domain knowledge in order to achieve better performance and low computational costs. However, a more recent school

of thought in machine learning posits that building models that are finely tuned to the domain often ignores the fact that computation over time becomes less expensive, and as such, it should be leveraged more [30].

3.4. Criteria for Evaluation Used by Models

Studies [17,26,31] use several metrics (e.g., precision and recall) to evaluate the performance of an identity deception detection model. These metrics attempt to represent how the model would perform in a real-world OSN. To evaluate the effectiveness of a detection approach, there are two essential components that are considered: the statistical measure (e.g., 80% recall), which reflects how well the model detected malicious accounts in some dataset of OSN traffic, and the dataset that is representative of real-world traffic, which affects the external validity of a model.

The predominant metrics that are used in the literature for identity deception detection relate to binary classification and are precision, recall, and F_1 score [32]. Precision is the ratio of true positives to the sum of true and false positives. It effectively measures the true positive rate. Recall, on the other hand, is the ratio of true positives to the sum of true positives and false negatives. It is thought to be an indicator of the degree of malicious accounts that are identified out of the complete sample. In practical terms, a high recall would mean that a model detects most malicious accounts, whereas a high precision would mean that moderators can trust the model to correctly identify accounts as malicious. The F_1 score is a weighted harmonic average of precision and recall. ROC curves are also frequently used to measure performance [33].

The second criterion for the evaluation of a model's performance is the dataset that the model uses and on which it is evaluated. This will influence the ability of the model to generalize well in other contexts [19]. Typically, researchers develop and subsequently test their model on numerous datasets of OSN traffic. These datasets usually contain data extracted from an OSN and contain primarily benign users alongside a small portion of malicious users (i.e., the low prevalence of malicious users will result in a low count of malicious account cases in a dataset). These datasets typically contain labels of what users or actions are considered "normal" or "not normal." Since these datasets are needed to confirm a model's performance, their quality and integrity are essential for the external validity of a model.

3.5. Detection Technique Objectives

Based on our survey of papers, we identified two main contexts for which models have been developed: either for industry or as part of academic research. This does not mean that there are not many studies that include both academic and industry professionals, but rather that the objective of a study is often influenced by the individuals who are involved in the work.

Studies conducted for industry use have models that have been developed in order to improve user experience while simultaneously minimizing work for OSN administrators [22]. As a consequence of trying to meet these goals, an industry-oriented model tends to have a low false positive rate (i.e., high precision), which may come at the expense of letting a few malicious users remain undetected (i.e., low recall).

Academic researcher models are being developed with a different perspective. Since the model's objective is not tied to the performance of any one OSN, the focus is more on the general soundness of the detection method, rather than having to frame the model in the context of a marketable product. As a result, studies from academia tend to focus more on making sure that their model can select all of the malicious actors who may be present in the dataset. This intuitively leads to a "sound" model, making sure that the detection mechanism is as correct as possible (i.e., high recall).

4. Research Approach Vulnerabilities

In our analysis, we identified several over-arching themes among the studies in identity deception detection. The state-of-the-art research that was mentioned in the previous section has methodology,

as well as assumption shortcomings that have predictable effects on the performance of models and their validity. These shortcomings include: relying on weak datasets, having data selection biases, selecting faulty features for use in the model, and overemphasizing precision rather than recall. We itemized these issues along corresponding examples of models that have these shortcomings. Table 1 presents a summary of these issues.

Table 1. Primary types of models, along with their common research vulnerabilities.

	Faulty Feature Selection	Weak Datasets	Precision Bias	Data Selection Bias
Social feature models	Select features that relate to the social context of an OSN (e.g., following relationships); the ground truth comes from a seed of pre-identified malicious users [2]	Datasets tend to contain easily separable clusters of users; users with an ambiguous social context are removed from the dataset [3]	Prioritize identifying components of the graph that are more homogenous rather than seeking out components that are more ambiguous [4]	Limit the size of a graph due to collection or computational limitations [2]
Atomic feature models	Selecting features relating to a user's online behavior (e.g., clickstream); the ground truth comes from a pre-existing "profile" for malicious behavior [6]	Datasets tend to contain manually selected users from real-world traffic, easily separable and free of much "noise" [8]	Tend to allow a few "hard to classify", but malicious users in order to minimize false positives [7]	Downsizing a dataset of real-world traffic in order to boost the model's performance with regards to "hard to classify" data points [3]

4.1. Weak Datasets

We identified several studies that used weak, outdated, or unrepresentative datasets for developing and testing their model. As discussed in the previous section, these datasets validate the model and become the ultimate frame of reference for understanding the model's performance. If a researcher evaluates his/her model based on an overly simplistic and unrepresentative dataset, the validity of the model becomes questionable when it comes to applying the identity deception detection approach in a real-world scenario. This is a common practice in identity deception research. One reason is because of the lack of recent or publicly available datasets that represent diverse OSN or network traffic. Another major contributing factor to the weakening of datasets relates to pre-processing due to the desire to have two distinct classes of "normal" and "malicious" users in the dataset. Every member of the two classes should display consistent behavior that is in line with what the model's approach to detection posits these classes as. Naturally, human behavior is non-deterministic, and trying to abstract the plethora of possible human activity on OSNs into two categories of "normal" and "irregular" behavior runs the risk of eliminating much nuanced contextual information from real-world data in order to condense the domain on which the model operates. A direct consequence of this categorization is the fact that users in the dataset who display an ambiguous set of features are removed from the dataset entirely, so that their ambiguity does not interfere with the "cleanliness" of the dataset. This decision to remove users with ambiguous social context (in the case of a social feature model [2]) or confusing individual behaviors (in the case of an atomic feature model [8]) is effective at removing the inherent "noise" to any dataset at the expense of weakening the dataset by removing real-world observations from it. Developing a model against an overly sanitized dataset can weaken its external validity. Furthermore, the limitation of available datasets can often lead to studies utilizing simulated data that are based on real-world observations and assumptions. The effect of these practices is similar to data sanitization [34] with the additional risk that the assumptions about user behavior that deception detection studies make to generate datasets can also influence both internal and external validity.

Finally, another dataset weakness we found in recent literature is the fact that all studies we referenced in this work have exclusively used a single OSN dataset; as such, the identity deception detection models may be highly customized for a specific OSN, and such models are unlikely to be able to generalize to other OSNs.

4.2. Data Selection Bias

One of the biggest factors of data selection bias is the reduction of real-world data in order to create a dataset that contains labeled ground truth about the individuals present in the dataset. To establish a ground truth, labels annotating the data points within the dataset are added. This labeling is an important part of creating an annotated dataset and can be made manually or by using automated tools. Even when these labels are assigned using some automated heuristic, they are further processed by a human annotator after the initial filtering. As such, biases can be introduced in this step. Datasets that are manually annotated by a human can introduce two issues: the filtering done by the human annotator very frequently selects the most obvious demonstrators of the features he/she is trying to categorize (e.g., selecting the most obvious fake account or the most obvious real account), as well as creating a very disproportionate ratio of real to fake users.

The emphasis on data points that clearly display the desired trends can result in studies that report an extremely high precision (e.g., 99%), because of the fact that the data points that comprise the dataset are all unambiguous displays of the behavior they are trying to predict. Furthermore, a disproportionate malicious-to-normal user ratio in a dataset can skew and misrepresent a model's performance. Real-world OSNs typically contain a significant number of normal users and a small number of malicious users. Since precision is the ratio of true positives to false positives selected, a high ratio of malicious to normal accounts in a dataset results in high precision and, to a lesser extent, a high recall. In other words, the dataset determines the performance of the model and not the other way around. This is because, with a larger ratio of malicious user to normal users, a model can make a few errors and still maintain a very high precision. This technique is done to effectively "pad out" the domain of malicious users that are trying to be detected in order to decrease the weight on precision that each potential misclassification might have. As a result, selection bias in a dataset may inevitably lead to the model under-performing in real-world scenarios.

4.3. Faulty Feature Selection and Construction

Studies tend to select or construct features that do not properly reflect the trends that they are trying to predict or are based on faulty assumptions.

A faulty assumption that is made for some features is misinterpreting time-dependent features as static. An example of this is using finite measures of activity (e.g., edits on Wikipedia [17] or number of Twitter followers [33]) without incorporating a measure of time for these metrics. This is often done because of a limitation in obtaining a proper dataset (e.g., we cannot obtain through Twitter's API the number of followers at time X after an account has been registered). As such, if current values for these features are obtained, the classifier model will learn to classify what can be observed at a specific point in time. Moreover, since a dataset includes legitimate and malicious users (some of which are banned or abandoned accounts), the time disparity for these metrics it is bound to be extreme. In other words, a legitimate user can have years of activity, whereas a malicious user may have just weeks of activity. This leads to an easier classification of users, but based on faulty assumptions. Ideally, most detection models should be able to detect identity deception after observing a specific amount of activity for an account (e.g., five days after registration). Without incorporating time in these features, this approach is not possible.

Spurious relationships are also likely to lead to features that perform well during experimental tests, but under-perform in real-world implementations. These are relationships that appear to exist between features and the predicted label (e.g., legitimate or malicious user), but in reality, an underlying third feature exists in between. In other words, a study fails to identify a key feature and instead identifies a proxy feature while assuming what has been observed is the true predictor of malicious account behavior. For example, a low count of followers that is used in studies as a feature (e.g., [33]) is not necessarily an indicator of a fake account's behavior, but it could be the result of account inactivity or administrative restrictions placed on new or suspicious accounts. Without this extra information (i.e., features), the follower count is just a proxy for predicting identity deception in this case.

4.4. Precision Bias

We already discussed that the metric of precision does not accurately provide a complete picture of how a model performs. Frequently, studies will highlight a high precision (e.g., 95% or higher) and will not equally pursue a high recall [4,7]. This is a phenomenon observed in industry led studies where a high precision can have a practical impact on releasing administrator time to address more intelligent adversaries that can evade systems. Placing a large value on precision also violates long-standing principles of security that tolerate high false positive rates as long as a larger emphasis is placed on eliminating false negatives.

5. Recommendations

Based on the aforementioned shortcomings that we have identified in the literature, we propose a few recommendations on future identity deception detection research.

5.1. Dataset Quality and Sharing

We recommend the development and release of high-quality real-world, as well as simulated datasets that are publicly available. Many of the shortcomings that we highlighted have been caused by the lack of publicly available datasets of realistic OSN activity. Many of the datasets being used in research are either highly outdated [23], comprised of simulated data [27], or exclusively contain clusters of users that are very easily separable [17]. On the other hand, real-world online platforms do not have an equal distribution of users and contain data points that are difficult to distinguish due to non-deterministic behavior by the users on these platforms. To address this issue, we would need a large-scale initiative to publicly distribute datasets comprised of real-world OSN that is minimally pre-processed as much as possible. Further, the frequency of the distribution of these datasets is also a relevant component. The datasets need to be frequently updated. Increasing the accessibility of these realistic, diverse datasets will not only improve the efficacy of identity deception detection models, but it will also make it easier to compare and contrast different approaches to identity deception detection. This way, we can further understand how the design and policy of OSNs influence user behavior and in effect malicious user behavior.

5.2. Emphasizing Recall

As mentioned before, increasing the emphasis on precision can lead to under-performing models and result in conditions of "security theater" in which the reported model performance does not accurately reflect the model's ability to detect the entire population of malicious users. As such, we recommend that studies prioritize achieving a high recall and make a concerted effort to report the recall values of newly developed models. We believe that although reporting recall may come at the cost of having an identity deception detection model appear less attractive, reporting this metric is necessary in order to develop models that have external validity (i.e., can perform outside a lab setting). For example, a study [31] demonstrated that an identity deception detection model can be fine-tuned to shift the emphasis between precision and recall, effectively finding a suitable medium between overemphasizing precision and overemphasizing recall, neither of which is optimal.

5.3. Realistic Feature Selection

Selecting features that are stronger indicators of malicious behavior will predictably shift the focus of a model toward the aspects of user data that are more relevant for detection. Since this phenomenon is intimately tied to the fact that weak datasets guide studies to using sanitized and filtered data points, improving the diversity and authenticity of publicly available datasets will inherently improve the features that researchers use to detect malicious users. However, even when the dataset quality increases, there are still some choices about feature selection that must be made in order to improve the soundness of model detection. Further, assumptions that are made about how features are being

derived should be realistic. Models that are designed to be introduced into real-time identity deception detection processes cannot use static data or otherwise data that have not been adjusted in terms of time. For example, aggregating Internet Protocol (IP)-related features over some sort of window of time is a meaningful step because these may change over time. This can also help to smooth out the noise inherent in the data [29]. Similarly, other assumptions need to be disclosed when features are selected. If accounts that are used in the dataset have already been banned and their complete history has been used, then the model will inevitably be able to distinguish banned accounts from accounts of legitimate users (as opposed to a malicious user who is still actively operating in the OSN).

5.4. Increasing Computational Overhead

There is a correlation between data and prediction accuracy. Overall, a large set of examples will produce better and more realistic performance results for identity deception detection models. As a result, this can potentially lead to an increase in computational overhead that is however bound to be alleviated over time as processing power increases and becomes less expensive. In particular, when using deep learning, the algorithms are often able to determine trends in the data that are completely imperceptible to human researchers because of the high dimensionality of the data. We further highlight that the risk of overfitting is not eliminated with any algorithm, and that includes deep learning models. A larger dataset that contains "weak" qualities will lead to an overfitted model that will under-perform in real-world scenarios.

6. Conclusions

This paper reviewed the current state-of-the-art results on identity deception detection research and highlighted a few key areas where researchers are making incorrect generalizations about the domain they are operating on, or do not have the appropriate data required to generate models that can generalize well. We highlighted several key issues that stem either from the lack of good datasets, faulty feature selection and construction, and inappropriate methodology. In order to address these issues, we provided several recommendations on future identity deception detection research. We recognize that this analysis is naturally subject to the authors' inherent biases, and the scope of the domain that was covered during our analysis. For this work to have an impact on mitigating research vulnerabilities in this domain, it needs to exist within the context of a greater discussion about how to improve practices in this field.

7. Acknowledgments

We thank the anonymous reviewers for their valuable comments, which helped us improve the content, organization, and presentation of this paper.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Tsikerdekis, M.; Zeadally, S. Online deception in social media. *Commun. ACM* **2014**, *57*, 72–80. doi:10.1145/2629612.
2. Yang, C.; Harkreader, R.; Zhang, J.; Shin, S.; Gu, G. Analyzing spammers' social networks for fun and profit. In Proceedings of the 21st International Conference on World Wide Web—WWW '12, Lyon, France, 16 April 2012; ACM Press: New York, NY, USA, 2012; pp. 71–80. doi:10.1145/2187836.2187847.
3. Zheng, X.; Zeng, Z.; Chen, Z.; Yu, Y.; Rong, C. Detecting spammers on social networks. *Neurocomputing* **2015**, *159*, 27–34. doi:10.1016/j.neucom.2015.02.047.
4. Wang, B.; Gong, N.Z.; Fu, H. GANG: Detecting Fraudulent Users in Online Social Networks via Guilt-by-Association on Directed Graphs. In Proceedings of the 2017 IEEE International Conference on Data Mining (ICDM), New Orleans, LA, USA, 18–21 November 2017; pp. 465–474. doi:10.1109/ICDM.2017.56.

5. Tsikerdekis, M.; Zeadally, S. Multiple account identity deception detection in social media using nonverbal behavior. *IEEE Trans. Inf. Forensics Secur.* **2014**, *9*, 1311–1321. doi:10.1109/TIFS.2014.2332820.
6. Shi, P.; Zhang, Z.; Choo, K.K.R. Detecting Malicious Social Bots Based on Clickstream Sequences. *IEEE Access* **2019**, *7*, 28855–28862. doi:10.1109/ACCESS.2019.2901864.
7. Cresci, S.; Petrocchi, M.; Spognardi, A.; Tognazzi, S. On the capability of evolved spambots to evade detection via genetic engineering. *Online Soc. Netw. Media* **2019**, *9*, 1–16. doi:10.1016/j.osnem.2018.10.005.
8. Concone, F.; Re, G.L.; Morana, M.; Ruocco, C. Twitter Spam Account Detection by Effective Labeling. In Proceedings of the ITASEC 2019, Pisa, Italy, 13–15 February 2019.
9. Tsikerdekis, M.; Zeadally, S. Detecting and Preventing Online Identity Deception in Social Networking Services. *IEEE Internet Comput.* **2015**, *19*, 41–49. doi:10.1109/MIC.2015.21.
10. Sanzgiri, A.; Joyce, J.; Upadhyaya, S. The Early (tweet-ing) Bird Spreads the Worm: An Assessment of Twitter for Malware Propagation. *Procedia Comput. Sci.* **2012**, *10*, 705–712. doi:10.1016/j.procs.2012.06.090.
11. Huber, B.; Barnidge, M.; Gil de Zúñiga, H.; Liu, J. Fostering public trust in science: The role of social media. *Public Underst. Sci.* **2019**, *28*, 759–777. doi:10.1177/0963662519869097.
12. Lovari, A. Spreading (Dis)Trust: Covid-19 Misinformation and Government Intervention in Italy. *Media Commun.* **2020**, *8*, 458–461. doi:10.17645/mac.v8i2.3219.
13. Allcott, H.; Gentzkow, M. *Social Media and Fake News in the 2016 Election*; National Bureau of Economic Research Working Paper Series; NBER: Cambridge, MA, USA, 2017; No. 23089. doi:10.3386/w23089.
14. Badawy, A.; Ferrara, E.; Lerman, K. Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign. In Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Barcelona, Spain, 28–31 August 2018; pp. 258–265. doi:10.1109/ASONAM.2018.8508646.
15. Lever, R. *Fake Facebook Accounts: The Never-Ending Battle Against Bots*; AFP: France, Paris, 2019.
16. Graham, M. Fake Followers in Influencer Marketing Will Cost Brands $1.3 Billion This Year, Report Says. 2019. Available online: https://www.cnbc.com/2019/07/24/fake-followers-in-influencer-marketing-will-cost-1point3-billion-in-2019.html (accessed on 30 August 2020).
17. Yamak, Z.; Saunier, J.; Vercouter, L. Detection of Multiple Identity Manipulation in Collaborative Projects. In Proceedings of the 25th International Conference Companion on World Wide Web—WWW '16 Companion, Montréal, QC, Canada, 11–15 April 2016; ACM Press: New York, New York, USA, 2016; pp. 955–960. doi:10.1145/2872518.2890586.
18. Ferguson, L. External Validity, Generalizability, and Knowledge Utilization. *J. Nurs. Scholarsh.* **2004**, *36*, 16–22. doi:10.1111/j.1547-5069.2004.04006.x.
19. Nicholson, W.K. Minimizing threats to external validity. In *Intervention Research: Designing, Conducting, Analyzing and Funding*; Melnyk, B.M., Morrison-Beedy, D., Eds.; Springer Publishing Company: New York, NY, USA, 2012; Chapter 7, pp. 107–120.
20. Sanders, C.; Smith, J. *Applied Network Security Monitoring: Collection, Detection, and Analysis*; Syngress: Waltham, MA, USA, 2014.
21. Cao, Q.; Yang, X.; Yu, J.; Palow, C. Uncovering Large Groups of Active Malicious Accounts in Online Social Networks. In Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security—CCS '14, Scottsdale, AZ, USA, 3–7 November 2014; ACM Press: New York, New York, USA, 2014; pp. 477–488. doi:10.1145/2660267.2660269.
22. Stein, T.; Chen, E.; Mangla, K. Facebook immune system. In Proceedings of the 4th Workshop on Social Network Systems—SNS '11, Salzburg, Austria, 10 April 2011; ACM Press: New York, New York, USA, 2011; Volume 8, pp. 1–8. doi:10.1145/1989656.1989664.
23. Daya, A.A.; Salahuddin, M.A.; Limam, N.; Boutaba, R. A Graph-Based Machine Learning Approach for Bot Detection. In Proceedings of the 2019 IFIP/IEEE Symposium on Integrated Network and Service Management (IM), Arlington, VA, USA, 8–12 April 2019; pp. 144–152.
24. Cao, Q.; Sirivianos, M.; Yang, X.; Pregueiro, T. Aiding the detection of fake accounts in large scale social online services. In Proceedings of the 9th NSDI'12 USENIX conference on Networked Systems Design and Implementation, San Jose, CA, USA, 25–27 April 2012; USENIX Association: Berkeley, CA, USA, 2012; p. 15.
25. Viswanath, B.; Post, A.; Gummadi, K.P.; Mislove, A. An analysis of social network-based Sybil defenses. In *ACM SIGCOMM Computer Communication Review*; ACM: New York, NY, USA, 2010; Volume 40, p. 363. doi:10.1145/1851275.1851226.

26. Tsikerdekis, M. Identity Deception Prevention Using Common Contribution Network Data. *IEEE Trans. Inf. Forensics Secur.* **2017**, *12*, 188–199. doi:10.1109/TIFS.2016.2607697.
27. Ruan, X.; Wu, Z.; Wang, H.; Jajodia, S. Profiling Online Social Behaviors for Compromised Account Detection. *IEEE Trans. Inf. Forensics Secur.* **2016**, *11*, 176–187. doi:10.1109/TIFS.2015.2482465.
28. Zeadally, S.; Tsikerdekis, M. Securing Internet of Things (IoT) with machine learning. *Int. J. Commun. Syst.* **2020**, *33*, e4169. doi:10.1002/dac.4169.
29. Sommer, R.; Paxson, V. Outside the Closed World: On Using Machine Learning for Network Intrusion Detection. In Proceedings of the 2010 IEEE Symposium on Security and Privacy, Berkeley/Oakland, CA, USA, 16–19 May 2010; pp. 305–316. doi:10.1109/SP.2010.25.
30. Sutton, R. The Bitter Lesson, 2019. Available online: http://www.incompleteideas.net/IncIdeas/BitterLesson.html (accessed on 30 August 2020).
31. Nazer, T.H.; Davis, M.; Karami, M.; Akoglu, L.; Koelle, D.; Liu, H. *Bot Detection: Will Focusing on Recall Cause Overall Performance Deterioration?*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 39–49. doi:10.1007/978-3-030-21741-9_5.
32. Hossin, M.; Sulaiman, M.N. A Review on Evaluation Metrics for Data Classification Evaluations. *Int. J. Data Min. Knowl. Manag. Process* **2015**, *5*, 1–11. doi:10.5121/ijdkp.2015.5201.
33. Cresci, S.; Di Pietro, R.; Petrocchi, M.; Spognardi, A.; Tesconi, M. Fame for sale: Efficient detection of fake Twitter followers. *Decis. Support Syst.* **2015**, *80*, 56–71. doi:10.1016/j.dss.2015.09.003.
34. Anagnostopoulos, I.; Zeadally, S.; Exposito, E. Handling big data: research challenges and future directions. *J. Supercomput.* **2016**, *72*, 1494–1516. doi:10.1007/s11227-016-1677-z.