Is this the Era of Misinformation yet? Combining Social Bots and Fake News to Deceive the Masses

Patrick Wang
Institut Supérieur d'Électronique de Paris
Paris, France
patrick.wang@isep.fr

Rafael Angarita
Institut Supérieur d'Électronique de Paris
Paris, France
rafael.angarita@isep.fr

Ilaria Renna
Institut Supérieur d'Électronique de Paris
Paris, France
ilaria.renna@isep.fr

ABSTRACT

Social media is an amazing platform for enhancing public exposure. Anyone, even social bots, can reach out to a vast community and expose one's opinion. But what happens when *fake news* is (un)intentionally spread within a social media? This paper reviews techniques that can be used to fabricate fake news and depicts a scenario where social bots evolve in a fully semantic Web to infest social media with automatically generated deceptive information.

CCS CONCEPTS

• **Information systems** → **World Wide Web**; **Social networks**; *Internet communications tools*; • **Human-centered computing** → *Social content sharing*;

KEYWORDS

Social bots, Social media, Deceptive information, Semantic Web

1 INTRODUCTION

Social media can be defined as a group of Internet-based services built on top of the ideological and technological foundations of allowing users to create and exchange content [18]. There is no doubt about how social media impacts each and every aspect of our society. Some of these aspects include: social communication, the creation of friendships and communities, and even professional advertisement. Possibly one of the most disruptive technologies that have emerged in recent times, social media sometimes has a positive impact on society. For instance, Massive Open Online Courses (MOOCs) provide free first-rate education to anyone having an Internet connection. Social media also helps to increase job satisfaction [14], foster social relationships [8], and organize efforts to face the aftermath of disasters [22].

However, social media is also the scene of three types of abuses: excessive use, malicious use, and unexpected consequences. For

example, [21] signaled the risks of addiction to social media and "Facebook depression"; [31] showed the correlation between a user's time spent on social media and her anxiety; [21] warned about the dangers of cyberbullying, cyberstalking, or online harassment; and [4] recently revealed the case of a social activity tracker disclosing the locations of US military camps in war zones. Some scenarios can overlap multiple types of abuses. For example, *Tay*, a Twitter chatbot, became "an evil Hitler-loving, incestuous sex-promoting, Bush did 9/11-proclaiming robot" after a single day of existence [16]. Malicious users took advantage of Tay's learning algorithm to promote offensive ideas, and the scientists behind Tay were not prepared for such a turn of event. This example is somewhat comical, but social media abuses can have otherwise more dramatic fallouts.

In 2016, the world witnessed the storming of social media by *social bots* spreading fake news during the US Presidential election [3, 5, 27]. In a study presented in [5], researchers collected Twitter data over four weeks preceding the final ballot to estimate the magnitude of this phenomenon[1]. Their results showed that social bots were behind 15% of all accounts and produced roughly 19% of all tweets. These figures are worrisome and suggest the role these social bots had in twisting the public debate. Spreading like infectious diseases [19], deceptive information reaches us more and more often. Though techniques exist to prevent the spread of deceptive information, users – and especially the *influential* ones – must think critically when confronted with news on social media. But what would happen if social media were to get so contaminated by fake news that trustworthy information hardly reaches us anymore?

This paper envisions such a scenario. We first define deceptive information, then list existing technologies to fabricate it, and describe how information diffuses in social media. We then present social bots and how they can operate to fabricate and spread deceptive information. We conclude by proposing some actions which could refrain such a scenario from occurring.

2 DECEPTIVE INFORMATION

2.1 Fabrications, satire, large-scale hoaxes

The nature of social media allows any user to create and relay content with little third-party filtering or fact-checking. A consequence is that anyone can engage in amateur citizen journalism with the risk of (un)willingly participating in the dissemination of deceptive

---

[1]This represented approximately 2.8M accounts generating 20M tweets.

information. Indeed, a majority of adults get news on social media [10] and research has shown that people exposed to fake news find them to be accurate [17].

Journalistic deception corresponds to either the act of withholding true information or providing false information [7]. This phenomenon is not new: The Sun published in 1865 the *Great Moon Hoax*, a story about the discovery of life and civilization on the Moon. Serious fabrications, humorous fakes and large-scale hoaxes are the three categories of deceptive information listed in [26].

Serious fabrications concern the creation of deceptive information characterized by a relatively small outreach. Examples of distribution channels are the yellow press, tabloids, or online websites [3, 26]. Their contents are centralized in a unique location and are still mostly produced by human means, reducing the capacity of serious fabrications to spread rapidly.

Humorous fakes are conveyed online and oftentimes presented in a format reminiscent of mainstream journalism. A popular example is *The Onion*, which presents itself as *America's finest news source*. Although they may seem unsettling at first[2], we consider that humorous fakes can sometimes be flagged because of their heavily satirical nature (and sometimes popularity).

Finally, large-scale hoaxes are characterized by coordinated and complex fabrications that disseminated on multiple platforms so as to get relayed by mainstream or social media. As suggested by its designation, this type of deceptive information distinguishes itself from serious fabrications by its operation scale. Indeed, the objective is to quickly reach a large audience and to effectively disseminate false information. A recent case involving multiple channels is the Columbian Chemicals plant explosion hoax that propagated using text messages, Twitter, Facebook, and YouTube.

2.2 Creation of deceptive information

Digital information can take the shape of text, audio tracks, pictures, or videos. Lots of popular tools already exist to doctor pictures or videos (e.g., Photoshop or After Effects). In this section, we present state-of-the-art techniques that can be used to fabricate deceptive information and which are lesser known to the public.

Natural language generation is the research field interested in designing computer systems that produce meaningful texts in natural language on a specific topic. Such systems can generate texts using a variety of input sources such as numerical data, corpus, or taxonomies [25]. However, a difficulty still resides in generating longer texts, as they require more vocabulary and can present more grammatical or semantic flaws. As a result, generating fake news automatically could be feasible but the resulting text might present poorly constructed sentences, thus raising suspicion.

Voice transformation is a technique used to alter the voice of a person. In a survey [28], the author lists three types of applications: voice conversion, voice modification, and voice morphing. Voice conversion concerns the transformation of the voice of a source speaker to imitate the one of a target speaker. Voice modification and voice morphing are focused on altering the voice of one or

more source speakers without specifying any target. Voice conversion presents a risk of fabricating audio discourses, but several techniques can detect such frauds [2, 24].

Videos also can be fabricated. For example, in [29], authors demonstrated the ability to learn lip sync from audio and video footages[3]. Face2Face [30] is another technique which transposes in real-time facial expressions from an actor onto faces present in videos[4]. Recently, we also witnessed the emergence of *deepfakes* [6] which uses deep learning techniques to swap faces in videos and sparked a vivid interest in the pornographic scene. These techniques can open the door to fake speeches, degrading videos of celebrities, or even the alteration of historical footages.

2.3 Information diffusion in social networks

Understanding how information diffuses within a social network provides insights on how deceptive information spreads. Information diffusion analysis is concerned with three main objectives [13]: the modeling of information diffusion, the detection of popular topics, and the identification of influential users.

Out of the models presented, we focus on [20] which studied the diffusion of information in Twitter through the lens of internal and external influences. In particular, this study sought to determine if a tweet containing a URL resulted from network diffusion or factors external to the network. Three points in this article are worth mentioning. First, out of a month worth of Twitter data[5], only 71% of URL mentions were attributed to internal exposure (e.g., commenting, citing, or retweeting a tweet) and 29% were driven by external exposure (e.g., sharing an external link on Twitter). Second, the analysis of information diffusion on Twitter showed that some topics, and in particular Politics, were more prone to external influence. Third, information diffusion usually follows these steps: (1), the network is *exposed* to external information and gets *externally infected* whenever a node shares this information; (2), by sharing this external information, external infection causes a burst of internal exposure; (3), internal exposure can lead nodes to relay the information resulting in an internal infection; and (4), external and internal exposures continue to feed the contagion after the initial burst of infection.

In the light of this model, we can better understand how popular topics emerge and how influential users can reach a massive audience. In the case of deceptive information, malicious software agents can be used to generate an initial burst of exposure by relaying deceptive information or participating in discussions [5]. However, these actions still depend on the fact that deceptive information is generated by humans. In the following sections, we first describe the current limitations to the creation of social bots. Then, we examine the consequences of having social bots generating and spreading credible deceptive information.

3 NOT SO *AUTONOMOUS* SOCIAL BOTS

Social bots are software agents that cohabit with humans in the social media ecosystem. They can interact with other users via social media functionalities; for example, they can tweet, send emails, or

---

[2]We can also cite *Le Garofi*, a French satirical news website whose name is a pun to the mainstream journal *Le Figaro*. The graphical identity of the former seems strongly inspired by the latter.

[3]See this video clip made by [29]: https://youtu.be/9Yq67CjDqvw?t=5m41s.
[4]See this video clip made by [30]: https://youtu.be/ohmajJTcpNk.
[5]After data cleaning, this represented 18,186 different URLs being diffused on Twitter.

engage in conversations using natural language via instant messaging applications. With these functionalities, social bots (or human users) can disseminate deceptive information. However, a social bot can only do what its social media platform and human creator allow it to do. For example, a Twitter social bot can act just as any other human user (re)tweeting or sending direct messages. On the contrary, a Facebook social bot can be, by its creator's choice, prevented from interacting using instant messaging or group posts.

We have seen that social bots can diffuse deceptive information but they still rely on humans to produce meaningful fake news. Before becoming *fully* autonomous, social bots need to overcome three obstacles. The first issue concerns the automatic creation of social bots for current or new social media platforms. The second issue relates to automatic adaptation to comply with changes in the social media platforms. The third issue concerns the impossibility for social bots to detect and understand content that can be useful for their objectives. We argue that social bots will have a greater impact once these issues are resolved. We name such social bots *intelligent social bots*. Currently, *political* and *technical* barriers exist to prevent their emergence.

Personal accounts are the sole propriety of human users. The creation of personal accounts is hard to automate. It usually requires a human user to interact with Web-based interfaces and to confirm being human indeed. To enforce this rule, social media implement CAPTCHAs and/or email and SMS verifications that add even more complexity to the process of creating accounts. A social bot could create and use a personal account by simulating the use of a Web browser programmatically. However, it is a time consuming and error-prone process and still must be implemented by a human.

Social media platforms occasionally propose *social bots accounts* to give social bots the possibility to interact with users in a controlled way. These accounts are found under different names[6]: *apps*, *business pages*, *bots*, etc. In any case, the creation of such accounts also necessitates passing verification tests similar to the ones mentioned previously. Moreover, there always need to be a human responsible for a social bot account.

Political & legal barriers are imposed by social media platforms or governments to limit the reach of social bots. For example, the Terms of Service of a particular social media platform may forbid social bot accounts from interacting with personal accounts[7]. This is true with Facebook as it does not allow social bots to initiate a conversation with a human user. Another example is the 2016 BOTS Act[8] which prohibits the use of software agents to buy online tickets for concerts or sporting events. These political and legal barriers influence important aspects of the *technical barriers*. A first consequence of the enforcement such political and legal barriers is that social bots own a distinct type of account. A second consequence relates to how social bots can use social media platforms.

Using social media functionalities requires understanding the semantics of the social media APIs. For instance, consider the development of a social bot for reading emails using Gmail[9]. An engineer will have to create a Gmail account, register as a developer to get an API key, and implement a social bot coordinating the following

operations: get the list of new emails, extract their identifiers, get each email individually by its identifier, extract and process their contents, and mark each processed email as "read". It can be a trivial task for an engineer, but automating the development of these steps can be difficult.

These technical barriers still exist for two reasons. First, human developers are the ones to specify how social media APIs must be used. Second, social bots cannot fully understand the semantics of these APIs yet. As a consequence, humans are still responsible for creating social bots and defining their behaviors. But in a world where the Web is fully semantic, we argue that intelligent social bots can appear, develop themselves, and adapt to changes in their social media platforms. In the following section, we attempt to portray this world and describe how such intelligent social bots can massively spread misinformation during critical political events.

4 A POSSIBLE POLITICAL OVERTHROW?

In this section, we present the following hypothetical scenario: intelligent social bots massively spread deceptive information to destabilize a government in place and eventually overthrow it. We set the context to an era in which machines would be able to browse a fully semantic Web. We provide a step-by-step description of how this scenario could unfold. This description is also backed up by information diffusion theories and technological progress.

The first step relates to intelligent social bots scanning and searching the Web for information about the current political climate. In parallel, social bots can also learn to create deceptive information using techniques described in Section 2.2 and publish it on the Web. This step has a crucial objective that is to select the contents that will be used to disseminate fake news. In this context, implementing opinion mining or sentiment analysis algorithms would allow social bots to recognize and share pieces of information that endorse their claims.

The second step consists in social bots spreading deceptive information on social media. The objective is to create initial bursts of internal exposure in social media platforms by relaying fake news. In Figure 1, we illustrate the dynamics of the diffusion of deceptive information in Twitter according to the model presented in Section 2.3 [20]. The choice of Twitter serves as an illustration, but the same development of events could occur in other social media as long as social bots can publicize deceptive information on their platforms. In [5], we have seen that Twitter bots represented 14% of active users while generating approximately a fifth of all tweets during the 2016 US Presidential election. Considering the impacts social bots had in this event, these figures are already not negligible. Increasing either one of these two figures could give a boost to the initial burst of internal exposure, but not without side effects. On the one hand, increasing the number of malicious social bots could lead to social bots infecting each other. On the other hand, a suspiciously high publication frequency could give away the fact that the authors are not human.
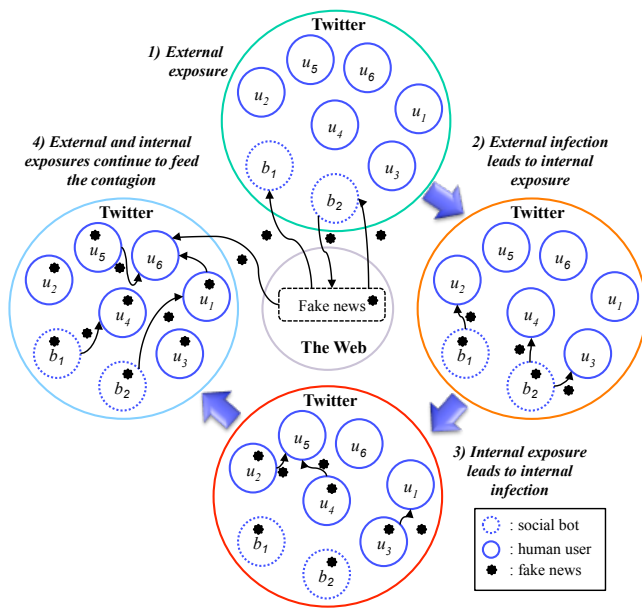
The objective of the third step is to infect as many human users as possible, i.e., to have them share or comment on a deceptive information created by an intelligent and malicious social bot. This step corresponds to enabling the internal influence mentioned in [20]. Note that here, the use of natural language processing techniques

---

[6]See, for example, https://developers.facebook.com.
[7]See Facebook terms of service: https://www.facebook.com/terms.php.
[8]BOTS Act: https://www.congress.gov/bill/114th-congress/senate-bill/3183
[9]https://developers.google.com/gmail/api/

Figure 1: Hypothetical scenario of Twitter social bots generating, fetching, and spreading deceptive information based on the dynamics of information diffusion described in [20].

can be leveraged to interact with human users in an attempt to plant even deeper the seed of deception. For example, social bots could implement opinion mining and sentiment analysis techniques [23] to locate users whose views are in line with the deceptive information being scattered. Moreover, social bots could evaluate, thanks to the social network APIs, the size of the audience of each infected human users. This should allow social bots to specifically target influential human users in the future.

After the initial burst, these three steps could repeat several times more with a lesser and lesser impact. Ultimately, a well-timed strike comparable to the one described in this section could have devastating consequences if the deceptive information conveys hate messages, for example. In a period when society seems more and more divided despite being hyper-connected, when nationalism experiences a renewed interest, the scenario depicted in this section looks quite dreary.

5 DISCUSSION AND CONCLUSION

The scenario depicted in the previous section is not completely hypothetical. Nowadays, we witness more and more reports of social bots interfering in political or societal debates [1, 11]. This situation could get worse with breakthroughs in Semantic Web or Natural Language Processing. In a society where all computing devices will be able to communicate with each other, the proliferation of both intelligent social bots and fake news could influence the world in ways we still cannot fathom.

Methods exist to detect social bots, or at least to discriminate between them and humans. In [9], these methods are divided into techniques based on (1) social network information, (2) crowd-sourced information, and (3) behavioral information. Techniques

based on social network information rely on community detection algorithms making the assumption that, within a community, social bots are more densely interconnected with each other than humans do. Such techniques were proved to perform poorly because of the assumption not being well grounded. Crowd-sourcing the detection of social bots can be both time-consuming and expensive, without even mentioning the lack of automation. Finally, techniques based on behavioral information attempt to apply supervised learning algorithms to distinguish humans from social bots. When performed on Twitter datasets, these algorithms showed that social bots possess recent accounts, tend to tweet a lot less and retweet a lot more than humans, or simply have longer usernames. This last set of techniques also presents a flaw as social bots can evolve to adopt a human behavior.

Alongside the detection of social bots, fact-checking is a key technique to assess the veracity and correctness of a claim. After all, fake news can be fabricated and spread by both humans and social bots. Given the scale of information diffusion in social media, manual fact-checking is a tedious task. The need for automated tools is imperative. ClaimBuster [15] is a fact-checking system that aims towards that goal. ClaimBuster comprises five elements: (1) a claim monitor that collects data from websites, social media, and broadcast media; (2) a claim spotter that identifies check-worthy factual sentences in a dataset; (3) a claim matcher that finds fact-checks of similar claims on other fact-checking websites such as Politifact; (4) a claim checker that gathers supporting or disproving evidence from the Web; and (5) a fact-check reporter that gathers all the information previously mentioned and displays it to the user. Despite showing promising results, there is still room for improvements in the design of ClaimBuster. Indeed, Hassan et al. noted some differences in the evaluations formulated by human checkers or the claim spotter. Reducing these differences would mean that there will be fewer errors when identifying check-worthy factual sentences.

To conclude, we insist on the importance of planning for both malicious use and unexpected consequences of social media. In [12], authors insisted on gathering all stakeholders to think about how to prevent such abuses, and qualified the "ignorance, willful or unintentional, [of an agent's eventual behavior to be] itself an ethical lapse, a lapse that is shared [amongst all stakeholders]".

REFERENCES

[1] David Alandete and Daniel Verdú. 2018. Russian interference: How Russian networks worked to boost the far right in Italy | In English | EL PAÍS. https://elpais.com/elpais/2018/03/01/inenglish/1519922107_909331.html?id_externo_rsoc=FB_CC. (2018). (Accessed on 03/02/2018).

[2] Federico Alegre, Asmaa Amehraye, and Nicholas Evans. 2013. Spoofing countermeasures to protect automatic speaker verification from voice conversion. In *Acoustics, Speech and Signal Processing, 2013 IEEE International Conf. on*. IEEE, 3068–3072.

[3] Hunt Allcott and Matthew Gentzkow. 2017. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives* (2017).

[4] Laignee Barron. 2018. U.S. Soldiers are Accidentally Revealing Sensitive Locations by Mapping Their Exercise Routes. Time. http://time.com/5122495/strava-heatmap-military-bases/. (2018). (Accessed on 01/31/2018).

[5] Alessandro Bessi and Emilio Ferrara. 2016. Social bots distort the 2016 US Presidential election online discussion. (2016).

[6] Samantha Cole. 2017. AI-Assisted Fake Porn Is Here and We're All Fucked - Motherboard. https://motherboard.vice.com/en_us/article/gydydm/gal-gadot-fake-ai-porn. (2017). (Accessed on 02/09/2018).

[7] Deni Elliott and Charles Culver. 1992. Defining and analyzing journalistic deception. *Journal of Mass Media Ethics* 7, 2 (1992), 69–84.

[8] Nicole B Ellison, Charles Steinfield, and Cliff Lampe. 2007. The benefits of Facebook friends: Social capital and college students use of online social network sites. *Journal of Computer-Mediated Communication* 12, 4 (2007), 1143–1168.

[9] Emilio Ferrara et al. 2016. The rise of social bots. *Commun. ACM* 59, 7 (2016), 96–104.

[10] Jeffrey Gottfried and Elisa Shearer. 2016. *News Use Across Social Medial Platforms 2016*. Pew Research Center.

[11] Erin Griffith. 2018. Pro-Gun Russian Bots Flood Twitter After Parkland Shooting | WIRED. https : / / www . wired . com / story / pro-gun-russian-bots-flood-twitter-after-parkland-shooting/. (2018). (Accessed on 03/02/2018).

[12] Frances S. Grodzinsky, Keith W. Miller, and Marty J. Wolf. 2008. The ethics of designing artificial agents. *Ethics and Information Technology* 10, 2 (2008), 115–121. https://doi.org/10.1007/s10676-008-9163-9

[13] Adrien Guille, Hakim Hacid, Cecile Favre, and Djamel A Zighed. 2013. Information diffusion in online social networks: A survey. *ACM Sigmod Record* 42, 2 (2013), 17–28.

[14] Brittany Hanna, Kerk F Kee, and Brett W Robertson. 2017. Positive impacts of social media at work: Job satisfaction, job calling, and Facebook use among co-workers. In *SHS Web of Conf.*, Vol. 33. EDP Sciences.

[15] Naeemul Hassan et al. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by ClaimBuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1803–1812.

[16] Helena Horton. 2015. Microsoft deletes 'teen girl' AI after it became a Hitler-loving sex robot within 24 hours. The Telegraph. http : / / www . telegraph . co . uk / technology / 2016 / 03 / 24 / microsofts-teen-girl-ai-turns-into-a-hitler-loving-sex-robot-wit/. (2015). Accessed: 2018-01-31.

[17] Ipsos. 2016. Ipsos/BuzzFeed Poll - Fake News. https://www.ipsos.com/en-us/news-polls/ipsosbuzzfeed-poll-fake-news. (2016). (Accessed on 02/06/2018).

[18] Andreas M Kaplan and Michael Haenlein. 2010. Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons* 53, 1 (2010), 59–68.

[19] Adam Kucharski. 2016. Post-truth: Study epidemiology of fake news. *Nature* 540, 7634 (2016), 525.

[20] Seth A Myers, Chenguang Zhu, and Jure Leskovec. 2012. Information diffusion and external influence in networks. In *Proceedings of the 18th ACM SIGKDD international conf. on Knowledge discovery and data mining*. ACM, 33–41.

[21] Gwenn O'Keeffe et al. 2011. The impact of social media on children, adolescents, and families. *Pediatrics* 127, 4 (2011), 800–804.

[22] Leysia Palen and Amanda L Hughes. 2018. Social Media in Disaster Communication. In *Handbook of Disaster Research*. Springer, 497–518.

[23] Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval* 2, 1–2 (2008), 1–135.

[24] Patrick Perrot, Guido Aversano, and Gérard Chollet. 2007. Voice disguise and automatic detection: review and perspectives. In *Progress in nonlinear speech processing*. Springer, 101–117.

[25] Ehud Reiter and Robert Dale. 2000. *Building natural language generation systems*. Cambridge university press.

[26] Victoria L Rubin, Yimin Chen, and Niall J Conroy. 2015. Deception detection for news: three types of fakes. *Proceedings of the Association for Information Science and Technology* 52, 1 (2015), 1–4.

[27] Chengcheng Shao et al. 2017. The spread of fake news by social bots. *arXiv preprint arXiv:1707.07592* (2017).

[28] Yannis Stylianou. 2009. Voice transformation: a survey. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conf. on*. IEEE, 3585–3588.

[29] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing Obama: learning lip sync from audio. *ACM Transactions on Graphics* 36, 4 (2017), 95.

[30] Justus Thies et al. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *Computer Vision and Pattern Recognition, 2016 IEEE Conf. on*. IEEE, 2387–2395.

[31] Anna Vannucci, Kaitlin M. Flannery, and Christine McCauley Ohannessian. 2017. Social media use and anxiety in emerging adults. *Journal of Affective Disorders* 207 (2017), 163–166.