Detecting Online Content Deception

Michail Tsikerdekis
Western Washington University

Sherali Zeadally
University of Kentucky

***Abstract*—The surge of content (such as fake news) in the last few years has made content deception an important area of research. We identify two main types of content deception based on either fake content or misleading content. We present a classification of deception attacks along with their delivery methods. We also discuss defense measures that can detect deception attacks. Finally, we highlight some outstanding challenges in the area of content deception.**

Web-based technological innovation for supporting social media applications has been fueled by the ever-increasing adoption of such services by users who are eager to utilize new Internet-based software. The availability and easy access to a wide range of social media tools have increased social media users' abilities to generate content as well as disseminate and increase the content's reach. In 2018, there were 3.2 billion social media users out of whom 2.4 billion were Facebook users. Many messages are publicly available or can be further forwarded to other social media platforms (e.g., Twitter). However, the ease of content generation and propagation has also opened up opportunities for malicious users who introduce fake or otherwise deceitful content in this "ecosystem."

The most recent documented case of a major content deception attack occurred in 2016 during the U.S. presidential election, and the surge of fake news in social media platforms has not stopped growing since then.[1] These fake news attacks are often well coordinated and involve a hybrid approach to online deception.[2] Hybrid approaches combine fake content, fake identities, and faking or tampering with communication channels (e.g., faking someone else's identity in order to deliver a fake message while hijacking a hashtag on twitter). Adversaries create new online accounts (often behind proxies or virtual private networks) and establish identities for these accounts using online data that are available for a demographic or geographic group

Digital Object Identifier 10.1109/MITP.2019.2961638
Date of current version 27 March 2020.

Published by the IEEE Computer Society

by utilizing profiling strategies. The aim is not only to make these identities appear as legitimate (to evade detection) but also to make such identities relatable to potential targets. The second step of such attacks involves the generation of fake content and its subsequent dissemination through and to potential victims. Sybil attacks (distributed fake accounts) may be utilized as an intermediate step to increase the status of the account utilized by the creator of fake content as well as those who assist with the dissemination of deceptive content.[3] Such information is not just restricted to fake news that is a subset of fake content. Since the Internet's early days, various research efforts have attempted to detect and mitigate the creation and dissemination of fake content. What is arguably a new phenomenon is the use of social software to expand the reach and, as such, the effect of content deception. Modern cases involve (besides fake news) product promotion (e.g., fake reviews), social engineering (e.g., password retrieval), and profile elevation among other attack vectors.[4,5] Content representation for such attacks uses both verbal and nonverbal forms.[6] The delivery methods of the fake content, as well as its effectiveness, varies, depending on the design of a social media platform and the respective interactions that are available to users (i.e., the digital environments affordances to its users).[2] Attacks from deceptive content can affect individuals, businesses, or even states making their impact broader and difficult to measure. In many ways, content deception can be seen as a form of online pollution .[5]

Much of the recent literature has identified some of the complexities of detecting fake content and have proposed defensive techniques that can protect against such deception attacks. However, there is no clear and universally accepted definition of what content deception is, how different attack vectors would be classified if there were a definition, and how subsequent defense measures would be implemented to take into consideration the wide range of factors among social media platform designs. Such structured analysis can help information technology professionals to better address the problem of content deception holistically and invest their efforts in mitigating the effects of content deception through high impact changes in their online platforms. For example, although some studies[7–9] provide a 1-D performance evaluation of detection methods, they still cannot sufficiently validate the effectiveness of these methods due to the variance in the design of social platforms. For example, if a fake content detection method utilizes the followers of a user as an indicator of deception for one platform, it is unclear whether this would hold true on another platform that applies stricter criteria to following others, or social net etiquette may not allow users to follow others as freely.

We summarize the main contributions of this article as follows.

- We explain how content deception is normally achieved, and we describe two categories of content deception that make use of fake content and misleading content.
- We classify common content deception attacks as well as delivery methods that attackers use.
- We discuss metrics that can be used for content deception and compare their relative efficiencies for different types of content deception.
- We discuss outstanding challenges in the field of content deception detection.

DEFINITION

To understand what content deception formally represents, we need to first establish what an ideal online communication looks like. For simplicity, we are using a single sender, although the process can involve multiple senders and receivers. A typical communication model has a message M submitted by sender S to a set of intended recipients R through a communication channel C. M is said to have been derived by a lossless function f such that $f(t) = M$, where t is an objective immutable truth with respect to a state of the world W. $f(t)$ transforms the truth in a communicable form and is said to be successful (lossless) if the "essence" of proposition t can be fully recovered from M. R is said to recover the message using a function d, such that $d(M, e) = t_d$, where e is the receiver's bias (prior belief) regarding the hypothetical nature

of the content before recovering the original message. If e is 0, then a perfect recovery occurs, which extracts the "essence" of t submitted by S. In communication, it is often the case that if the bias remains low, the derived t_d by R will be $t_d = t$ or $t_d \approx t$. It is worth noting that, even though for simplicity, we consider e to be a numeric representation of bias, and in different contexts, the data type of e may vary.

Content deception falls under the umbrella of deception, which is classified as a deliberate act to convey false information to one or more parties who are not aware that such an act is taking place.[2] As such, content deception is deliberate and requires at least two parties (an attacker and a victim). The unawareness requirement also allows for cases of self-deception for a victim. The goals and motivations for content deception that apply to social platforms include instrumental (e.g., fake reviews), relational (e.g., revolving around relationships and social capital), and identity based (e.g., fake profiles). Well-coordinated attacks may be primarily motivated by instrumental goals and rely on secondary relational and identity-based goals to ensure their success.

We define two distinct categories of content deception that we will examine in this article: fake content and misleading content. These relate to a sender's action as well as the bias of the victim, and they are formally defined in the following.

In the case of fake content, S produces t', which is an altered or fabricated version of t such that $t \neq t'$. Encoding occurs in a lossless manner $f(t') = M'$. In turn, R decodes the received message using $d(M,' e) = t'_d$ while assuming that $t = t'_d$, and thus, deception occurs. Deception occurs because the model of S states that $t \neq t'$ while the model of R states that $t = t'_d$. If t can be obtained by R and compared with the derived t'_d, then the conflict between the two models becomes apparent. For example, truth t: all elephants are grey contrasts with truth t': all elephants are red.

In contrast, misleading content is a much more cognitive demanding case of content deception for a sender. In this case, S uses t with a lossy function $g(t) = M'$, and bias e from R is such that $d(M,' e) = t'_d \neq t$. The success of this attack depends on S to correctly estimate the recipient's bias e and formulate a lossy function g accordingly. Misleading content is what colloquially one may refer to as a half-truth. For example, suppose a truth t states *50% of sexual violence victims are transsexuals*. $f(t)$ may result in M: *victims of sexual violence are largely trans*. However applying a lossy function $g(t)$ will result in M': *trans are largely involved in sexual violence*. The bias e of R toward the original "essence" of t can result in a decoding such that $d(M,' e) = t'_d$ states: *many sexual predators are trans*. However, arriving t is also a possibility if e remains "small."

Figure 1 illustrates the visual definitions as well as the differences between these two types of content deception attacks.

Additionally, deception attacks can fall into the categories of identity deception or communication channel deception. The two can often be used along with content deception as a form of a hybrid attack that can increase the effectiveness of an attack vector.[2] Examples of such attacks involve cases of plagiarism or faking a bank's website in order to obtain a potential victim's password. Both examples alter the identity of a sender (or point of origin) for a message M and truth t rather than the message M itself. This article focuses on methods of content deception and introduces types of hybrid attacks when necessary in order not to omit any necessary context.

CONTENT DECEPTION ATTACKS

Several types of attack vectors have been identified that include the two definitions mentioned above. Attacks can be further identified as targeted or nontargeted (e.g., whether an individual or set of individuals can be thought of as ultimate targets). A nontargeted attack can experience higher rates of collateral: victims who were not the original targets of attackers. Table 1 presents a summary of content deception attacks.

Deceptive Website

Deceptive websites have been, historically, the most common type of attack since the inception of the world wide web. The attack involves multilayered content deception where the aim is

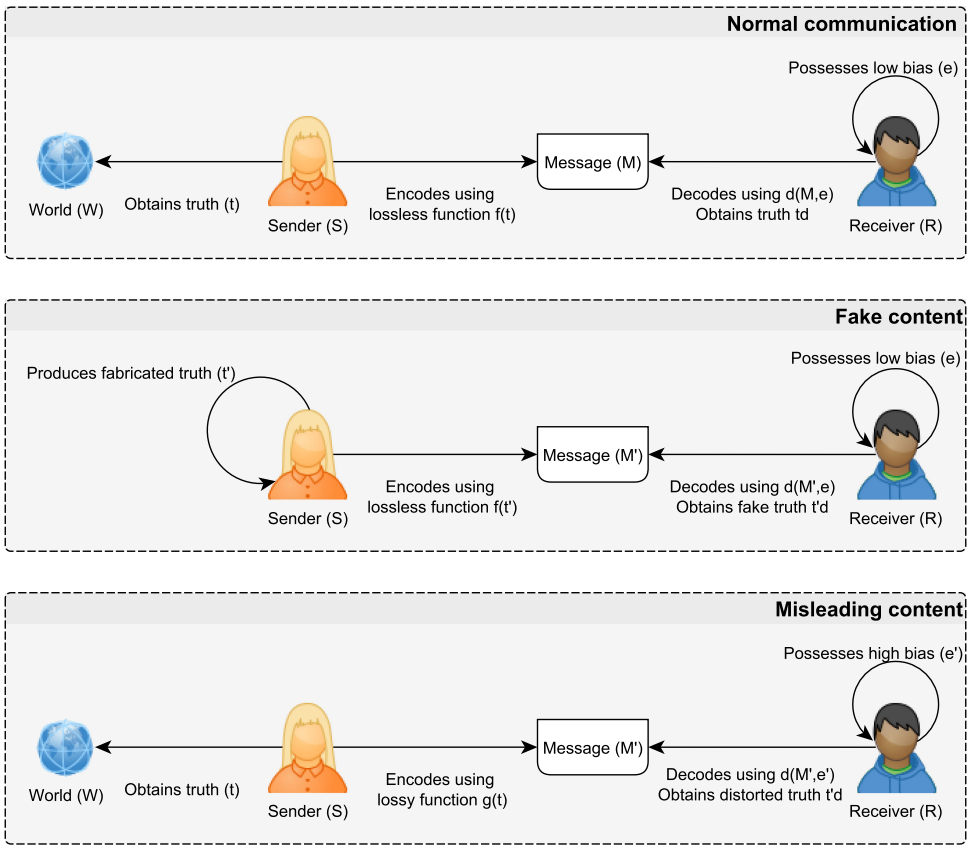

Figure 1. Visual depiction of normal communication as well as two types of content deception.

to facilitate the dissemination of fake or misleading content. In this context, multilayered comprises the various components that make up the website content. In a website, these layers can support its structure that can implicitly convey a deceptive message and various multimedia contents that can explicitly convey the deceptive message. In other words, the content can involve verbal (text, audio, or video) as well as nonverbal (e.g., images). The method is effective for attacks that aim to promote or damage an entity's (an object or individual) reputation. Automated methods that enable the automatic creation of fake websites[5,10] have also been observed. When website creation is automated, large databases containing genuine content are used to derive and alter content using techniques such as natural language processing and deep learning. *Web Spam*, the injection of artificially created web page into the web in order to

Table 1. Summary of content deception attacks.

Type	Formal definition	Examples	Delivery method
Deceptive website	M' is multilayered, e.g., content, structure, metadata, and domain name	Product promotion, password retrieval (social engineering)	Website creation, promote via social bookmarking
Deceptive metadata	Metadata is distant semantically with its associated content (M) relating to truth t	Profile elevation of a post by using popular keywords	Existing infrastructure (e.g., Twitter's hashtags)
Deceptive post	M' is single layered and self-contained	Fake news, fake reviews, paid posts	Existing infrastructure (e.g., Facebook's posts)
Deceptive personal message	M' is single layered and self-contained	Personal message phishing attack, e-mail phishing	Existing infrastructure or e-mail relays

influence results from search engines, is another use of such an attack.[11] The method is also popular as a "redirect" point for further phishing and other social engineering attacks. Delivery methods tend to be rather expensive because a website hosting infrastructure is needed, which also includes bandwidth allocation and domain acquisition, among other costs. Furthermore, website creation is often seen as the first step since promotion of content needs to occur through social bookmarking or other dissemination avenues (covert or noncovert).

Deceptive Metadata

Deceptive metadata aims to increase the profile of some content and has become popular due to its extensive use in Web 2.0 applications. Formally, the method utilizes at least one semantically distant metadatum (e.g., tag) in relation to message M from set of metadata $t = \{a, b, c, \ldots\}$, where $a, b, c, \ldots$ are examples of elements. The content is textual and machine readable. The method can be used in conjunction with other content deception attacks (e.g., deceptive website). An example of such an attack is the promotion of a paid product or service via a social media website that allows the use of tags for indexing. Misleading tags (e.g., "free") can elevate the profile of a post in order to reach more people within the bounds of a website's user network. The attack is considered cheap in terms of resources that are used because much of the infrastructure is already in place for an adversary to abuse.

Deceptive Post

A modern type of content deception attack in Web 2.0 involves the use of posts to deliver fake or misleading content.[4] Post refers to self-contained content both thematically and visually. It can contain both verbal as well as nonverbal content. Just like in the case of deceptive metadata, a deceptive post will use existing infrastructure as the delivery method, which makes it a low-cost type of attack. Examples of these attacks involve fake news or fake reviews, and the aim is to influence opinions much like deceptive website approaches. Since these attacks use the existing infrastructure, collateral can often be much higher because it becomes impossible to predict the victims of such an attack (apart from identifying the probability of dissemination through the complete connected component of a user's network). Delivery methods for these attacks can often be manual as well as automatic through the use of botnets or social bots.[12] There are two subsequent types of attacks based on the size of the content that we have further identified: deceptive microposts and deceptive macroposts. Deceptive microposts are found on microblogging platforms, where limitations on the size of a "message" apply.[13] For example, Twitter restricts messages to 140 characters or a single URL link to a video. In contrast, messages on Facebook can run as long as approximately 63 000 characters. The distinction is important because time can often be an important detection factor based on the cues left by a deceiver, and in the case of text, this often translates to quantity.[14,15]

Deceptive Personal Message

A similar type of attack to deceptive posts is deceptive personal messages. These posts also aim to deliver a self-contained single-layer content M', but often through targeted means. Examples of these are phishing e-mail attacks or phishing personal messages through forums or online social networks. Although collateral is also likely, senders usually have specific targets in mind, and propagation of a message through a user network is less efficient because many users do not attempt to forward these messages. Personal messages are also less restricted in terms of the size limitations and are often largely compared to microposts. Although an existing infrastructure can be used for personal messages, e-mails can also utilize a private server or e-mail relays to launch an attack. Depending on the approach, the method can be more expensive for a deceiver.

DETECTION METHODS

We have identified several methods that have either been directly tested on a specific type of content deception or that could be applied given their required parameters. At a high level, most approaches aim to detect fake content by either looking for signs of lossy "encoding" for a message M or by utilizing "inference databases" that can decipher whether a content has be misrepresented with respect to a truth t or an expected baseline (e.g., presentation quality). Additionally,

there are other methods that investigate the qualities of a sender's (or author's) message. We list them in the following and present their necessary requirements, performance benefits, as well as some of their drawbacks.

Deceptive Dictionary

Dictionaries are one of the most elementary techniques that can be used in detecting deceptive content. The technique requires a ground truth sample of deceptive content. The ground truth is used to derive an itemized list of elements that are indicative of deceptive content. The dictionary can then be used as baseline for fake content deception by means of statistical or machine learning approaches.

Dictionary size and quality can vary depending on the medium (text, audio, or video), as well as the domain. The simplest type of such a dictionary is a word list. In this case, a dictionary can be constructed based on several text analysis techniques such as Boolean (binary classification), term frequency (TF), or TF–inverse document frequency.[16] More advanced dictionaries involve utilizing parts of speech, deep syntax, or a variety of bigrams (or *n*-grams). For cases where audio is present, prosodic feature (e.g., voice pitch) dictionaries can also be utilized.[17] Similar approaches can be found in cases of fake pictures or video where an analysis for features of the image may be used to extract statistical qualities that will help build a proper baseline.

The approach is considered to be computationally efficient once a baseline is constructed because each post is checked against the baseline that has been derived from the dictionary. For some statistical and machine learning approaches, a response for a particular post may be constant ($O(1)$) once features are analyzed from a post. However, the analysis of the post itself, as well as the analysis of the ground truth that serves as a training dataset, can vary, depending on the algorithmic time complexity for extracting each metric. As a result of its simplicity, the approach has been found to be accurate at detecting social bookmarking site spam.[5]

A major drawback of this method is the need for a good ground truth (training data) that has been documented, which can be difficult to find.[16] Datasets can often include content that is thought to be fake, although that may not be really the case. Even more challenging is the case of misleading content detection where clear dictionary features may not be enough to help in the detection of fake content. This is because much of the work in a misleading attack focuses on the omission of information with respect to a truth t. Moreover, the subsequent dictionaries that may be developed by such techniques are often domain specific and may not translate well to other domains. For example, a dictionary may be accurate at detecting fake review language but inefficient in detecting fake news that is disseminated through personal messages. Essentially, detection using this method relies on observing only the message M, which is under the complete control of the sender giving the latter an asymmetric advantage.

Deceptive Content Cohesion

Deceptive content cohesion approaches aim to identify consistency within the content, as opposed to utilizing an external structure or content in order to infer deceptive intent. Cohesion means internal consistency between parts of content (or metadata associated with it) and the whole content. Such metrics (e.g., comparing a title with a body of text and any associated URL link) have been developed to look for deviations between such consistency, which could be indicative of deceptive content. The method aims to identify deceptive cues that are inadvertently leaked by lossy encoding functions for a message and, therefore, are more likely to reveal cues for misleading content rather than fake content. An example of a technique in this category is a measuring tag similarity, which is looking for semantic variance among tags.[5] If the dissimilarity between tag words is high, then we can often assume the content itself is also fake or misleading. The particular time complexity in this example is quadratic. However, given the limited number of tags or even metadata associated with content, the computational overhead in practice is minimal. The aforementioned method is less accurate compared to the use of deceptive vocabularies.[5] However, it is a cheaper method to implement given that there is no need for a ground truth and training data. That is, as long as a heuristic rule is established,

the approach can be applied to any content. The major limitation is that heuristics are likely to be domain specific. For example, phishing websites may differ substantially compared to phishing e-mails. However, there are some features that may be fairly predictable across domains (e.g., grammatical errors in professional texts are rare).

Deceptive Structure

Fingerprinting content and structure have been found to be an effective way for identifying deceptive content.[5,10] Although content dependent, the overall approach involves looking at the structural properties of content and establishing through statistical methods or heuristics what properties can be used as inferential points for deceptive content. This could involve features such as word length distribution or, in the case of images, frequency bins for pixel color ranges.[10] A particular use for such an approach can be found in cases where well-defined structures are identifiable (e.g., through hypertext). In cases of deceptive websites, one can look for the occurrence of HTML elements and establish cues indicative of deception. In one study, Markines *et al.*[5] obtained an area under curve of 0.86, and in another study, Abbasi and Chen[10] obtained a detection accuracy of 0.85. Given that these are results of binary classification, they are significantly high. Similar to the deceptive content cohesion approach, the aim is to detect structural elements that indicate the underlying motive (e.g., placements of many ads for generating revenue).[5] Depending on its implementation, the method has the potential to be more computationally efficient than approaches that use natural language processing techniques. The main limitation of the method is that it may not be applicable to some types of content (e.g., deceptive microposts).

Deceptive Account History

The detection methods we have presented so far look at only a message M and its associated metadata. A further step that can increase the detection accuracy focuses (whenever possible) on the history of the sender S associated with M and traverse through all M_n submitted by S. The subsequent evaluation may vary, depending on the type of content deception. For example, message history (M_n) cohesion detection where past messages are evaluated for consistency is one possible objective. Another approach is to look for the presence of past hyperlinks that are still valid in M_n.[5] For example, in microblogging applications, many users post deceptive links as part of a phishing attack. Messages containing older domain names that are invalid can raise red flags about future messages submitted by S.

The approach is more computationally expensive than other methods because we need to traverse and analyze historical data. However, for repeated offenders who still maintain active accounts (due to a nonviolation of the terms of agreement), this may be an effective way to identify deceptive content.

Deceptive Behavioral Indicators

One type of nonsemantic analysis aims to identify cues associated purely with the senders account S.[4] Methods under this category look at patterns associated with the account from which content originated in order to infer behavior. This category includes indicators such as number of posts to number of replies ratio, time delta between posts, active days, number of posts in which one comments, and geographic origin of S, among others.

The approach can often be cheaper than methods that need to dissect content because much of the structured user data is usually also available in databases that support online social media platforms. However, detection indicators can have a varying effect on the accuracy due to the ease for manipulating some of them. For example, if a certain number of posts is necessary in order for the account to appear active and exceed the threshold that would make it suspect, a user can just post more content; that is, often, unrelated content or even autogenerated content using bots. In fact, most of the example provided by indicators can be manipulated. Other studies have demonstrated that it is more difficult to manipulate some complex user behavior metrics such as the social network structure of a user[18] or the use of like to post ratio,[7] but these metrics come at the expense of substantial computational costs.

Deceptive Database

A final method, which is the most expensive in terms of the infrastructure needed, is the use

of databases in order to infer directly the truth associated with a message. This involves a semantic analysis of M and utilizing a trusted database containing various truths. We refer to the term database as a metaphor that may encompass a sociotechnical system of users and the information infrastructure that supports the process. Due to the ambiguity of language and the lack of well-structured data, the task often involves a human evaluating content directly against the database to reach a conclusion. This expert opinion process is often outsourced to third-party vendors who invest the effort and responsibility needed for such a process. A major example in this category is Facebook's effort to flag fake news based on this system. The process relies on the reports sent by several users and third-party fact-checking services. The method does not address cases of misleading content that contains truthful information in a lossy or misrepresented fashion. In such cases, there is subjectivity based on the experience level of the human evaluator.

FUTURE CHALLENGES AND RESEARCH OPPORTUNITIES

We have identified several challenges and opportunities that need to be addressed to perform content deception detection efficiently and accurately. These challenges include the absence of universal datasets, lack of universal benchmarks, and the need to establish informational metrics for a method's tolerance to reverse engineering.

Universal Datasets

There is a lack of access to datasets that are available to researchers involved in the area of content deception detection. As such, most studies result in a nomadic data collection process that is often never publicized or is publicized in isolated locations (e.g., personal website). This is in sharp contrast to other fields, such as machine learning or network security, where centralized dataset repositories exist and upon which algorithms and methods can be tested and contrasted against each other. In addition to the challenges of fake content detection, for most cases, data collection is the first step in what subsequently requires cleanup as well as coding analysis in order to establish ground truths (i.e., verified fake content or misleading content).

Universal Benchmarks

Another challenge relates to the lack of a unified testing solution for different methods used in fake content detection. Algorithmic analysis of many solutions is rarely informative on what resources may be necessary for some detection methods in practice. For example, some methods scan all user records but only do a "high-level" (as opposed to an in-depth) evaluation, whereas others utilize in-depth analyses to construct inference systems. In terms of time complexity, both of these approaches may appear to scale similarly, but in practice, the computational resources required (e.g., CPU utilization) are much higher for some methods than others. In addition, so far, much of the focus has been on data volume, whereas online platforms also need to deal with issues related to data velocity. For detection to be relevant and accurate, metrics that describe real-time applications are needed. Future work should focus on identifying novel methods that can better contextualize not only the accuracy but also the performance of detection solutions.

Metrics

Finally, innovative solutions for detecting deceptive online content need to incorporate countermeasures that will ensure that they cannot be easily influenced, bypassed, or reverse engineered by adversaries. Existing detection studies are extremely limited on this aspect. Many of the current solutions utilize statistical or machine learning models, but even rule-based (heuristic) methods have inherently limited protection against attacks from an adversary. For example, using a dictionary for detection of deceptive content can quickly be reversed engineered by an intelligent attacker, but the method also becomes outdated as the norms and language of the online community evolve. To a large extent, this issue stems from the fact that such methods are built with stationary environments in mind. However, this assumption is violated in online communities by legitimate users and, subsequently, adversaries because they are

considered intelligent and adaptive agents. As such, the content deception detection methods need to be adaptable and flexible from two perspectives: adapt against an adversary that aims to influence the defense model (e.g., attacks on tainting datasets[19]) and adapt against an adversary that attempts to reverse engineer and circumvent the detection method.[20]).

CONCLUSION

The detection of deceptive online content has been a challenge for researchers from many years. We lack solutions (especially automated ones) that can mitigate the ease that existing online infrastructures allow adversaries to engage in deceptive content creation and dissemination. In this article, we have provided a formal definition of what online content deception is. We have identified several content deception attacks, and we classified how several detection methods may be applied to these types of attacks, along with highlighting the main challenges that the domain faces. We believe that, given the multifaceted nature of the problem, we need a unified detection approach that incorporates detection solutions to address the problem from multiple perspectives and may eventually become feasible if some the challenges we have identified in this article are solved in the near future.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their valuable comments, which helped them to improve the content and presentation of this article.

REFERENCES

1. H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," Nat. Bureau Economic Res. Working Paper Series, vol. 23089, 2017. [Online]. Available: http://www.nber.org/papers/w23089 http://www.nber.org/papers/w23089.pdf
2. M. Tsikerdekis and S. Zeadally, "Online deception in social media," *Commun. ACM*, vol. 57, no. 9, pp. 72–80, Sep. 2014.
3. B. Viswanath, A. Post, K. P. Gummadi, and A. Mislove, "An analysis of social network-based Sybil defenses," in *Proc. ACM SIGCOMM Comput. Commun. Rev.*, vol. 40, no. 4, pp. 363–374, 2010.
4. C. Chen, K. Wu, V. Srinivasan, and X. Zhang, "Battling the internet water army: Detection of hidden paid posters," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, 2013, pp. 116–120.
5. B. Markines, C. Cattuto, and F. Menczer, "Social spam detection," in *Proc. 5th Int. Workshop Adversarial Inf. Retrieval Web*, 2009, pp. 41–48.
6. S. L. Humpherys, K. C. Moffitt, M. B. Burns, J. K. Burgoon, and W. F. Felix, "Identification of fraudulent financial statements using linguistic credibility analysis," *Decis. Support Syst.*, vol. 50, no. 3, pp. 585–594, Feb. 2011.
7. M. Tsikerdekis, T. Morse, C. Dean, and J. Ruffin, "A taxonomy of features for preventing identity deception in online communities and their estimated efficacy," *J. Inf. Secur. Appl.*, vol. 47, pp. 363–370, Aug. 2019.
8. M. A. Wani, N. Agarwal, S. Jabin, and S. Z. Hussain, "Analyzing real and fake users in Facebook network based on emotions," in *Proc. 11th Int. Conf. Commun. Syst. Netw.*, 2019, pp. 110–117.
9. J. Wang, W. Zhou, J. Li, Z. Yan, J. Han, and S. Hu, "An online sockpuppet detection method based on subgraph similarity matching," in *Proc. IEEE Int. Conf. Parallel Distrib. Process. Appl., Ubiquitous Comput. Commun., Big Data Cloud Comput., Social Comput. Netw., Sustain. Comput. Commun.*, 2018, pp. 391–398.
10. A. Abbasi and H. Chen, "A comparison of fraud cues and classification methods for fake escrow website detection," *Inf. Technol. Manage.*, vol. 10, no. 2/3, pp. 83–101, Sep. 2009.
11. A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly, "Detecting spam web pages through content analysis," in *Proc. 15th Int. Conf. World Wide Web*, 2006, pp. 83–92.
12. E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," *Commun. ACM*, vol. 59, no. 7, pp. 96–104, Jun. 2016.
13. S. Lohmann, M. Burch, H. Schmauder, and D. Weiskopf, "Visual analysis of microblog content using time-varying co-occurrence highlighting in tag clouds," in *Proc. Int. Working Conf. Adv. Visual Interfaces*, 2012, pp. 753–756.
14. H. S. Park, T. Levine, S. McCornack, K. Morrison, and M. Ferrara, "How people really detect lies," *Commun. Monographs*, vol. 69, no. 2, pp. 144–157, Jun. 2002.
15. L. Zhou, J. Burgoon, J. F. Nunamaker, and D. Twitchell, "Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications," *Group Decis. Negotiation*, vol. 13, no. 1, pp. 81–106, 2004.

16. A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, "Fake review detection: Classification and analysis of real and pseudo reviews," Dept. Comput. Sci., Univ. Illinois Chicago, Chicago, IL, USA, Tech. Rep. UIC-CS-2013-03, 2013.
17. S. Benus, F. Enos, J. Hirschberg, and E. Shriberg, "Pauses in deceptive speech," *Proc. ISCA 3rd Int. Conf. Speech Prosody*, 2006, vol. 18, pp. 2–5.
18. M. Tsikerdekis, "Identity deception prevention using common contribution network data," *IEEE Trans. Inf. Forensics Secur.*, vol. 12, no. 1, pp. 188–199, Jan. 2017.
19. L. Huang, A. D. Joseph, B. Nelson, B. I. P. Rubinstein, and J. D. Tygar, "Adversarial machine learning," in *Proc. ACM Workshop Secur. Artif. Intell.*, 2011, pp. 43–58.
20. D. Lowd and C. Meek, "Adversarial learning," in *Proc. 11th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2005, pp. 641–647.

Michail Tsikerdekis is currently an Assistant Professor with the Computer Science Department, Western Washington University, Bellingham, WA, USA. His research interests include deception, data mining, cybersecurity, and social computing. He received the Ph.D. degree in informatics from Masaryk University, Brno, Czechia. Contact him at michael.tsikerdekis@wwu.edu.

Sherali Zeadally is currently an Associate Professor with the College of Communication and Information, University of Kentucky, Lexington, KY, USA. His research interests include cybersecurity, privacy, Internet of Things, computer networks, and energy-efficient networking. He received his bachelor's degree in computer science from the University of Cambridge, U.K. He also received a doctoral degree in computer science from the University of Buckingham, U.K. Contact him at szeadally@uky.edu.