New Dimensions of Information Warfare

Part I: Society

Roberto Di Pietro
Hamad Bin Khalifa University - CSE
rdipietro@hbku.edu.qa

Simone Raponi
Hamad Bin Khalifa University - CSE
sraponi@hbku.edu.qa

Maurantonio Caprolu
Hamad Bin Khalifa University - CSE
mcaprolu@hbku.edu.qa

Stefano Cresci
National Research Council - IIT
stefano.cresci@iit.cnr.it

Contents

List of Figures

List of Tables

Part I

Society

Since the last thirty years, humanity has been witnessing an incessant global metamorphosis of society. The development and the widespread diffusion of new, faster, more powerful communication technologies – such as social media – have brought profound changes to multiple aspects of social, economic, political, and cultural life, completely redefining the concepts of time, space, and identity. We are living in an age where much of the information and knowledge of mankind can be digitally reproduced in the form of text, image, music, or video. This progress, both in the technological and in the communication sphere, is radically changing the way of living, working, as well as producing and distributing goods and services.

Besides influencing interactions among people, this information society is forcing traditional organizational structures to become more participatory and decentralized. The opportunity to timely access information eventually translated into the simplification of many organizations' processes, thus leading to an increase in both their efficiency and the overall productivity. Contemporary society is characterized by a high dynamism that places information in a central position, making it a strategic resource that can heavily affect the systems' efficiency, becoming a determining factor of social and economic development, growth, and cultural wealth[1].

Definitions

Information Society. Society in which the creation, distribution, and manipulation of information has become the most significant economic and cultural activity.[a]

[a] `https://whatis.techtarget.com/definition/Information-Society` (Last checked August 2020)

Information can be seen as a new necessary capital, a valuable exchange commodity that can be accumulated – to enjoy more knowledge to be exploited, denied – to monopolize the information channels, or even imposed – to manipulate public opinion, thus becoming a new, powerful form of power.

[1] `http://www.treccani.it/enciclopedia/societa-dell-informazione_%28Enciclopedia-della-Scienza-e-della-Tecnica%29/` (Last checked August 2020)

1

Information Disorder

The rise of new technologies, including Online Social Network (OSN)s, media sharing services, online discussion boards, and online instant messaging applications, make information production and propagation increasingly fast. The number of Internet users, as well as the amount of available information, is continuously growing at an unprecedented pace. In 2014, 3,079 billion Internet users were populating the web. This number grew in the following years, reaching 4,157 billion Internet users in December 2017, an outstanding increase of 35% in three years, up to 4,648 billion Internet users in June 2020. The number of users populating the web over the years is depicted in Fig. 1.1. Accordingly, user-generated content created and shared by Web users grew proportionally. To give some figures of this massive phenomenon, the indexed Web nowadays contains more than 4.45 billion pages, representing only the tip of the iceberg. The *deep Web*, also called the hidden or invisible web, represents the part of the World Wide Web whose contents are not indexed by common Web search engines, and is estimated to be 500× the size of the indexed web, also known as *surface Web*.

People surfing the Web have an almost infinite amount of information at their fingertips. Some of this information is truthful, other is not. As a direct consequence, the ideas of individuals are no longer built autonomously (i.e., based on facts they know or elaborate), but they are increasingly based on hundreds of thousands of opinions read on the Web, of which only a negligible part appears to be authoritative. With these assumptions, shifting the attention of the masses and changing the opinions of people is a breeze – in the former case to make events of national importance go unnoticed, in the latter one, to set the agenda.

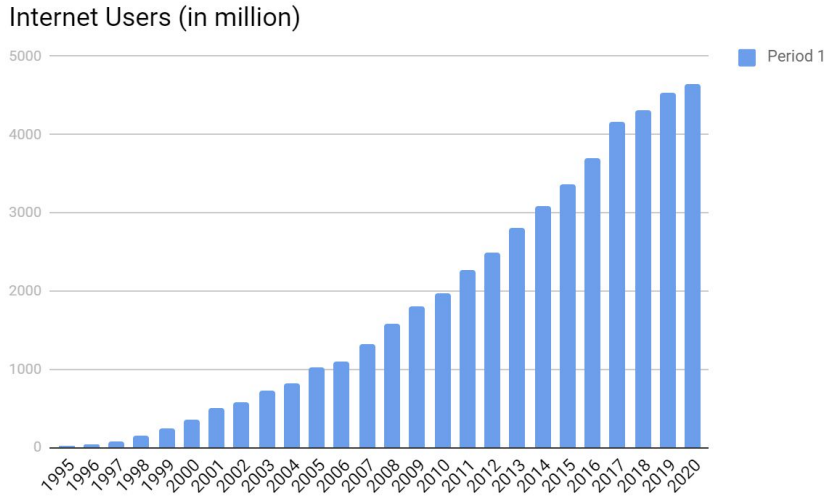

Figure 1.1: Users populating the web over the years

Definitions

Misinformation. Misinformation is defined as incorrect or misleading information. Regardless of any intention, misinformation is inaccurate, or incorrect information that causes people to be misinformed.

Disinformation. A subset of misinformation, it is defined as false information *deliberately* and *intentionally* spread to influence public opinion or mislead the truth.

Misinformation (or worse, disinformation) grows in step with information, making it difficult to distinguish authoritative information from misleading ones. To make matters worse, people do not use to double-check the content they incur in while surfing the Web, due to either lack of time or will, and sometimes they share it without having even read or understood it, participating in an unintended spread of unauthoritative information that uncontrollably spread through the Web. Furthermore, technology is automatically separating users from the information they disagree with, virtually isolating each user in its ideological and cultural bubble.

Definitions

Filter bubble. Tendency of social networks and recommendation algorithms to lock users into personalized feedback loops, each with its news sources, cultural touchstones, and political inclinations [1].
Confirmation bias. A pervasive cognitive bias resulting in human tendency to search for, favor, and recall information that supports (or confirms) one's prior personal values or beliefs.
Homophily. Tendency to form strong social connections (or to be attracted) to people with similar characteristics (e.g., ethnicity, gender, age, political leaning).
Echo chamber. Situation in which beliefs are reinforced (or amplified) by communication and repetition inside a closed system, while alternative ideas are not considered.
Polarization. Sharp division into contrasting groups or sets of beliefs.

While surfing the Web, users are overwhelmed by waves of personalized content that has been artfully-crafted based on their interests, their location, and their browsing history. This phenomenon tends to lower the critical spirit of individuals, placing them in front of a vision of a personalized world, absolutely not in conformity with reality. It is scientifically proven that the exposition to already known facts (or to facts people are already convinced of) exposes them to less cognitive effort if compared with the one generated when something new should be figured out. This tendency is known to the public as *confirmation bias*. Moreover, the homophily tendency explains how users are more likely to socially connect to people with similar thoughts and characteristics. With such a Web organization, people that share the same thoughts/tastes are grouped and are never exposed to ideas and contents of people who think differently from them. Similar people are closer and closer, while the "different" ones are virtually moved further and further away. This means that on every topic, instead of having confrontations with people who have different opinions, several communities with a few intersection points are formed, as if they were different clusters, i.e., the polarization concept.

In this chapter, we will first introduce the building blocks of this new social media ecosystem, analyzing what changes they are bringing to the current ecosystem. Then, two original scenarios are taken into account to describe how information warfare could pose a threat to the fabric of society of a country. Scenarios include running disinformation campaigns to cause unrest and discredit either people or governments, and piloting the elections in democratic states. Finally, we report current open issues and we propose some innovative countermeasures.

1.1 The New Social Ecosystem

The latest Web technologies, including OSNs, social media, and discussion forums, are eventually leading to a process of democratization of information. News, information, and statements, that years ago could only be produced by a few authoritative sources, can be nowadays produced by literally everyone, without any control or restriction. Indeed, while in the past sending a message to the mass was a pretty troublesome business, nowadays there is no need to go through authoritative sources: posting a tweet would be enough to climb over them. The producers of information themselves, publicly deprived of the role of information gatekeepers, are forced to compete against every individual to obtain public attention. If on the one hand, this *disintermediation* is allowing everyone to participate in public debates and have their voice heard, on the other hand, it is bringing to a steep decrease in both the quality and the truthfulness of the information itself.

Although these episodes have always occurred throughout history, the dynamics are drastically changing. Indeed, while in the past this phenomena could be easily dammed to prevent it from spreading, nowadays the technological catalyst is allowing it to spread on a large scale, without dams that hold. As a consequence, different entities are emerging, intending to exploit this amplified sounding board with malicious purposes. Among the most important actors, we can surely find social bots, cyborgs, trolls, spammers, and sock puppets.

Definitions

Social bot. Social bots are fully- or partially-automated accounts (i.e., pieces of software) that often act like humans on social media. Because of their automation, they are able to create content, share content, and interact with the other registered users [2].

Troll. A troll is a human-operated account who creates and shares offensive, provocative, or inconsistent posts and comments online.

Cyborg. A cyborg is defined as an account that is either a human-assisted bot or a bot-assisted human, inheriting characteristics from both.

Spammer. A spammer is a person (or an organization) that sends (or spams, in the jargon) irrelevant and unwanted messages indiscriminately to a plethora of users over the Internet.

Sock puppet. A sock puppet is a fake person employed to interact with other people, particularly in online discussions or blog comments sections.

When a social bot is used for political reasons (e.g., to artificially inflate the number of the followers of politicians or the popularity of a political

post, to spread propaganda, to subtly but effectively influence people political opinions, to target political opponents with a flood of enraged tweets) they are called political bots [3]. One significant difference between political bots and humans regards the pace of publication on social media. Indeed, even if compared with the most boisterous activists, bots do not need to sleep and have, as the sole purpose of life, the one assigned by the programmer. This makes a political bot an impressive tool when it comes to shaping public opinion, given its omnipresence in every targeted political post. The employment of political bots can be found in almost every recent political election worldwide [2], in many cases to advertise and give parties a more sonorous voice, in others, to spread fake news about the opponent and attempt to radically change the opinion of voters.

Cyborgs fall exactly in the middle between humans and bots and are becoming increasingly present in OSNs such as Twitter and Facebook [4, 5]. Examples of cyborgs are accounts that, by exploiting RSS feeds or widgets, automatically post content on behalf of the user. When the content published is inflammatory, provocative, or offensive, these actors take the name of trolls. The reasons behind the behavior of trolls are manifold: to get attention, to start discussions, to distract from legitimate discourses, to create unnecessary arguments, all of the reasons eventually leading to either the troll's amusement or to specific gains[1]. The online trolling phenomenon, already critical on its own, is lately getting enormous attention because of the alarming consequences it provokes. Indeed, there are several cases of teenagers who committed suicide after being victim, without any reason, of hate speeches online.

Definitions

Hate speech. A hate speech is defined as a violent and abusive speech intended to offend, insult, or intimidate an individual because of some traits, such as religion, race, origin, disability, or sexual orientation.

Other actors, in turn, have the role of spreading messages indiscriminately to the Web (i.e., spamming), with the aim of advertising products or spread malware [6]. The messages sent by spammers, often enclosed within e-mails, have different purposes, including the advertisement of a person/product/organization (i.e., unsolicited commercial email). The writer George Orwell, in his masterpiece "1984" defined the term "spam" as "pink meat pieces", giving this word the meaning "something disgusting but inevitable".

Furthermore, the Web is becoming increasingly populated by fake profiles whose purpose is to create disagreements and, in general, to provoke

[1] `https://techterms.com/definition/troll` (Last checked August 2020)

chaos. Sock puppets and meat puppets are proud representatives of this category of attackers. The term sock puppet comes as a reference to the manipulation of a hand puppet made from a sock, and it was originally coined to refer to the false identities assumed by Internet users who pretend to be someone else. Nowadays, sock puppets are employed for a variety of shady reasons, including (i) honoring, defending, supporting, attacking, insulting either people or organizations on the Web; (ii) manipulating the public opinion[2], and (iii), circumventing bans/suspensions from websites[3]. A meat puppet, in turn, by pretending to have experience with the subjects discussed on the Web, can direct conversations and manipulate the public opinion[4]. Meat puppets have been defined as "guns for hire able to be marshaled at a moment's notice", and the transition from sock puppets to meat puppets has been described as the elevation from subscribers with few fake personas to the business of invoking hundreds of automated fake followers at will [7].

Besides creating chaos and affecting online discussions, these entities are also employed to start targeted campaigns. Astroturfing campaigns allow increasing the credibility of statements and organizations without revealing the (id)entity of the supporters[5]. Political astroturfing refers to either politically-motivated individuals or micro-blogging platforms that make use of centrally-controlled accounts (i.e., bots, trolls, or sock puppets) to create the impression of widespread grassroots support for a candidate or opinion and to create widespread negativity for opposing opinions. When taken to the extreme, this mechanism allows to exaggerate the success and deny the failures of individuals or governments, or even worse, to portrait opponents or critics as "traitors" or national security risks[6] [8].

Definitions

Astroturfing. Astroturfing refers to the action of creating impressions of widespread support for a policy, individual, or product, where a little or (most of the time) none of such support really exists.

These actors, as soon as they find breeding ground (i.e., users that do not use to double-check content and tend to trust every information they find on the Web) can become extremely powerful tools to change the public opinion

[2] `https://guardianlv.com/2013/11/china-uses-an-army-of-sockpuppets-to-control-public-opinion-and-the-us-will-too/` (Last checked August 2020)

[3] `https://en.wikipedia.org/wiki/Sockpuppet(Internet)` (Last checked August 2020)

[4] `https://www.yourdictionary.com/meat-puppet` (Last checked August 2020)

[5] `https://en.wikipedia.org/wiki/Astroturfing` (Last checked August 2020)

[6] `https://www.thefridaytimes.com/political-astroturf/` (Last checked August 2020)

and to opportunely manipulate the thoughts of individuals. Unfortunately, the information they trust could belong to the sizeable fake news set.

Definitions

Fake news. Fake news is an ambiguous and informal umbrella term including hoaxes or disinformation purposely distributed via either traditional news media or online social media. It is used in scientific literature increasingly sporadically, in favor of more rigorous terms such as mis/disinformation, rumors, etc.

Hoax. A hoax is defined as a falsehood artfully fabricated to conceal the truth.

Rumor. A rumor is an unverified opinion, or a talk, that has been widely disseminated without a discernible source.

Propaganda. Expression of opinion or action by individuals or groups deliberately designed to influence the opinions or actions of other individuals or groups with reference to predetermined ends [9].

Although hoaxes, rumors, and disinformation have always been around, the sounding board has never been so loud. Indeed, online social media and Web services are contributing to create an ideal habitat for the ruthless sharing of information, allowing an immediate spread in every corner of the Web. There are different reasons behind the draft and the distribution of fake news, hoaxes, and rumors. Besides trying to convince citizens of the news itself, they may be used to manipulate the public opinion and, for example, to induce people to question democracies, states, or those entities in which we place our trust, that are at the basis of the functioning of a state, thus increasing the fragmentation and the polarization.

For the sake of simplicity, in the rest of the book, the umbrella term "low-quality information" will refer to mis/disinformation, hoaxes, and rumors, while every adversary will be called "malicious account", indistinctly.

1.2 Scenario 1: Freedom of Information

This scenario takes into account a State that does not make use of censorship techniques to silence the citizens. People are allowed to publicly express their opinion, in traditional ways as well as with modern means, such as OSNs, blogs, and online discussion forums.

The social platforms do not rely on traffic filtering techniques, that are usually applied to deny access to specific websites. Users are allowed to freely adopt any communication service, such as real-time messaging applications and mail services. Moreover, the government has no control over the content of the transmitted and received information, thus allowing users to express their opinion without being incurred in fines or punishments.

In this scenario, that can be applied without loss of generality to all, and not only, the western democratic countries, citizens are free to express both their thoughts and their opinion about any topic, whatever they are. The information spreading through social media can be of any kind: true or false, trusted or not trusted, accurate or inaccurate.

1.2.1 Threat: Disinformation Campaign

One of the first documented examples of alleged low-quality information takes us to Ancient Rome in July 64 b.C. The emperor Nero set fire to an entire district of the city to make room for new buildings, accusing the Christian community of the crime. He artfully created a piece of low-quality information to not turn the public opinion against himself, and to continue his persecution campaign against the Christian community. Going forward over the years, several other famous examples can be found. In 1933 the palace of the Reichstag – seat of the German parliament – was set on fire. The leaders of the Nazi party took advantage of this opportunity to blame the opponents of the Communist party, gaining consensus that led to their final rise to power. These two cases make it clear that the invention of news, as well as the alteration of them, make it possible to maneuver public opinion and, consequently, to obtain illicit advantages out of it.

Nowadays the same principle still applies, with the social media creating a sounding board that has never been so wide and loud. According to a famous study from researchers at Ohio State University, fake news probably played a significant role in depressing Hillary Clinton's support on United States' Election Day. This study, that provides a look at how fake news may have affected voter choices, suggests that about 4 percent of President Barack Obama's 2012 supporters were dissuaded from voting for Clinton in 2016 because of fake news stories[7] [10]. This problem was already clear some years ago when the World Economic Forum listed "the rapid spread of misinformation online" as one of its top 10 problems facing the world[8]. Some years later, in 2016 and 2017, the Oxford dictionary elected "Post-Truth" as the word of the year, and Collins Dictionary did the same for "Fake News", respectively [11].

The lack of truthfulness of information makes the recognition of truthful and false content hard for citizens, creating doubts and confusion among the population. Artfully built news is usually mixed with any size fragments of truth over time, escaping the control of the creator, who usually manages to

[7] https://www.washingtonpost.com/news/the-fix/wp/2018/04/03/a-new-study-suggests-fake-news-might-have-won-donald-trump-the-2016-election/?noredirect=on&utm_term=.d6e63f61fa06 (Last checked August 2020)

[8] https://reports.weforum.org/outlook-14/top-ten-trends-category-page/10-the-rapid-spread-of-misinformation-online/?doing_wp_cron=1583915074.7138180732727050781250 (Last checked August 2020)

govern the spread only for a short time. Then, this news assumes realistic contours, becoming in effect truthful news (as accepted by all as such), ignoring denials or not granting replication rights. Foreign governments, as well as terrorist groups and activists, may exploit these uncertainties on social media to undertake several kinds of disinformation campaigns, to undermine the credibility of the state or to control public opinions, with the aim of generating chaos and destabilizing the population. In the course of history there have been numerous cases in which the use of disinformation campaigns has caused discontent among the population, disagreements, and revolts, giving the history a presumed truth, impossible to ascertain.

1.2.2 Attacks

Throughout history, low-quality information played a determining role in the decision-making process of states and governments. This section analyzes four misinformation campaigns that have had effects on the population first, to then have heavy repercussions on the countries. The campaigns are related to (i) vaccines, (ii) immigration, (iii) conspiracy theories (e.g., flat earth), and (iv) climate change, respectively.

It is worth to notice that several countermeasures from the psychological, social, and cultural viewpoints can be employed to mitigate the aforementioned campaigns (e.g., media literacy [12]), but we consider them out of scope for this book. Instead, technological countermeasures are evaluated in Section 1.4

Vaccine hesitancy

"Even though it is a vaccine-preventable disease, measles kills over 100,000 people every year. Worldwide cases tripled in the first three months of 2019. The causes of these outbreaks are diverse: from health infrastructure to civil strife or vaccine hesitancy. In some countries, vaccine-skepticism and populism are increasing together" [9]

Lola García-Ajofrín

Low-quality information has had a huge impact on the health system and, as a consequence, has heavily affected the health-related choices citizens make. According to the World Health Organization, this phenomenon threatens the progress made throughout history in tackling vaccine-preventable disease, enough to be identified as one of the top ten global health threats of 2019[10].

[9] `https://outride.rs/en/vaccines-fake-news/` (Last checked August 2020)

[10] `https://www.who.int/news-room/feature-stories/ten-threats-to-global-health-in-2019` (Last checked August 2020)

Definitions

Vaccine hesitancy. Also known as Anti-Vax, it has been defined by the World Health Organization as "the reluctance or refusal to vaccinate despite the availability of vaccines".

To further highlight the benefits of being vaccinated, the United States Department of Health & Human Services listed the five main reasons to vaccinate a child: (i) immunizations can save the life, protecting against diseases that vaccines have been eliminating partially or completely, such as polio; (ii) safety and effectiveness of vaccines, the benefits of the disease-prevention of vaccines are greater than any possible side effect (redness, pain, or tenderness at the site of injection); (iii) immunization protects others, preventing the spread of diseases; (iv) immunization can save time and money; and (v) immunization protects future generations[11]. Even just looking at the numbers, vaccination is still one of the most cost-effective ways to avoid diseases, by preventing more than 2-3 million deaths per year. Furthermore, according to the World Health Organization, a further 1.5 million could be avoided if the global coverage of vaccinations improved[12].

However, despite the obvious benefits and negligible consequences, whole communities are arising to protest against the immunization and to try to prevent people from doing it. There are several reasons behind the anti-vaccinationists choice, ranging from conspiracy theories to concerns about safety. In the following, the main myths are reported[13].

Autism. Doubts about a possible relationship between autism and vaccines started as a result of the publication of a scientific paper by Andrew Wakefield, dated back 1998, defined in [13] as "the most damaging medical hoax of the last 100 years". Indeed, the scientific community, after extensive investigations, proved this theory to be false, not having found any relationship or causal mechanism between vaccines and the incidence of autism. However, the anti-vaccination activists continue promoting myths and conspiracy theories about the risk of autism, with misinformation acting as the glue between the two.

Vaccine overload. Vaccine overload, i.e., the theory that inoculating a large number of vaccines could negatively impact (e.g., damage or weaken) the immune system of children, is another myth embraced by anti-vaccination

[11] `https://www.vaccines.gov/getting/for_parents/five_reasons` (Last checked August 2020)

[12] `https://www.who.int/news-room/facts-in-pictures/detail/immunization` (Last checked August 2020)

[13] `https://en.wikipedia.org/wiki/Vaccine_hesitancy` (Last checked August 2020)

activists. According to several scientific studies, this belief is not based on solid foundations. Indeed, the improvements in the design of the vaccines of the last years have strongly reduced their immunologic load. Despite this factual evidence, vaccine overload remains one of the crucial points on which anti-vaccinationists base their campaign.

Ingredient concerns. In 2005, Rolling Stone and Salon magazines co-published an article by Robert F. Kennedy Jr., an environmental lawyer nephew of former President John F. Kennedy, alleging a government conspiracy. According to the article, the government was trying to cover up evidence that Thiomersal, or specifically, the mercury contained in Thiomersal, may cause brain problems, including autism. Thiomersal is an antifungal component used in some vaccines to prevent their contamination. Endless scientific evidence excludes the fact, showing that there are no common clinical symptoms and that the phenomena are definitely uncorrelated. Another ingredient that has been considered dangerous by vaccine-hesitant people is the Aluminum, used as immunologic adjuvants in many vaccines. Even in this case, several scientific studies revealed that there is no evidence of serious health risks or changes to immune systems.

Sudden infant death syndrome. Sudden Infant Death Syndrome (SIDS) is the sudden and unexplained death of a child during the first year of age. Unfortunately, such a syndrome leaves no traces, thus making autopsies fruitless. This led the anti-vax activists to think that vaccination might be a determining cause. In 1999, the ABC news program 20/20 broadcasted a story claiming that a vaccine (specifically, the hepatitis B one) was one of the causes of SIDS. The story was enriched with a picture of a 1-month-girl girl who experienced SIDS only 16 hours after receiving the second dose of hepatitis B vaccine. However, not only did several studies find no evidence of correlation, but also they found evidence that vaccination may protect children again SIDS.

Muslims. After several deaths among vaccinated workers, some militant groups and Islamists started seeing vaccination as a plot to either kill or sterilize Muslims. In 2003, Imams in Nigeria (i.e., Muslim leaders who lead prayers in Mosques) warned their followers and advised them not to proceed with the vaccination against polio for their children, perceived to be a plot to decrease Muslim fertility. This initiative caused a steep increase in polio cases both in Nigeria and in the neighboring countries that, in turn, stopped immunizing their children even against other diseases. As a consequence, in 2005 Nigeria reported over 20,000 cases of measles, counting 600 deaths, and in 2006 accounted for over half of all new polio cases worldwide.

Other. Other myths include: (i) vaccination during illness, many parents

are worried of vaccinating their children if they are sick; (ii) natural infection, people think that natural infection will provide their children with better immune protection against the future illnesses if compared to vaccination; (iii) vaccine schedule, people do not agree with the schedule recommended by the Advisory Committee on Immunization Practices (ACIP) that is built to protect children when they are most vulnerable. Unfortunately, a delay in vaccinating the children will uselessly expose them to illnesses.

These myths, fueled by misinformation found on the Web and onsocial media, led several parents to delay (or worse, to not require) the vaccination for their children. Clearly, this behavior had several consequences. To make an example, the World Health Organization no longer considers measles to be eradicated in the United Kingdom, as well as in the Czech Republic, Greece, and Albania. Measles is a highly contagious and potentially deadly disease caused by the measles virus. Measles killed approximately 2.6 million people per year until 1963, when the first measles vaccine was introduced. From that moment on, the number of measles cases steadily declined, leading to 21 million lives saved until the end of the century. However, the widespread anxieties about vaccine safety led to a dangerous return of the disease. Only taking into account Europe, there were 4240 cases of measles in 2016, 21,315 in 2017, and 82,596 in 2018. Things got worse in 2019 where 104,248 measles cases have been identified. Figure 1.2 summarizes the trend of measles cases identified in the last 5 years in the European region, provided by the Regional Office for Europe of the World Health Organization[14].

Immigration

"I'd like to live in a world where immigration is just called moving"

Stefan Molyneux

Although immigration has always been of crucial importance throughout history, the race for unbridled sharing of information in recent years is giving it a new and dangerous face. The presence of wars and the geopolitical scenarios led to an increasingly growing economic disparity between industrialized and developing countries. This disparity is only one of the reasons that brought people to leave their hometown and search for a better quality of life somewhere else. Among the other reasons, we may find the escape from persecution and violence, extreme poverty, and natural disasters.

In the following, some of the most widespread low-quality information about immigration are reported:

[14] `https://www.euro.who.int/en/health-topics/disease-prevention/vaccines-and-immunization/publications/surveillance-and-data/who-epidata` (Last checked August 2020)

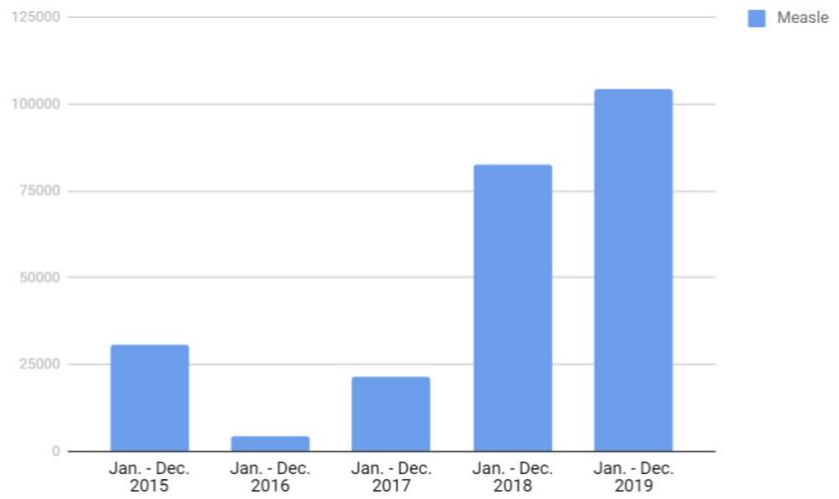

Figure 1.2: A 5-year summary of the epidemiological data on measles disease in the European Region.

- **Migrants moving for economic reasons do not need protection.** As mentioned above, the reasons that led individuals to leave their hometown are manifold. Some think that migrants moving for economic reasons do not need any kind of protection. However, despite the reasons that led people to move, immigrants face countless difficulties during the travel: from sexual abuse (where the victims are mostly women and child) to child labor, from detention to war enrollment, not to mention the atmospheric conditions. Being aware of those risks does not change migrants' minds, having few (sometimes no) hopes in the departure country.

- **Europe is the favorite destination of migrants.** People in Europe, partly motivated by the weaponized propaganda, think that all of the migrants want to cross the Mediterranean to seek asylum in European countries. This belief does not absolutely go along with the statistics. Indeed, in 2018, more than 90% of African migrants never left their region, preferring developing countries within their borders. Even if they are facing difficulties, migrants, if allowed to choose, prefer places where they can find a similar culture.

- **Do they have a smartphone? They surely do not need help.** It is common for a newscast to broadcast images and videos of migrants piled on a boat, but intent on typing on a smartphone. The first thought of an individual would surely be "does he have a smartphone? Clearly, he does not need any protection". Even in this case,

the reality is different. It is indeed crucial for a migrant to have a smartphone, that is both a way to communicate with her/his family, living thousands of kilometers away, and a source of information. Information is a synonym of protection when it comes to immigration, as it allows to guide the migrants through less risky routes and to be aware of the laws of other countries to understand their rights. It is worth noting that most of the time, to afford a smartphone, the migrant has probably traded some food, clothes, or accommodation.

- **All the migrants are criminals.** Throughout history, migrants have been portrayed as criminals from media and news. Let's suppose there is a crime in the city. A possible algorithm to make immigrants look bad is the following: if the person to commit the crime is a migrant the nationality will be specified on the news, otherwise, it will be omitted. In the medium (and long) term, viewers and listeners will be well impressed with the fact that crimes in the city have been always carried out by migrants that, in turn, will begin to be held up as dangerous and "intruding". This process may be used for different purposes: redirect hatred, manipulate the public opinion, establish fear in people who will then be easier to control.

- **There are too many immigrants, they will end up taking the power.** For the same reasons as before, media and news may spread inaccurate information about the number of immigrants within a Country. Several political parties all over the world have taken advantage of this citizens' fear to organize ad-hoc election campaigns, thus inciting hatred, misinformation, and the fear toward the "different". Facebook, Twitter, and other social media platform have incredibly increased the resonance of the rumors, thus paving the way for the spread of low-quality information such as "Illegal immigrants live in five-starred hotels entirely paid by Italian taxpayers" [14]. Such information is predictably false, but nonetheless may lead voters to have a better view of a political party that is less inclined to accept migrants within the country.

Flat earth society

The disintermediation made possible by social media allowed everyone to participate in online debates. If on the one hand this gives everyone the chance to speak their minds thus enjoying the freedom of expression, on the other hand, the information authoritativeness is undergoing a considerable decrease. In the past, without online technologies, conspiracy theorists (or people who had different opinions with respect to those already affirmed), were forced to organize physical meetings or rallies to meet. These limitations did not allow them to have a direct impact on the population. As

technology progressed, those rallies became groups on social networks, and meetings turned into discussion groups that millions of people can read and actively take part in. This facilitated the creation and advertisement of several "sects" which, without the echo provided by technologies, remained only local realities. An example is given by modern Flat Earth Societies.

Definitions

Modern flat earth societies. Organizations that dispute the Earth's sphericity by promoting the belief that the Earth is flat.

People have been knowing that the Earth is round at least since the sixth century. However, the first doubts in people have been instilled by Samuel Rowbotham, an English writer, that published a pamphlet called "Zetetic Astronomy" based on the Bedford Level experiment. This experiment, carried out in the summer of 1838, consisted of measuring the position of the flag of a boat slowly descending a river for 9.7km. The scientist reported that the flag constantly remained in his view for the whole journey. This made the scientist claim that the Earth is flat since, if it were not so, the top of the mast should have been below his line of sight. After Samuel Rowbotham's death, Lady Elizabeth Blount founded a Universal Zetetic Society intending to propagate the Natural Cosmogony knowledge in confirmation of the Holy Scriptures (according to [15], "Bible, alongside our senses, supported the idea that the earth was flat and immovable and this essential truth should not be set aside for a system based solely on human conjecture"), based on practical scientific-based investigations. After the Zetetic Society, several related associations came to light (e.g., Flat Earth Research Society) with a common goal: convince people (scientifically or not) of the flatness of the Earth.

The Flat Earth Society lived a severe bankrupt in the nineties and was losing, one after the other, all of its supporters, but then the Internet appeared, transforming the group into an online global entity. Indeed, while it took 50 years for the Flat Earth Society to reach 3,500 members in the pre-Internet era, nowadays their website gets 300,000 unique visits every day[15]. At the time of writing, the Flat Earth Society had gathered 89K followers on Twitter, 209K likes and 224K followers on Facebook. Among the most famous supporters of the flat Earth theory there are the basketball players Shaquille O'Neal[16] and Kyrie Irving, the football players Geno Smith and

[15] `https://medium.com/s/world-wide-wtf/how-the-internet-made-us-believe-in-a-flat-earth-2e42c3206223` (Last checked August 2020)

[16] `https://www.forbes.com/sites/trevornace/2017/03/28/shaq-thinks-earth-is-flat-because-it-doesnt-go-up-and-down-when-he-drives/#4ab8f4187233` (Last checked August 2020)

Bills Watkins, Homer (i.e., the author of the Iliad and the Odyssey), Hesiod, and Herodotus of Halicarnassus. In the Internet era, social technologies such as Facebook, Twitter, and YouTube have given activists, journalists, and people in general, a new way to connect and exchange theories, stories, and ideas, no matter how illogical they are. YouTube, in particular, has been accused of allowing the spread of misinformation through its platform. Indeed the American video-sharing platform in 2019 announced changes to its recommendation algorithm, to mitigate the spread of conspiracy theories[17].

Among the extravagant theories, Flat Earth Society claims that NASA, together with other government agencies, conspires to make the humanity believe the Earth is spherical. They claim that NASA uses to edit (i.e., photoshop) the images got from its satellites. Evidence of this, according to the aforementioned society, is found in the change of the color of the ocean observed in different pictures and the position of the continents. Furthermore, they think NASA is similar to Disneyland and that cosmonauts are actors, thus taking for granted that the moon landing never happened.

Climate change

Although the terms "Global Warming" and "Climate Change" are sometimes used interchangeably, they refer to slightly different concepts. One victim of the similarity of these terms was the current president of the United States of America Donald Trump who, according to a study conducted in 2018 [16] where his tweets have been analyzed, confuses the terms weather, global warming, and climate change. To clarify, we give the definitions of both Global Warming and Climate Change, provided by the National Aeronautics and Space Administration (NASA) agency[18]. In both cases, the changes are due to the increased levels of atmospheric carbon dioxide produced through the intensive use of fossil fuels.

Definitions

Global warming. Global Warming is a long-term shift in global (or regional) climate patterns. The term is also used to define the rise in global temperatures from the mid-20th century to the present.

Climate change. Climate change is an umbrella term that encompasses global warming but, specifically, it refers to the range of changes that are happening to Earth. These changes include sea level rise, melting glaciers, and shifts in flower/plant blooming times.

[17] `https://edition.cnn.com/2019/01/25/tech/youtube-conspiracy-video-recommendations/index.html` (Last checked August 2020)

[18] `https://climate.nasa.gov/faq/12/whats-the-difference-between-climate-change-and-global-warming/` (Last checked August 2020)

At first glance, it may seem that, given its objectivity, the problem, together with the underlying causes, is universally recognized. However, if on the one hand the scientific community reached a unanimous consensus on the reality of the human-caused climate change, on the other hand, the general public is becoming increasingly polarized on the issue. Indeed, the scientific consensus is continuously questioned by ideologically-motivated groups, such as the Merchants of Doubt. This group, supported and incited by others, have organized influential disinformation campaigns in which they publicly dispute the scientific consensus on several issues, including the human-caused climate changes [17], thus increasing the polarization and limiting the societal engagement with the issue.

The U.S. President Donald Trump, besides confusing the meanings of the terms as mentioned above, is not new to episodes of skepticism towards the concepts of climate change and global warming. Indeed, he defined climate change as "mythical", "nonexistent", and "an expensive hoax" and shared tweets like "The concept of global warming was created by and for the Chinese in order to make U.S. manufacturing non-competitive", "It's freezing in New York – where the hell is global warming", "I don't believe it", and "The badly flawed Paris Climate Agreement protects the polluters, hurts Americans, and cost a fortune. NOT ON MY WATCH"[19]. Some of the aforementioned tweets and many others are depicted in Figure 1.3. Given the number of supporters, these statements may be dangerous and convey the wrong message, leading most Americans to underestimate and scoff at an established problem.

In the opposite extreme, the young Swedish environmental activist Greta Thunberg entered the debate. The seventeen-year-old girl has been internationally recognized for her determination to make humanity aware of the existential crisis the climate is facing. The straightforward and unambiguous speeches of Greta aimed at world (political) leaders are reverberating all over the world, in which she criticizes their failure to take action to address the climate change emergency. However, even against the figure of the activist criticisms and judgments arose. Among the most popular there is the question "Who is behind Greta Thunberg?". Indeed, conspiracy theorists think that Greta is only a convincing facade of a very complex marketing organization, whose goals are far from beneficial.

1.3 Scenario 2: Democratic Election in a Country

In this scenario, we will take into account the political election of a democratic country. During this election, the state promotes transparency and fairness, providing the candidates with a fair media space and controlling their advertising according to principles of equality and correctness. Accord-

[19] `https://www.bbc.com/news/world-us-canada-51213003` (Last checked August 2020)

Figure 1.3: Some of the tweets written by Donald Trump about the climate change and the global warming.

ing to the definition of Jeane Kirkpatrick, the scholar and former United States ambassador to the United Nations, "Democratic elections are not merely symbolic. They are competitive, periodic, inclusive, definitive elections in which the chief decision-makers in a government are selected by citizens who enjoy broad freedom to criticize the government, to publish their criticism and to present alternatives"[20].

Democratic elections are competitive, since the mere right to participate in the ballot is not enough. Indeed political, and other, groups involved in the elections must guarantee fairness, avoiding censorship, partisan media, and respecting the rules. Both opposition parties and candidates must enjoy the freedom of speech, as well as bringing alternative policies and candidates to the voters. Democratic elections are also definitive, since they determine the leadership of the government. The party leader will have the burden of leading the country, possibly respecting the political program presented during the election campaign.

1.3.1 Threat: Interference in Political Elections

Political elections within a country are not only reflected in the interests of citizens. Companies and institutions (local or foreign) may have an interest in illegally interfering with the electoral campaign, with the aim of piloting its results and obtaining profits in the short, medium, or long term.

[20] `https://usa.usembassy.de/etexts/gov/democracy-elections.htm` (Last checked August 2020)

Companies and institutions, especially governments, may rely on social media to profile users and manipulate their attitudes and behaviors through the use of hate speech, fake news, and manipulative campaigns. This user profiling allows the companies and institutions to build targeted (possibly fake) advertising, to manipulate the vote of individuals. Brad Smith, the President of Microsoft, realized the extent of such a threat, and in a recent article that appeared in the Microsoft blog entitled "We are taking new steps against broadening threats to democracy" stated "It's clear that democracies around the world are under attack. Foreign entities are launching cyber strikes to disrupt elections and sow discord. Unfortunately the internet has become an avenue for some governments to steal and leak information, spread disinformation, and probe and potentially attempt to tamper with voting system"[21]. In this article, dated back to August 2018, the President of Microsoft refers to the United States general election in 2016, the French presidential election in 2017, and those that would have been the midterm elections in November of the same year.

In this section, we provide a thorough analysis of the state of the art studies and online articles that portray how new technologies interfered, and are continuing to interfere with many of the political elections worldwide.

1.3.2 Attacks

Over time, technological developments and innovation have provided the population with means that were unthinkable only a few decades before. Hand in hand, to not fall behind, politics needs to adapt to this new ever-increasing accelerated pace of communication. The politician on duty, if focused exclusively on offline election campaigns as in past years, might be crushed by contenders that are more familiar with social media and that know how to rely on them to manipulate the mass. This familiarity, when placed in the wrong hands, brings to the introduction of political bots. These automatic actors may be employed to give the politician a louder voice, by automatically sharing the electoral program to a wider audience. However, political bots may also be employed to raise misinformation campaigns about the contender, by relying on credible low-quality information aimed at demolishing the credibility of the opponent. Nowadays social media turned into political war camps, where propaganda and disinformation became the most effective, modern, worldwide political strategies[22].

In this subsection, we explain how, and to what extent, low-quality information can affect people's opinions. Finally, we describe the most impactful alleged political scandals in the literature, by reporting them within cate-

[21] `https://blogs.microsoft.com/on-the-issues/2018/08/20/we-are-taking-new-steps-against-broadening-threats-to-democracy/` (Last checked August 2020)

[22] `http://techpresident.com/news/25374/bad-news-bots-how-civil-society-can-combat-automated-online-propaganda` (Last checked August 2020)

gories according to the Nation in which the presumed scandal happened.

Public naivety

In 2016, the most tweeted topics on Twitter have been revealed to the public and, not surprisingly, four out of the ten positions were occupied by political topics, with Brexit, Trump, Election 2016, and BlackLivesMatter all making the list[23]. According to a Pew Research Center article, entitled "News Use Across Social Media Platforms 2016", it emerges that 62% of United States adults get news on social media. Compared to 2012, where Americans acquiring information on social media were (only) 49%, this boost gains importance and marks the beginning of a significant historical period. As depicted in Figure 1.4, the same study highlights that 64% of Americans usually get news from one social media only, without even double-checking the content to assess the reliability of the news. Regarding the other 36% of the United States citizens analyzed, 26% get news from two sites, while only 10% get news from three or more sites[24]. Furthermore, an interesting study published in 2016 showed that 60.66% of Twitter's users reshare articles without having read them, only relying on the "appealing" headline [18]. This information, together with the fact that viral fake election news stories outperformed real news on Facebook (in the last three months of the United States presidential campaign, the fake election news stories generated more engagement than the top stories from New York Times, Huffington Post, Washington Post, etc.)[25], and the fact that fake news headlines fool United States adults about 75% of the time[26], make it possible that low-quality information can opportunely reshape people's opinion nowadays.

Many studies in the literature confirm these beliefs. To consider a practical example, in [19], the authors face the problem of identifying social influence effects in observational studies. In particular, they report results from a randomized controlled trial of political mobilization messages sent to 61 million Facebook users during the 2010 US congressional elections. Results are striking, political self-expression, information seeking, and real-world voting behavior of millions of people have changed accordingly. The messages, besides having directly influenced the recipients, also had an im-

[23] `https://www.independent.co.uk/news/twitter-most-tweeted-moments-2016-donald-trump-brexit-black-lives-matter-rio-a7466236.html` (Last checked August 2020)

[24] `https://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/` (Last checked August 2020)

[25] `https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook` (Last checked August 2020)

[26] `https://www.buzzfeednews.com/article/craigsilverman/fake-news-survey` (Last checked August 2020)

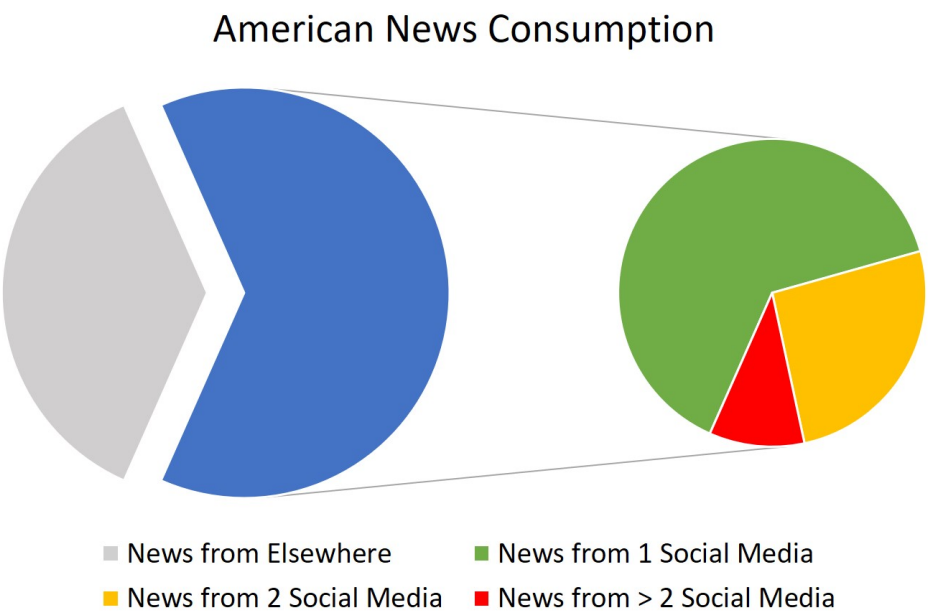

Figure 1.4: News consumption of Americans according to the study conducted by journalism.org

pact on the recipients' friends and friends of friends. However, on the other hand, several studies in the literature do not agree with these theories, and suggest that Americans may not be easily susceptible to online influence campaigns [20].

Alleged political scandals

In the following, we report some of the alleged political scandals in the literature. To allow a more effective reading, we organize the studies in paragraphs, each referring to a different Nation whose political elections have reported misinformation, fake news, or political bots employment. The results are summarized in Table 1.1

France. In 2017, during the French presidential election, a hacking attack led to the leakage of more than 20,000 e-mails related to the Emmanuel Macron's campaign. This leak immediately became viral due to the quickness in the spreading of related news throughout the Web. Indeed, US activists, WikiLeaks, and bots, consistently helped to amplify the leak by spreading the information on Twitter, Facebook, and 4chan [25, 26, 27], while shifting the attention to the Russian government (under Vladimir Putin) as responsible. In the same year, Emilio Ferrara published an interesting article [28], where he provided an extensive statistical analysis of

country	alleged facts	references
Argentina	Fake Twitter Accounts to support politicians, Astroturfing, Fake News	[21, 22]
Australia	Fake Tweets, Social Botnets, Fake Online Users, Cyborgs, Sock Puppets, Meat Puppets to support political leaders	[7, 23]
Austria	Disinformation	[24]
France	Hacking Attack that led to a massive leakage of politics-related e-mails, Bots, Disinformation	[25, 26, 27, 28, 29]
Germany	Political Bots to influence the election campaign	[30, 31, 32, 33, 34]
Italy	Political Bots to promote politicians, Fake News, Hate Speech, Political Propaganda, Misinformation	[35, 36, 37, 38, 39, 40]
Mexico	Spammers, Disinformation, Political Propaganda, Political Bots	[41, 42, 43, 21, 44, 45, 46]
Spain	Political Bots, Hate Speech	[47]
Russia	Political Bots, Trolls to manipulate the public opinion, Disinformation	[48, 49, 50, 51, 52, 53]
Turkey	Political Bots	[54]
Ukraine	Political Bots	[55]
UK	Political Bots	[56, 57, 58]
US	Fake News, Disinformation, Astroturfing, Spammers, Political Propaganda, Fake Online Users, Sock Puppets, Meat Puppets	[59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69]
Venezuela	Political Bots	[70]

Table 1.1: Alleged political scandals documented in literature.

the MacronLeaks disinformation campaign that occurred in the run-up to the French presidential election. He collected a Twitter dataset composed of 17 million posts related to the election that occurred in the period April 27 - May 7, 2017. Then, thanks to a combination of cognitive-behavioral modeling and machine learning techniques, he was able to effectively distinguish humans from social bots. As a result, out of 99,378 users participating in MacronLeaks, the proposed model classified 18,324 as social bots and 81,054 as human users, respectively. Furthermore, according to the author, "prior interests of disinformation adopters pinpoint to the reasons of scarce success of this campaign", considering that most of the audience was composed of English-speakers Americans rather than French users. The author highlighted that some automated Twitter accounts had already been used to discredit Hillary Clinton during the United States presidential election.

The second round of campaigning of the French presidential election has been analyzed by [29]. The authors, relying on a set of specific hashtags,

collected a 3-day Twitter dataset. The analysis of these tweets led to the following interesting discoveries: (i) the Twitter content related to Macron was dominating Le Pen traffic, although the latter was slowly growing; (ii) the amount of traffic generated by automated accounts doubled between the first and the second round of voting; and (iii), the ratio "professionally produced news content" to "other political content" passed from 2 to 1 (in the first round of voting) to 1 to 1 (in the second round of voting).

Italy. In a series of online articles [35, 36], it is reported the story of an accusation against a famous Italian comedian and politician, Beppe Grillo, leader of the Five Star Movement. According to the algorithm designed and implemented by Marco Camisani Calzolari, 54% of Beppe Grillo's 600,000 Twitter and Facebook followers might be bots. This news provoked an immediate reaction from the followers of the aforementioned politician, who started campaigns to prove themselves humans (e.g., "I am real", "I am not a bot" campaigns). The news also angered the politician who threatened to sue the Professor. Subsequent scientific studies proved that the algorithm by Camisani Calzolari was simplistic and highly inaccurate, thus likely to yield unreliable results [71].

Other articles, on the other hand, claim that bots have been adopted to help populists win elections in Italy [37]. According to a recent article named "#ElectionWatch: Italy's Self-Made Bots" [38], Lega's followers are automating themselves. Indeed, some bots have been created by the account-holders themselves, a phenomenon that is called "selfbot".

Moving on to another election, Facebook intervened personally to shut down multiple Italian pages that were accused to spread fake news, hate speech, political propaganda, and misinformation in the run-up to the 2019 European elections [39]. According to Avaaz, 23 pages with 2.46 million followers were shut down, where half of the accounts supported either the League or the Five Star Movement parties.

In [40], the authors discussed the problem of detecting social spambots. By performing several experiments they showed that neither humans nor state-of-the-art spambot detection applications can recognize them (i.e., they are erroneously labeled as genuine human-operated accounts). This study led to the construction of several datasets, that the authors have kindly made available to the scientific community. One dataset, that is called "social spambots #1", contains the activities of a group of almost 1,000 automated accounts that was discovered on Twitter during the 2014 Rome Mayoral election. Those automated accounts were almost impossible to distinguish from genuine accounts (e.g., they had a picture, a short bio, thousands of followers/friends, and used to publish few tweets posts every day, including quotes from popular people, songs, and YouTube videos). However, every time the candidate posted a new tweet in his/her personal profile, all the bots were triggered and retweeted it a few minutes later, to

reach a wider audience.

Resources

The dataset used in [40] as well as in [72, 73, 74, 75] for the analysis of traditional and modern, sophisticated social bots, is publicly available online[a]. The dataset contains Twitter data about legitimate and automated accounts, annotated by contributors of a crowdsourcing platform.

[a] `http://mib.projects.iit.cnr.it/dataset.html` (Last checked August 2020)

Russia. When it comes to Russia in these contexts, the Internet Research Agency fully deserves to be mentioned. The Internet Research Agency, sometimes referred to on the Web as "Troll Farm"[27], is a Russian company known for its commitment to online propaganda operations on behalf of Russian companies and the political interests of the Kremlin. The literature boasts several interesting scientific articles that have studied the Internet Research Agency's moves and analyzed the impact it had on national and international political elections. In [76], the authors investigated the scope of the Internet Research Agency activities in 2017. The study aims to show how easy it is for malicious actors to infiltrate social media to launch propaganda campaigns. However, it also shows that it is possible to track and understand this kind of activity by fusing content and activity resources from multiple internet services. Other studies focused on the content of the posts written by the alleged trolls. In [49], for example, the authors analyzed the content of 1.8 million images posted to Twitter by Russian trolls, as well as their posting activity. Among the many interesting findings, they: (i) showed that the image posting activity is strongly coupled with real-world events, with targets that were automatically changing; (ii) provided evidence of Russian trolls general targets, like Ukraine and United States of America; and (iii) showed how co-occurrence of these images are found in many popular OSNs. In [77], instead, the authors analyzed 27,000 tweets from 1,000 users suspected to be correlated with Russia's Internet Research Agency. The aim was to understand the differences with respect to a random set of Twitter users in terms of the disseminated content, the evolution of the accounts, and the general behavior. According to the study, although Russian trolls use to stay active for long periods and reach a conspicuous number of Twitter users, the action of making viral the spreading of news has a minor effect on social platforms. In [51], the authors analyzed the relationship between Russian Internet Research Agency's organized trolling efforts and

[27] `https://www.wired.com/story/facebook-may-have-more-russian-troll-farms-to-worry-about/` (Last checked August 2020)

political homophilia. The analysis of the Russian troll accounts participating in the #BlackLivesMatter debate shows that these conversations were divided into political groups, where the Russian trolls systematically took advantage of these divisions to accentuate disagreement.

A New York Times article dated back to September 2017 claimed that Fake Russian Facebook accounts bought $100,000 in political ads [52]. 3,500 of these ads, purchased by the Russian Government and released in May 2018 by the United State Congress House Intelligence Committee, have been analyzed by [53]. According to the study, the ads were principally biased toward the Democratic party as opposed to the Republican party. Furthermore, the authors suggested that, given the duration and the promotion of the Republican ads effort, Russia was trying to cause racial, religious, and political ideologies division rather than swaying the election.

In [50], the authors studied how state-sponsored trolls manipulate public opinion on the Web. They analyzed 10 million posts created and shared by 5,500 Twitter and Reddit accounts, identified as Russian and Iranian state-sponsored trolls. Among the interesting discoveries, they pointed out that: (i) during the United States 2016 presidential elections Russian trolls were pro-Trump, while Iranian trolls were anti-Trump; (ii) the campaigns undertaken by both the parts are strongly influenced by real-world events; and (iii) the behavior of the accounts does not remain consistent over time, leading to increasing complexity for automatic detection.

The bot activity within Russian political discussion (from February 2014 to December 2015) has been studied by [48], where the authors presented a methodology that relies on an ensemble of classifiers to allow accurate detection of bots on Twitter. Among the many interesting discoveries, they highlighted how the proportion of tweets produced by bots exceeds 50% of the traffic and that one prominent activity of those bots was spreading news stories and promoting media.

However, as already mentioned above, some studies claim that troll intervention has had no impact on Americans political attitudes and online behaviors. To make an example, in [20], the authors studied the attitudes and online behaviors of 1,239 Republican and Democratic Twitter accounts from late 2017 and, exploiting Bayesian regression tree models, claimed to not have found evidence that their interaction with IRA accounts has had an impact. According to the study, "the findings suggest that Russian trolls might have failed to sow discord because they mostly interacted with those who were already highly polarized.". The authors concluded by discussing some limitations of their study, including their inability to determine whether the troll accounts influenced the 2016 presidential election.

United Kingdom. In [56], the authors analyzed the use of political bots during the United Kingdom Brexit referendum about the European

Union membership. The two most active accounts during the StrongerIn-Brexit debate were *@ivoteLeave* and *@ivoteStay*, respectively. According to the authors, the two accounts were following a similar algorithm: they automatically retweet messages from their side of the debate, without ever creating new content. Among the many interesting bots they found, the authors discovered *@Col_Connaughton*, a pro-Palestine bot suitably repurposed to support Brexit, and *@Rotenyahu*, another pro-Palestinian bot used to retweet messages from *@Col_Connaughton*, to reach a wider audience. The behavior of bots during the Brexit debate after the 2016 referendum have been studied also by [78]. According to the analysis, more than 1,962 bot accounts participated in the Brexit debate. Among them, the author identified three bots that promoted the independence of Scotland, *@StillYesScot*, *@IsThisAB0t*, and *@FAO_Scotbot*. The goal of these bots was to influence, both constructively and destructively, the opinion of their followers.

The behavior of the social bots during the Brexit referendum has been analyzed also by [57] and [58]. In the first work, the authors discovered the existence of a network of social bots on Twitter composed of 13,493 accounts that suddenly disappeared after actively participating. The detailed analysis of these bots led to many interesting discoveries: (i) Twitter bots were able to rapidly generate small/medium-sized tweet cascades; (ii) the retweeted content included user-generated hyperpartisan news; and (iii), a botnet may follow a detailed organization (tiers or clusters), that allows a more effective replication of both active users and other bots content. Many of these accounts are alleged to be involved in the state-sponsored manipulation of the American elections, as per a list released by Twitter. Such a list contains 2,752 accounts Twitter believed to be controlled by Russian operatives. Given that a similar list for the UK referendum has never been released, the authors in [58] analyzed the behavior of the accounts related to the American election that produced UK-EU related content. They found 3,485 tweets posted by 419 accounts, gathered between August 29th, 2015, and October 3rd, 2017. According to the study, during that period the behavior of the bots changed from the generalized disruptive tweeting to retweeting each other, to allow other troll accounts to reach a wide audience. Furthermore, the authors showed that some of the bots were geographically located in Germany.

United States of America. Although Mark Zuckerberg, the Facebook's founder, considered foolish to believe that fake news shared on Facebook have caused changes in the choice of the leader to be elected during the US elections [79], the fact that the most discussed fake news tended to favor Donald Trump over Hillary Clinton brought the commentators to start wondering whether, without the influence of low quality information, Donald Trump would have been able to win the elections [59, 2]. Several studies

in the literature had the same doubt, and question the claim of the Facebook leader. For example, the authors in [60] studied the fake news impact on the 2016 United States Presidential campaign. They highlighted that: (i) approximately 1 American out of 4 visited a fake news website during the presidential campaign (i.e., from October 7th to November 14th); (ii) the majority of users that visited fake news websites are Trump supporters; (iii) almost 6 out of 10 visits to fake news websites came from the 10% of people with the most conservative online information diets; and (iv), Facebook was a key vector of exposure to fake news.

An article from Intelligencer entitled "Donald Trump Won Because of Facebook", dated back to November 2016, explains how one of the giants of social networks helped the current President of the United States to reach his position [80]. Among the reasons, the article mentions, there is a massive impact the spreading of fake news had on the voters. In fact, during the election campaign, there were many fake news reports against the current President's contender, Hillary Clinton, e.g., "Pope Francis Shocks World, Endorses Donald Trump for President"[28], "Hillary Clinton Bought $137 Million Worth of Illegal Arms"[29], "WikiLeaks: Clintons Purchase $200 Million Maldives Estate"[30], "Hillary Clinton's Alleged 'Lolita' Child Pedophile Sex Slave Island Ring"[31], and many others.

In [61], the authors discovered that a large fraction of users population, suspected to be social bots, accounted for a considerable portion (i.e., approximately one-fifth) of the content generated during the entire political conversation. The authors stated that the presence of social bots harmed the democratic political discussions, by altering public opinion and endangering the integrity of the Presidential election itself. After analyzing 14 million tweets during (and following) the presidential campaign and election, in [62] is highlighted that: (i) there is evidence that social bots were playing a key role in spreading the fake news; (ii) automated accounts were particularly active and tend to target influential users; (iii) many humans are vulnerable to the manipulation and tend to retweet the fake news shared by bots.

A Wired article entitled "Bots Unite to Automate the Presidential Election", dated back to May 2016, after providing an overview of the impact that bots have on the presidential elections [63], tells the story of Pepe Luis Lopez, Francisco Palma, and Alberto Contreras, three of the 7 million

[28] `https://www.snopes.com/fact-check/pope-francis-donald-trump-endorsement/` (Last checked August 2020)

[29] `https://www.snopes.com/fact-check/hillary-clinton-bought-137-million-worth-of-illegal-arms/` (Last checked August 2020)

[30] `https://www.snopes.com/fact-check/wikileaks-clintons-purchase-200-million-maldives-estate/` (Last checked August 2020)

[31] `https://www.inquisitr.com/3682274/hillary-clintons-alleged-lolita-child-pedophile-sex-slave-island-ring-scandal-5th-of-november-part-1-claims-by-anonymous/` (Last checked August 2020)

Twitter Trump's followers. Although being actively involved in supporting Donald Trump after his victory in the Nevada caucuses, these accounts are far from representing real human beings. Indeed they are automated accounts, employed to bring the Latino voters closer to Trump before the election. This is just one of the countless articles that claim the presence of bots among Donald Trump's supporters. Indeed, an article from Newsweek highlighted that nearly half of Donald Trump's Twitter followers are either fake accounts or bots [64] and, although many of them may be inactive, they played an active role in exaggerating the candidate's popularity during the US presidential election (i.e., astroturfing).

The low-quality information content during the US presidential election has been analyzed in [65], where the authors aimed at monitoring the traffic of websites that are known to create and share fake news in the months preceding the aforementioned election. Among their interesting discoveries, they pointed out that (i) social media was the main responsible for the circulation of fake news stories, and (ii) aggregate vote patterns were strongly correlated with the user visiting of websites serving fake news.

In the chaos caused by fake news, there is even someone claiming to be the reason behind Donald Trump's victory, as reported by a "The Washington Post" article [81].

An interesting study [66] analyzed the real-time search option implemented by the real-time websites, such as Twitter. According to the article, when considering political topics, the search provides results that would not be found in the first page while surfing the web. Results include fabricated content, personal opinions, unverified events, lies, and misrepresentations. To support this statement, the study provided evidence about the Massachusetts senate race between Scott Brown and Martha Coakley, showing that it is possible to predict the users' political orientation by solely relying on behavioral patterns of activity.

Other studies, as well as the United State Congress, investigated Russia's possible interference in the United States elections. Indeed, Russia has been accused of relying on both trolls and bots to spread misinformation and politically biased information. In [67], the authors collected a dataset containing 43 million posts shared between September 16th and September 21st, 2016. Among the interesting discoveries, they revealed that (i) about 4.9% and 6.2% of liberal and conservative users were bots and most of the trolls had a conservative, pro-Trump agenda; (ii) there were about 4 times as many Russian trolls conservative than liberal ones, and the former produced almost 20 times more content; (iii) conservative users retweeting these trolls messages were 30 times more than liberal users; and (iv) conservative users outproduced liberals users in terms of content at a rate of 35:1. A similar study has been performed by the same authors in [68]. The study in [69] investigated the influence of misinformation during the election by analyzing one of the largest repositories of Internet Research Agency's

tweets. According to the article, when the campaign reached its peak (i.e., between October and November 2016), most of the tweets drew the candidate Hillary Clinton in negative and moralizing terms, whereas Trump was depicted as a fighter and, potentially, a winner.

1.4 Countermeasures

In this section, we describe the countermeasures that have been proposed in the literature to mitigate low-quality information. Furthermore, we describe many state-of-the-art solutions that allow the identification of malicious actors, who are actively participating in the dissemination of low-quality information worldwide.

1.4.1 Low-quality Information.

Low-quality information has literally always been around, but the sounding boards are becoming wider and wider with the introduction of the latest technologies. The news that, in the past, used to circulate from mouth to mouth and could be stopped by either silencing an individual in the chain or isolating the source, have now turned into seemingly unstoppable distributed processes that bounce off the web reaching millions of people in seconds. The urgent need to defend against disinformation and low-quality information has also been identified by Microsoft that, in an online report[32], reported "electoral attacks" as the top tech issue for 2018. To proceed in a meaningful and effective direction, they introduced the Defending Democracy Program[33], aiming at: (i) protecting campaigns from hacking; (ii) increasing political advertising transparency online; (iii) exploring technological solutions; and (iv) defending against disinformation campaigns. According to the author, the global goal of the program is to protect the institutions and processes of the democratic countries in the years to come.

The problem has also been extensively studied in the literature. Indeed, in a report called "Combating Fake News: An Agenda for Research and Action" [82], the authors proposed some possible pathways to reduce the spreading of fake news, including: (i) providing the users with feedback related to the truthfulness of the news; (ii) providing alternative ideologically compatible sources that confirm (or deny) the authoritativeness of the news; (iii) detecting information promoted by both cyborgs and bots and tuning algorithm to manage those manipulations; and (iv) identifying the sources

[32] `https://blogs.microsoft.com/on-the-issues/2018/01/02/today-technology-top-ten-tech-issues-2018/` (Last checked August 2020)

[33] `https://blogs.microsoft.com/on-the-issues/2018/04/13/announcing-the-defending-democracy-program/` (Last checked August 2020)

of misinformation to reduce the promotion of information from those sources at the platform-level.

Researchers in both industry and academia are working hard to find innovative and effective ways to stop the spreading of low-quality information. Among the most interesting approaches, it is worth to mention:

- **Fake-News detection and removal**. Develop solutions to accurately identify low-quality information with the aim of clearing them out from the platform.
- **Credibility/Believability news evaluation**. Implement a mechanism to evaluate the credibility/believability of both the information and the source of the information. Supposedly, information with a low score would be identified as low-quality information and might be automatically filtered out by the platform.
- **Spreading the truth to counter low-quality information**. Implement a mechanism to spread authoritative information with the aim of fighting the already-spread low-quality information.

Fake-news detection and removal

In this section, we describe the possible steps that allow the identification of low-quality information. Since most of the existing studies that have addressed this problem are based machine learning for the identification of fake news, in this subsection we discuss the most commonly adopted machine learning pipelines for this task. Usually, machine learning approaches require a feature extraction phase, where specific characteristics of low-quality information are identified, followed by the construction of a model, where those features are exploited to build an effective classifier.

Feature extraction. While in traditional news media the low-quality information detection could be performed by relying on the text only, in social media, there is a bunch of auxiliary information that could be exploited. This information, that will be called *features* in this section, may be divided into two categories, depicted in Figure 1.5: News Content Features, and Social Context Features.

Definitions

News content features. News content features are generally the features that consider both the information and the meta information of the news, including the source, the title, and the text of the news itself.

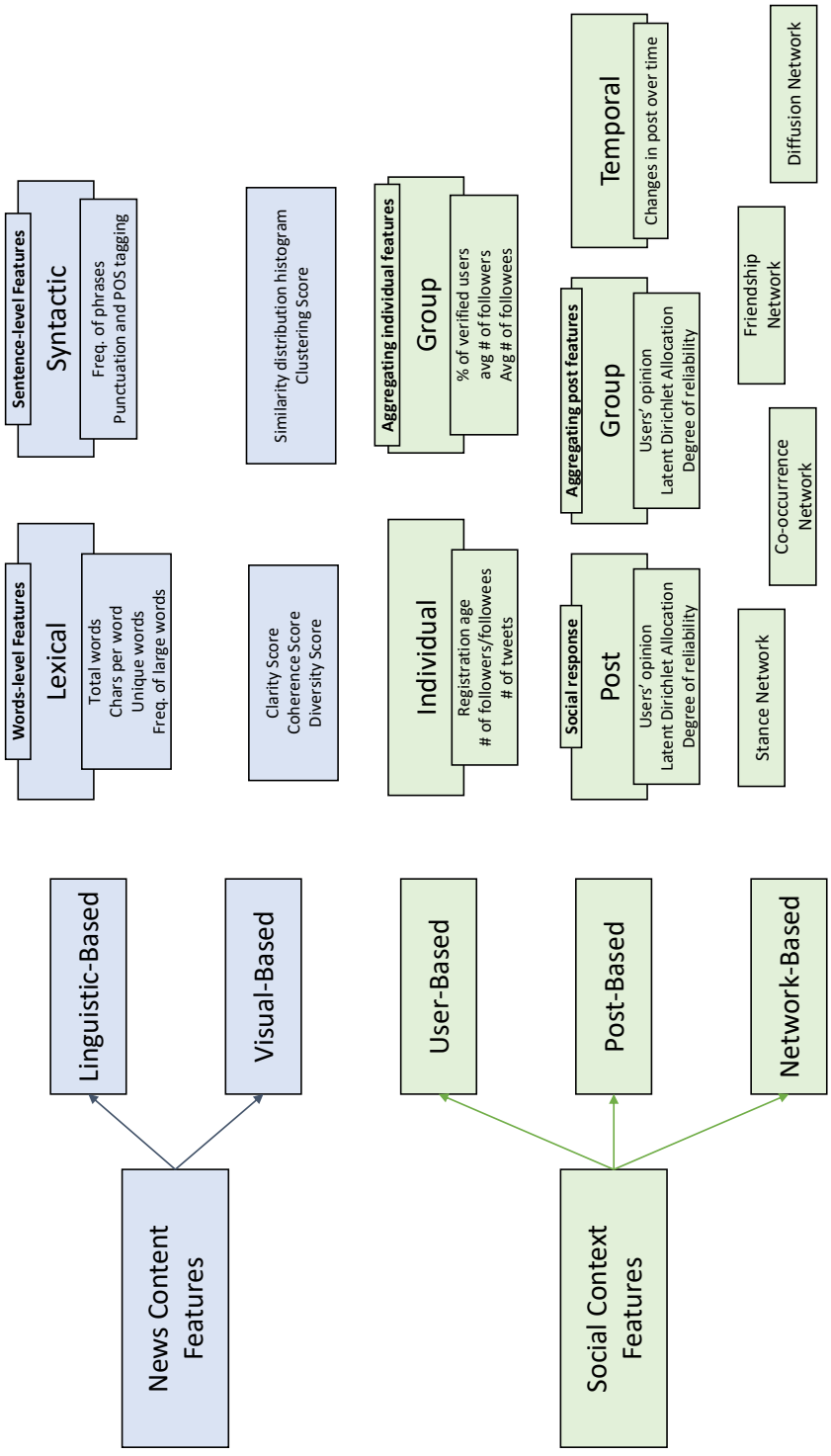

Figure 1.5: Categorization of News Content Features and Social Context Features.

News Content features may be, in turn, divided into two categories: *linguistic-based features*, and *visual-based features*, respectively. The idea behind the extraction of *linguistic-based features* is inspired by the fact that, usually, low-quality information is linguistically different when compared to authoritative information. To make an example, if we take into account spam messages, clickbait messages, or trolls hate speeches, we expect the text to be more appealing or more inflammatory when compared to everyday news [9]. *Linguistic-based features* may be extracted at different levels, from single characters, single words (lexical features), single sentences (syntactic features), and possibly others, according to the goal the researcher is willing to reach. *Visual-based features*, instead, are those who may be extracted from both videos and images. The rationale for this class of features stems from the growing importance of multimodal data in the spread of low-quality information and propaganda [9]. Even in this case, the intuition is that low-quality information images and videos present distinctive patterns, as they tend to impress the users to effectively attract their attention. Examples of *visual-based features* include: (i) visual clarity score; (ii) visual coherence score; (iii) visual similarity distribution histogram; (iv) visual diversity score; and (v), visual clustering score [83]. These features describe the characteristics of image distribution and allows to reveal hidden distribution patterns of images in news events.

We often consider low-quality information as multimedia traces (e.g., texts, audios, videos, images) from which we can extract news content features to evaluate their truthiness. Other determining roles are kept both by the author of the low-quality information and by the community that interacts with the information itself.

Definitions

Social content features. Social context features are the features that consider the user-driven social engagements of news consumption on social media platforms (e.g., the proliferation of the news over time, the veracity of the news, etc.).

Social context features are divided into three main categories: *user-based features*, *post-based features*, and *network-based features*, respectively.

Automated malicious actors, such as bots and cyborgs, are among the biggest low-quality information creators and spreaders. Being automatic accounts, their user-based characteristics appear to be different when compared to normal users ones. *User-based features* try to catch those differences and are divided into two categories: *individual-level features* and *group-level*

features, respectively. *Individual-level features* take into account single users' features, including the number of followers/followees, the age of the user, the number of messages commented, created, or shared, and possibly others [84]. *Group-level features*, instead, capture the characteristics of groups and consider the aggregation of *individual-level features* [2].

Post-based features aim at identifying information related to the veracity of the news contained in the social media posts. The intuition is that people express either sensational or skeptical reactions when commenting on low-quality information. These features could be divided into three categories: *post-level features*, *group-level features*, and *temporal-level features*. *Post-level features* identify the stance, the topic, and the credibility of each post, *group-level features* consider the aggregate information, while the *temporal-level features* aim to catch the temporal variations of the post over time [85].

Considering that social media expose users to the echo chambers phenomenon (i.e., beliefs are reinforced or amplified by communication and repetition inside closed systems), important features may be extracted from the network, to evaluate the structural relationships. Those features are called *network-based features*. Starting from the information available in the social media, several types of networks can be built, each with its unique characteristics. To make some examples, the friendship network highlights the closeness of individuals, the diffusion network helps to evaluate the trajectory of the spread of news, while the co-occurrence network shows off the user engagements with the news articles.

Model construction. After having extracted the relevant features, a low-quality information detection model can be trained. The approaches that are currently populating the literature could be divided into two categories: News Content Models, and Social Context Models, respectively, according to the features they rely on.

Definitions

News content models. News content models refer to the models that exploit news content features to perform the classification.

News content models can be divided into two main categories: *knowledge-based models* and *style-based models*.

Resources

Computational-oriented fact-checking:

- *ClaimBuster* [86]. Web-based automated live fact-checking tool that relies on NLP and supervised learning to identify low-quality information.
- Karadzhov *et al.* [87]. LSTM-based general-purpose framework for fully-automatic fact checking using external resources.

Considering that, by construction, most of the low-quality information tends to spread false claims, *knowledge-based models* aim at classifying them by relying on the evaluation of the truthfulness of the information. The truthfulness may be evaluated by using fact-checking approaches, that could be *Expert-oriented* (i.e., relying on human domain experts), *Crowdsourcing-oriented* (i.e., relying on the "wisdom of the crowd"), and *Computational-oriented* (i.e., relying on automatic systems) [88]. In the following, some of the most important fact-checking applications are reported.

Resources

Expert-oriented fact-checking:

- *PolitiFact*[a]. A fact-checking website that rates the accuracy of claims by elected officials and others on its Truth-O-Meter.
- *FactCheck*[b]. Nonprofit "consumer advocate" for voters that aims to reduce the level of deception and confusion in U.S. politics.
- *Snopes*[c]. A fact-checking website described as a "well-regarded reference for sorting out myths and rumors" on the internet
- *FullFact*[d]. The UK's independent fact-checking organization.
- *HoaxSlayer*[e]. Hoax-Slayer debunks email and social media hoaxes, thwarts Internet scammers, and combats spam.
- *TruthOrFiction*[f] *(Last checked August 2020).* A non-partisan website about eRumors, fake news, disinformation, warnings, offers, requests for help, myths, hoaxes, virus warnings, and humorous or inspirational stories that are circulated by email.

[a] `https://www.politifact.com/` (Last checked August 2020)
[b] `https://www.factcheck.org/` (Last checked August 2020)
[c] `https://www.snopes.com/` (Last checked August 2020)
[d] `https://fullfact.org/` (Last checked August 2020)
[e] `https://www.hoax-slayer.net/` (Last checked August 2020)
[f] `https://www.truthorfiction.com/`

Style-based models, instead, aim at capturing the writing style that has

been adopted to draft the low-quality information. The intuition is that, in order to persuade and manipulate people, the writing style of the low-quality information is different if compared with the "true" ones. There are two main categories of style-based models: *deception-oriented models*, and *objectivity-oriented models*, respectively. *Deception-oriented* models capture the deceptive statements or claims from news content, while *objectivity-oriented models* capture the style signals that bring to a decreased objectivity of news content.

Resources

Crowdsourcing-oriented fact-checking:

- *Fiskkit*[a]. Online commenting platform that improves the discussion of online articles by allowing users to tag, or call out, incorrect facts, bad reasoning, or uncivil behavior.
- Pennycook and Rand [89]. Fighting misinformation on social media using crowdsourced judgments of news source quality.
- Pinto *et al.* [90]. A Crowdsourcing-based process to perform the filtering, analysis, and classification of the news.

[a] `https://fiskkit.com/` (Last checked August 2020)

Definitions

Social content models. Social content models refer to those models who exploit social content features to perform the identification.

In social content models, auxiliary information, such as the user social engagements and the veracity of posts, are taken into account. Social content models can be divided into two categories: *stance-based models* and *propagation-based models*.

Stance-based models aim at capturing the users' viewpoints from the posts to infer their veracity, relying on stance detection. Stance detection is defined as the task of automatically determining if users are in favor, neutral, or against some target (i.e., a person, a policy, a product, an organization). For example, the task makes it possible to analyze the messages/speeches of Barack Obama to understand whether he is in favor of stricter gun laws in the United States [91].

The idea behind *Propagation-based models*, instead, is to evaluate the interrelations of the social media posts to predict their credibility. The

intuition is that the credibility of a news event is highly related to the credibility of relevant social media posts. The propagation process is analyzed by building credibility networks, that can be both homogeneous (i.e., consisting of a single type of entities, such as post or events) and heterogeneous (i.e., consisting of different types of entities, such as posts or sub-events) [88].

Credibility/believability news approaches

Another effective countermeasure to mitigate the low-quality information propagation is the adoption of credibility/believability approaches. Each post shared on social media is assigned a score, according to the content and the author who posted/shared it. The intuition is that posts containing low-quality information, as well as posts shared by questionable sources, are likely to be poorly scored. This low score will adversely affect the user's decision to share the post, thus limiting its propagation in the network.

The conception of a scoring mechanism for posts is challenging. If on the one hand, truthful information posted by truthful people should have a high score, on the other hand, the concept of freedom of expression should be guaranteed. This discussion opens the door to several philosophical questions:

- Does rating a post affect the author (disparagingly)?
- To what extent does filtering out low-scored posts affect freedom of expression of individuals?
- How would it be appropriate to score posts and users? Based on the expertise of people on the topic? Based on the number of "trusted" neighbors they have? Based on other metrics?
- How does an individual influence affect the scoring mechanism?

Twitter. Several studies in the literature focus on the Twitter platform which, in recent years, has conveyed several disinformation campaigns. For example, in [92], the authors introduced TURank, an algorithm to evaluate the users' authority scores on Twitter by relying on link analysis. In [93], instead, the authors introduce a machine learning-based approach to automatically identify rumors on Twitter by leveraging the concept of believability (i.e., "the extent to which the propagated information is likely to be perceived as truthful, based on the trust measures of users in Twitter's retweet network").

Expert finding. A possible way of assigning scores to posts and authors is based on the expertise the author has about the content of the post. An interesting research branch focuses on the identification of expertise on social networks. Authors in [94], for example, proposed a propagation-based

model that takes into account both person's local information and network information to measure the expertise of an individual, while in [95] the authors try to solve the problem "Given an expertise need and a set of social network members, who are the most knowledgeable people for addressing the need?".

Spreading truth

Another effective alternative consists of the spreading of truthful information to mitigate the low-quality information propagation. The intuition is to fight fire with fire, the propagation of low-quality information will be followed (or anticipated) by the propagation of truthful information, to provide the user with authoritative information immediately after (or before) receiving low-quality information. This approach, sometimes called *Rumor Debunking*, is far from being novel. Indeed, it has already been adopted by governments and other authorities throughout history, years before the introduction of social media. Their goal was to broadcast messages to the entire population to obscure any possible rumor.

However, when considering OSNs, several challenges arise, due to the huge number of users and their highly clustered network structure. In recent years, many researchers have become interested in the problem and have made significant contributions, looking at the problem from different perspectives. In [96], to make an example, the authors developed two models. In the first model, called "Delayed Start Model", a local authority discovers a rumor after a variable number of days and decides to start an anti-rumor campaign. In the second model, called "Beacon Model", instead, the authors disseminate several vigilantes inside the network able to both detect rumors and respond to them effectively.

Considering the problem from a theoretical point of view, authors in [97] aim at finding the smallest set of highly influential nodes whose decontamination would help to contain the viral spread of misinformation. This study, rather than identifying low-quality information messages, exploits the propagation mechanism to restore the "truth" within the platform. According to the authors, the key difference between the low-quality information propagation and the propagation of their truthful information lies in the propagation speed. Indeed, misinformation often starts from less influential nodes, thus being inevitably constrained by the structure of the graph itself.

When implementing the aforementioned alternatives, there are several trade-offs to take into account. For example, it may be the case to avoid overwhelming users with information they may not find interesting. To do so, every user should only receive truthful information about her topics of interest. This process, however, will eventually lead to the filter bubble problem, where the user only receives material related to her knowledge and would incur difficulties when she wants to be informed about new topics.

Furthermore, even this approach hides relevant philosophical questions. Among the many, who knows the truth? The fact of knowing the truth could not be taken for granted. Indeed, most of the time, the line between the truth and the non-truth is so subtle that it is not easily perceptible. Training someone to distinguish between truth and non-truth may require them to access private (or worse, classified) documents, or discuss with groups of experts. Both the resources may not be available in the time of need.

Datasets

The difficulty of finding accurate data to be used as ground truth remains one of the most critical open problems to face when it comes to creating proper classifiers. According to several studies on the web, a dataset should be:

- *Accurate.* The dataset should not contain erroneous elements;
- *Valid.* The data collected should reflect the requirements;
- *Reliable.* The instances collected in the dataset should not contradict the ones collected in similar datasets;
- *Timely.* The data has to be collected at the right moment in time and has to be updated when the reality changes;
- *Complete.* The data collected should provide the overall picture;
- *Available.* The data should be available for further future acquisition; and
- *Detailed.* The data collected should contain as much information as possible.

The following resource boxes contain lists of datasets available in the literature, on GitHub, and the Web in general.

Resources

Github:

- **BuzzFeedNews 2016**[a]. Dataset comprising 1,627 manually fact-checked (claim-by-claim) articles written on Facebook by 9 news agencies from September 19 to 23 and September 26 to 27 2016. (Released in 2016).
- **FakeNewsChallenge**[b]. Dataset released by fakenewschallenge.org comprising 49,972 posts associated with 4 stances: unrelated, discuss, agree, and disagree. (Released in 2017).
- **BuzzFeedNews 2017**[c]. Dataset released by BuzzFeedNews containing the 50 biggest fake news hits on Facebook in 2016 and 2017. (Released in 2017)

[a] `https://github.com/BuzzFeedNews/2016-10-facebook-fact-check` (Last checked August 2020)

[b] `https://github.com/FakeNewsChallenge/fnc-1` (Last checked August 2020)

[c] `https://github.com/BuzzFeedNews/2017-12-fake-news-top-50` (Last checked August 2020)

Resources

Literature:

- **CREDBANK** [98]. Large-scale crowd-sourced dataset containing more than 60M tweets grouped into 1049 real-world events, each annotated by 30 human annotators, covering 96 days from October 2016. (Released in 2016).
- **LIAR** [99]. Open-source dataset for fake news detection. 12.8K manually labeled short statements. (Released in 2017).
- **FacebookHoax** [100]. Dataset containing non-hoax and hoax posts, for a total of 15,500 posts from 32 pages (14 conspiracy pages and 18 scientific pages), collected by relying on Facebook Graph API. (Released in 2017).
- **BuzzFace** [101]. Dataset made up by extending the BuzzFeed dataset with Facebook news articles. Among the many possible uses, the authors claim that this dataset could be used for detecting social bots. (Released in 2018)
- **Goldbeck et al.** [102]. Dataset containing 283 fake news stories and 203 satirical stories from a diverse set of sources, posted between January 2016 and October 2017 regarding American politics. (Released in 2018).
- **FakeNewsNet** [103]. Fake news data repository containing two comprehensive datasets with diverse features in news content, social context, and spatiotemporal information. (Released in 2019).
- **Other datasets**. [104, 105, 106]. (Released in 2019)

Web: German fake news dataset[a], ISOT research lab datasets[b], and Kaggle dataset[c].

[a] `https://zenodo.org/record/3375714` (Last checked August 2020)
[b] `https://www.uvic.ca/engineering/ece/isot/datasets/` (Last checked August 2020)
[c] `https://www.kaggle.com/c/fake-news/data` (Last checked August 2020)

1.4.2 Malicious Actors

The extremely quick propagation of low-quality information is mainly due to bots (e.g., social bots, cyborgs) and malicious accounts (i.e., trolls, spammers, sock puppets, meat puppets). Depending on the context, their goal is to share the low-quality information with as many people as possible in the shortest time. Starting from this assumption, one of the first steps that comes into mind to implement an effective countermeasure to mitigate the

propagation of low-quality information consists of the distinction between malicious actors' accounts and the normal users' ones.

Social bots

Automated accounts have an enormous presence on social media platforms, and their number is increasing every year at an unprecedented rate. According to a study conducted in 2017 by CNBC[34], as many as 48 million Twitter accounts are not real, existing people. Although the social media operators are doing their best to identify them and clear them out from their platform (e.g., Facebook took down 2.2 billion fake accounts in the first quarter of 2019[35], Twitter purges 174,000 fake accounts linked to the Chinese government in the first quarter of 2020), these numbers seem to represent only the tip of the iceberg. Indeed, identifying a bot is far from being trivial[36], and there are several strategies that malicious actors can adopt to avoid (or at least slow down) detection [2]. As expected, these bots take part in political discussions and relevant events, trying to manipulate the mass or divert attention. To make an example, during the coronavirus pandemic in 2020, it is estimated that half of Twitter accounts pushing to reopen America might be bots[37]. According to the study, many of the accounts have been created in February and have started spreading and amplifying misinformation including false medical advice and conspiracy theories about the virus origin, to finally push to end stay-at-home orders and reopen the country. These accounts might had an impact on the resident population who, conditioned by the misinformation campaign on Twitter, could risk to unnecessarily expose themselves to the virus, increasing the number of infections in the best case, the number of the victims in the worse.

In order to mitigate the aforementioned misinformation campaigns there is the need to understand the extent of a bot network that, in turn, requires to be able to effectively distinguish automated agents from normal users. Several researchers worked in this direction, facing the problem from different perspectives and tailoring solutions for different platforms.

Twitter. Many of the studies in the literature focus on Twitter, being one of the platforms that are most populated by bots and one of the few for which collecting data is still possible. One of the first studies trying to address this challenge is discussed in [5]. To assist human users in identifying

[34] `https://www.cnbc.com/2017/03/10/nearly-48-million-twitter-accounts-could-be-bots-says-study.html` (Last checked August 2020)

[35] `https://variety.com/2019/digital/news/facebook-took-down-2-2-billion-fake-accounts-in-q1-1203224487/` (Last checked August 2020)

[36] `https://firstdraftnews.org/latest/the-not-so-simple-science-of-social-media-bots/` (Last checked August 2020)

[37] `https://www.technologyreview.com/2020/05/21/1002105/covid-bot-twitter-accounts-push-to-reopen-america/` (Last checked August 2020)

whom they are interacting with, the authors focused on the classification of humans, bots, and cyborgs accounts. During the study, they took into account legitimate bots (i.e., bots that generate a large number of benign tweets delivering news and updating feeds); malicious bots (i.e., bots that spread spam or malicious contents); and cyborgs (i.e., entities in the middle of humans and bots), respectively. The differences among humans, bots, and cyborgs have been evaluated in terms of *tweeting behavior*, *tweet content*, and the *related properties* of the account. The authors proposed a framework that is composed of four main components: (i) an entropy-based component; (ii) a spam detection component; (iii) an account properties component; and (iv) a decision-maker. The proposed solution allows discriminating a human with an accuracy of 98.6%, a cyborg with an accuracy of 91.7%, and a bot with an accuracy of 96%.

On the same platform, the study in [107] aims at understanding: (i) how the social botnet grows over time, (ii) how the tweets produced by automated accounts differ from the ones produced by normal users, and (iii), how social botnets may influence relevant discussions on the platform. The authors built a dataset containing about 3,000 tweets in English and Arabic from the famous Syrian social botnet. They highlighted that the behavior and the content of this particular botnet did not align with the general conception of botnets portrayed in the prior literature (e.g., the Syrian Social Botnet was exceptionally long-lived if compared to the life span of the other reported botnets, it is still not clear if the botnet was mimicking human behavior or if it was only interested in flooding Syrian civil war-related hashtags with topics that are not war-related), thus becoming harder to detect and interesting to study.

In [108], the authors presented a supervised machine learning framework, trained with more than a thousand features of different categories, including (i) user-based features; (ii) friend features; (iii) network features; (iv) temporal features; (v) content and language features; and (vi) sentiment features, respectively. According to the results, between 9 and 15% of active Twitter accounts are bots, and the simplest of them tend to interact with the users that present human-like behaviors. A subsequent application of unsupervised machine learning (i.e., clustering) techniques to the bots found by relying on the supervised machine learning framework described above allowed the authors to categorize the bots in several subclasses, including self-promoters bots and spammer bots.

However, not all bots have to be considered malicious. The authors in [109] pointed it out and developed a profiling framework that allows to effectively distinguish humans from consumption bots (i.e., automated agents aiming at aggregating contents from various sources and provide services for personal consumption), broadcast bots (i.e., automatic agents aiming at disseminating information), and spam bots (i.e., automated agents aiming at posting malicious content or aggressively promote content), respectively.

By relying on Machine Learning techniques, including Naive Bayes, Support Vector Machine, Logistic Regression, and Random Forest, they reached a Precision of 0.8432 with the Random Forest classifier, a Recall of 0.8254 and an F1-score of 0.8228 with the Logistic Regression classifier.

Facebook. Other studies in the literature focus on the Facebook platform. In [110], to make an example, the authors relied on the emotions that shine through from the posts to distinguish between bots and normal users. The work is based on the assumption that, while the posts of the real users reveal a variety of emotions such as sadness, joy, fear, and anger, dictated by life experiences, fake users (or bots) posts, that usually aim at accomplishing specific tasks, are likely to present a limited range of emotions. The proposed approach has been trained using 12 emotion-based attributes: the first eight features (i.e., (i) fear, (ii) anger, (iii) sadness, (iv) joy, (v) surprise, (vi) disgust, (vii) trust, and (viii) anticipation) are the Plutchik's basic emotions [111], while the remaining 4 have been introduced in this study, and include: (ix) the # of categories of emotions expressed by the user in their posts, (x) the variance in the emotions expressed by the user, (xi) the fraction of posts containing positive sentiment words, and (xii), the fraction of posts containing negative sentiment words, respectively. Several classifiers have been trained, including Support Vector Machine, JRip, Naive Bayes, and Random Forest, and tested on Facebook users, reporting accuracies of 87.66% (Support Vector Machine), 85.71% (JRip), 83.44% (Naive Bayes), and 90.91% (Random Forest), respectively.

Unsupervised learning. Most of the approaches described above rely on supervised learning techniques to distinguish between automated agents and humans. Although these approaches show very good performance, they require a labeled dataset of bots to be trained. According to other studies in the literature, this is a limiting requirement, since finding a reliable dataset is complex, and building one from scratch may take indefinite time [112, 2]. To overcome these limitations, in [113], authors proposed a seminal (unsupervised) approach to detect bots by relying on the correlation between accounts. Indeed, the approach is based on the assumption that if accounts are abnormally correlated, they are very unlikely to be human operated. They reached an accuracy of 94% and were able to identify bots that other methods did not detect. The proposed system produces a daily report about bots that the authors kindly make available online for further analysis and experimentations[38]. In [114], the authors relied on principal component analysis and K-means clustering techniques to identify Twitter bots during the 2019 Canadian elections. According to the study, the average number of daily tweets of bot accounts, as well as their percentage of retweets and

[38] `https://www.cs.unm.edu/~chavoshi/debot/` (Last checked August 2020)

daily favourites, are significantly higher when compared to human accounts. For a detailed survey of unsupervised bot detection technique, we point interested readers to the extensive literature reviews, such as [2].

Public services. Being the detection of social bots one of the most impactful and interesting fields of study of the last years, several studies in the literature provide online services accessible by everyone to identify automated agents. Botometer [115] (formerly BotOrNot) is one famous example, a publicly available and accessible service able to measure the similarity between a Twitter account and a social bot. After 2 years of the official release, that happened in May 2014, the Botometer service served over 1 million requests both via the website and the APIs. Another example, already mentioned in the previous paragraph, is DeBot [116], a parameter-free, unsupervised learning method able to identify bots in the Twitter network. In February 2017, DeBot collected more than 710,000 unique bots. The valuable performance of DeBot pushed the authors to provide further services, including a bot archive API[39] and an on-demand bot detection platform[40]. The latter allows the user to detect bots related to a given topic, into a specific geographical region, or starting from a set of users. Another interesting resource is Tweetbotornot2[41]. It consists of an open-source, out-of-the-box classifier that allows the detection of bots on the Twitter platform.

Online competitions and Web resources. The Defense Advanced Research Project Agency (DARPA), the agency of the United States Department of Defense responsible for the development of emerging technologies, gave several contributions to protect the democratic processes from alarming disinformation campaigns. Indeed, they highlighted the need for identifying and eliminating the political bots before they reach a critical level of influence. In February/March 2015, DARPA held a 4-week competition where the goal was to identify influence bots within Twitter. The description of this DARPA challenge, together with the description of the method used by the three top-ranked teams can be found in their white paper [117]. Another interesting challenge is the Author Profiling shared task at PAN 2019[42]. The goal was to determine whether the author of a feed on Twitter was a bot or a human. The overview of the approaches submitted by the 56 participating teams is reported in [118] in terms of preprocessing, feature selection, and classification approach.

[39] `https://www.cs.unm.edu/~chavoshi/debot/api.html` (Last checked August 2020)

[40] `https://www.cs.unm.edu/~chavoshi/debot/on_demand.html` (Last checked August 2020)

[41] `https://github.com/mkearney/tweetbotornot2` (Last checked August 2020)

[42] `https://pan.webis.de/clef19/pan19-web/author-profiling.html` (Last checked August 2020)

Furthermore, the Web boosts countless interesting articles identifying ways to fight social bots on online platforms. To make an example, in [119], DFRLab reported a list containing twelve clues that can help in the distinction between automated accounts and normal users on online platform, such as Twitter. The list is composed by: (i) activity, the number of tweets divided by the number of active days may be a good indicator (72 tweets per day can be considerable suspicious, threshold that the Oxford Internet Institute's Computational Propaganda set to 50); (ii) anonymity, in general the less personal information an account gives, the more likely it will be a bot; (iii) amplification, most of the bots present few or no original posts, but a never-ending list of retweets; (iv) low posts/high results, bots tend to amplify single posts; (v) common content, the content of the shared and published posts is similar; (vi) the secret society of silhouettes, the oldest bots did not have an avatar image, they were represented by the standard silhouette; (vii) stolen or shared photo, the bots without the standard silhouette have often stolen the profile image on the web, although nowadays modern generative adversarial networks are able to create new plausible faces, indistinguishable from real people's ones; (viii) bot's in a name, many bots have a @ symbol before the name, or an alphanumeric string generated by an algorithm; (ix) Twitter of Babel, many of the bots are political while others, although developed for commercial purposes, may be used to boost political tweets; (x) commercial content, advertising is seen as a classic indicator of botnets; (xi) automation software, the use of URL shortener is a possible clue, a frequent use of them is an indicator of automation; and (xii), retweets and Likes, an indicator of a botnet can be the ratio between the number of retweets and the number of "like" of a particular post, given that some bots are programmed to both like and retweet the same tweet.

Resources

One of the few publicly available repositories of datasets for training bot detection methods is hosted by the Indiana University[a].

[a] `https://botometer.iuni.iu.edu/bot-repository/datasets.html` (Last checked August 2020)

Fake profiles

Before starting an online propaganda campaign, as well as a misinformation/disinformation campaign on social media, one of the first steps is to create a fake profile, to avoid being identified and prosecuted in any way. In each campaign, thousands of plausible fake profiles are distributed among real people, with the aim of manipulating their opinions by sharing artfully

created low-quality information with them. Despite the clear repercussions that these fake profiles bring, there are still no definitive ways to prevent them from being created and used as part of information operations. To make an example, an article on ArsTechnica states that social media platforms leave 95% of reported fake accounts up[43]. Other recent results reported that only 5% of sophisticated fake profiles are actually removed from Twitter [40]. Scholars are therefore working hard to make their contribution to the cause, proposing innovative solutions that could be employed by social media platforms to mitigate the issue.

Facebook. Many studies in the literature are focusing on the Facebook platform to identify fake profiles. To make an example, in [120], Conti *et al.* explained how dangerous an adversary impersonating a real person on an OSN can be, aiming at acquiring as much personal information as possible to steal online identities. In this study, the authors investigated this problem and proposed an interesting approach to mitigate it. They first consider the time evolution of OSNs (i.e., Facebook, in this specific study), together with the characteristics of the growth rate of the network. Then, they showed that attackers aiming at impersonating a victim will eventually avoid people that are close friends of the victim in real life. This will change the evolution of the network and the interaction with the friends, thus allowing easier detection.

From manual to automatic detection. However, some companies prefer to rely on human employees to detect, verify, and shut down fake accounts. An example is given by Tuenti, one of the largest OSNs in Spain. According to [121], 14 Tuenti's full-time employees have been working exclusively on that task, thus resulting in substantial monetary costs. The difficulty in capturing the behavior of both fake and real OSN profiles is the main reason behind the non-automation of the task. To cope with this limitation, the authors introduced SybilRank, a scalable tool that relies on social graph properties and allows to rank the users according to the perceived likelihood of being fake. SybilRank has been deployed in the operation center of Tuenti and proved to have very good performance. Indeed, approximately 90% of the 200,000 accounts identified by SybilRank warranted suspensions, while Tuenti's current manual approach identified only 5% of the inspected accounts to be fake.

Dark Web. Fake profiles are also widespread in the Dark Web, where the anonymity guaranteed by the technology acts as additional protection for the users. To make an example, by relying on The Onion Router (TOR)

[43] `https://arstechnica.com/tech-policy/2019/12/social-media-platforms-leave-95-of-reported-fake-accounts-up-study-finds/` (Last checked August 2020)

anonymity mechanisms, users are allowed to anonymously access anonymous services (i.e., hidden services). If on the one hand, hidden services can be meeting places where dissidents of authoritarian countries may express their opinions without incurring in censorship or prosecutions, on the other hand, they often host questionable, controversial forums (e.g., CRD Club, a Russian site on computer hacking and technology frauds; or Dream Market, a forum in which the quality of drugs and related vendors sold in the associated marketplace is discussed). The first approach to detect and geographically deanonymize communities of the hidden services on the Dark Web has been proposed by [122], where the authors, by relying on a combination of Gaussian Mixture Models and the Expectation-Maximization fitting method, have been able to identify the time-zone of the users by only exploiting the timestamps of the comments they posted. The paper has been extended in [123] to validate and confirm the first results by applying Native Language Identification techniques. Native Language Identification is the task of determining the native language of an author based only on his writings in a second language. The authors applied Native Language Identification techniques to discover the geographical distribution of users in the Dark Web's hidden services by only relying on the English text of the messages they posted, thus validating the results obtained during the analysis carried out in the previous study.

Twitter. Several studies in the literature focus on the Twitter platform, being one of the social media that most suffer from the presence of fake profiles. In [124], to make an example, the authors analyzed 62 million publicly available Twitter user profiles to identify automatically-generated fake profiles. The authors rely on a combination of pattern-matching algorithms on screen names (with an analysis of tweet update times) and social graph analysis for detecting fake profiles on the OSN. An analysis of the characteristics of the fake profiles led to several interesting discoveries: (i) the fake profiles were created in batches, over intervals of less than 40 seconds; and (ii), all of the fake profiles were created on some weekdays during select times of the day, thus suggesting the existence of some manual element in either the generation or the maintenance of the profiles.

Survey. Many of the relevant studies in the literature about the detection of fake profiles have been collected in [125]. In this survey, the authors classified them under three major categories: network structure or graph-based defense, feature or content-based defense, and hybrid approaches, which is a combination of both. The studies, together with their assumptions, characteristics, datasets, and selection techniques, are effectively visually represented through timelines who also include their date of publication. Finally, the authors identified a list of open issues and proposed some interesting countermeasures that are worth investigating.

LinkedIn. Facebook and Twitter, although among the most famous and popular OSNs, are only two of the many platforms currently affected by the fake account epidemic. Indeed, the authors in [126] proposed an approach to identify fake profiles in LinkedIn, the American business and employment-oriented service. According to the authors, the approaches that have been proposed for other social networks rely on data that are not publicly available for profiles on LinkedIn. For this reason, in this study, the minimal set of profile data required to identify fake accounts on LinkedIn is introduced, together with a suitable data mining approach for this task. The proposed approach, when applied to the 3 datasets consisting of 37 profiles each, reached an accuracy of 87% with and a True Negative Rate of 94%.

Astroturf

The term astroturf was coined in 1966, when a form of artificial grass was installed in Houston, Texas. This turf, although designed to look like natural grass, is fake [127]. The term has been then extensively used to refer to the action of creating impressions of widespread support for policies, individuals, or products, where a little or (most of the time) none of such support really exists. When applied to political contexts – i.e., political astroturfing – it appears as a centrally coordinated disinformation campaign in which all the participants pretend to be ordinary citizens who act independently according to their will. Predictably, political astroturfing has the potential to heavily influence electoral outcomes, as well as any other forms of political behavior [128].

The early detection of astroturfing campaign is crucial, since it allows to timely manage a phenomenon that would soon become uncontrollable due to its ever-increasing nature. Astroturfing has been studied in the literature both from a theoretical and from a practical perspective.

Theoretical perspective. In [129], the authors were among the first to study the astroturfing phenomenon from a theoretical perspective. They provided an interesting theoretical definition of online astroturfing and discussed many of its key attributes. The motivations behind the employment of astroturfing, as well as the methods used to start an astroturfing campaign, and the mechanisms for effective astroturfing have been discussed in detail, to provide valuable insights for both practitioners and scholars.

Practical perspective. In [8], on the other hand, the authors studied astroturfing political campaigns on microblogging platforms. The study was followed by the implementation of a machine learning framework, called Klatsch, which relies on topological features, content-based features, and crowdsourced features to detect the early stages of the viral spreading of

political misinformation. The proposed model reached 96% accuracy when applied in the detection of astroturfing content in the run-up to the 2010 United States midterm elections. Authors in [130], instead, proposed a hidden astroturfing detection approach based on a combination of the analysis of emotions and the detection of unfair ratings. The system is composed of five modules: (i) a data crawling module; (ii) a pre-processing module; (iii) a bag-of-words establishment module; (iv) an emotion mining and analysis module; and (v), a matching module, respectively. The results show how this approach can detect implicit astroturfing under the prerequisite of an improvement in the classification accuracy of the emotions.

There are many ways to start an online astroturfing campaign. The goal, as defined above, is to increase the credibility of statements/organizations without revealing the identity of the supporters. A politician, for example, might buy an army of anonymous bots (or fake accounts) who would start following him and retweeting the tweets he shares, to reach a wider audience. An inexperienced eye may consider those followers real, and may unconsciously be affected by the ever-increasing popularity of the politician. The authors in [71] were among the first to propose an approach aiming at detecting fake Twitter followers. In this interesting paper, they: (i) reviewed many of the most relevant features for the detection of atypical Twitter accounts; (ii) created a dataset of verified humans and fake followers, that has been released open-source[44]; and (iii), trained a machine learning classifiers built over the features identified. The proposed approach demonstrated the effectiveness of the selected features and correctly classified more than 95% of the accounts of the original training set.

Survey. Many of the above-mentioned studies, together with other interesting ones, have been gathered into a recent survey about the astroturfing techniques [131]. The authors provided a detailed taxonomy of the approaches, thoroughly summarizing the studies in the literature, and pointed out interesting research challenges and research directions for future possible contributions. Challenges and potential research directions include the need for annotated astroturfing datasets, to create an effective ground truth; the consideration of the size and the dynamicity of the OSN's data; and the study of the privacy and security in online astroturfing. Furthermore, the authors made available the set of essential characteristics they believe a generic astroturfing detection framework should present, to address the challenges identified. According to the study, a framework for detecting astroturfing should present the following characteristics: (i) generalization across multiple OSN platforms, the framework should require no (or at max, a few) modifications to be applied to different OSNs; (ii) diverse

[44] `http://mib.projects.iit.cnr.it/dataset.html` (Last checked August 2020)

feature space design, a framework relying on hybrid feature spaces would be more robust and valuable with respect to a framework that relies on a single class of features; (iii) platform selection and data collection, simplicity in data collection should be a key indicator for choosing the platform; and (iv) facilitation of network interaction data, the framework should facilitate the extraction of the data at the network level to identify the graph of the interactions, a data structure that is crucial for the extraction of useful features.

Spammers

Although apparently they seem less threatening than the other malicious actors we have considered, spammers heavily outperform contenders when it comes to the number of accounts within social media. Indeed, according to an article published on Bloomberg, experts estimated that as many as 40% of social network accounts are used for spam[45]. Social media platforms are now appearing as heavens on earth for spammers who, instead of senselessly searching for email addresses on the Web to overwhelm with their content, have the opportunity of targeting specific demographic segments, particular groups, or specific subsets of people, according to the organization of the social media they are in. Social media has implemented various mechanisms to report the accounts that exhibit these behaviors, intending to clearing them out from the platform once a given threshold of reports has been reached (or once careful investigations have been carried out on the reported account). However, spammers use to frequently change their behaviors, thus being extremely difficult to be identified [2]. The literature boasts many contributions to mitigate the problem, with solutions that consider the identification of spammers either as a classification problem or as an anomaly detection problem, respectively.

Classification problem. In machine learning and statistics, classification is the problem of identifying to which set of categories a new observation belongs, based on the features extracted during the training phase. In literature, many contributions to identify spammers treated the problem as a classification problem. In [132], the author relied on three graph-based features (i.e., the number of followers, the number of friends, and the ratio between them) and three content-based features (i.e., the number of duplicate tweets, the number of HTTP links, and the number of replies/mentions) extracted from the top 20 most recent tweets of the users. Results showed an accuracy of 91.7% by using a Naive Bayesian classifier on a dataset proposed by the author. A new dataset is also proposed by the authors in [133]. To build the dataset, they created a set of honey-profiles on three large social

[45] `https://www.bloomberg.com/news/articles/2012-05-24/likejacking-spammers-hit-social-media` (Last checked August 2020)

networks (i.e., Facebook, Twitter, and MySpace) and they logged the contacts and the received messages. From the received messages, the authors identified four categories of spam bots:

- *Displayers:* automated agents that, instead of directly sending spam messages to other users, display spam content on their profile page. This spammer is considered the least effective one in terms of people reached;
- *Braggers:* automated agents that send messages to their feed (e.g., status updates on Facebook, tweets on Twitter);
- *Posters:* automated agents that send messages to other users (i.e., on users' wall on Facebook); and
- *Whisperers:* automated agents that send private messages to other users.

Having in mind these distinctions, the authors trained a classifier by relying on 6 features: the FF ratio (i.e., the number of friend requests sent over the number of friends), the URL ratio (i.e., the number of messages that contain URLs over the total number of messages), the messages similarity, the friend choice (i.e., the total number of names among the friends over the number of distinct first names), the number of messages sent, and the number of friends. The classifier allowed to detect 130 spammers in the Facebook Los Angeles and New York networks dataset and 15,392 spammers on Twitter.

This problem has been considered as a classification problem also by [134], where five types of spammers have been identified:

- *Sole spammers:* accounts created with the sole purpose of spreading malicious scripts;
- *Twitter followers market merchants:* accounts that send both information and promotional links;
- *Pornographic storytellers:* accounts that disseminate pornographic content;
- *Fake profiles:* users that impersonate the profiles of genuine users; and
- *Compromised profiles:* existing genuine profiles that have been stolen by attackers.

The dataset is composed of approximately 20 million tweets written and shared by around 100,000 users. By using a machine learning classifier that relies on Bayes Net, Logistic Regression, J48, Random Forest, and AdaBoostM1, the authors have been able to predict spammers with an accuracy of 92.1%.

Anomaly detection. All the studies described in the previous paragraph considered the detection of spammers in OSNs as a classification problem. In [135], instead, the problem is treated as an anomaly detection problem. Anomaly detection consists of identifying unexpected events or items in data sets, which differ from the norm. Besides the features identified in the previous studies (i.e., the follower count, the friend count, the favorites count, the listed count, the tweet count, the re-tweet count, user verified, age day, follower ratio, link count, reply/mention count, and hashtag count), the authors took into account further 95 one-gram features extracted from the text of the tweets. Then, they modified and applied two stream clustering algorithms to adapt to the streaming nature of tweets, StreamKM++, and DenStream, respectively. Even if the algorithms were effective by themselves (i.e., StreamKM++ performed 99% recall and 6.4 false positive rate, while DenStream performed 99% recall and 2.2% false positive rate), the authors introduced a model that exploits the conjunction of the two approaches, reaching striking performance (i.e., 100% recall and 2.2% of false positive rate).

Evading techniques. Hand in hand with the introduction of machine learning algorithms to detect spammers inside social networks, several evading techniques have been developed, to allow spammers to deceive the classifiers and continue their activity undisturbed [2]. Early evading techniques have been studied in [136]. The authors analyzed the difficulties of using machine learning features to detect spammers and carried out an interesting analysis aiming at evaluating both the features used in the literature (24 features in total) and new features, proposed in the same study. Results showed that the newly designed features are more effective when it came to detecting Twitter spammers (even the ones that employ evasion techniques). Indeed, the new features allowed to increase the detection rate to 85%, a noteworthy improvement if compared with a detection rate of 51% (the worst existing detector evaluated) or 73% (the best existing detector evaluated).

Blacklisting. One seminal study about spammers on Twitter is discussed in [137]. Authors found that 8% of 25 million URLs posted on Twitter are related to malware, scams, or phishing. Furthermore, according to their analysis, most of the accounts that are guilty of sending spam messages have been compromised and are puppeteered by spammers. After discovering that the URLs identified were already listed on popular blacklists, as a countermeasure, the authors proposed to rely on these blacklists to detect spammers on the social platforms. However, they pointed out several limitations: (i) 90% of visitors view a page before it is blacklisted, since the procedure to blacklist a page is too slow compared to the spreading of spam;

and (ii) even by reducing the delays of the blacklisting procedure, spammers might rely on URL shortening services to obfuscate the original link and circumvent the defense mechanism.

An in-depth analysis. To analyze the behavior of spammers, together with the tools, the techniques, and the support infrastructure they boast, the authors in [138] carried out an interesting analysis over 1.1 million accounts suspended by Twitter in a period of seven months. The dataset is composed of 1.8 billion tweets, of which 80 million are related to spammer accounts. They discovered that: (i) Twitter banned 77% accounts the same day of their first tweet; and (ii) less than 9% of spammers form social relationships with Twitter users, 17% of spammers performed hijacking activities, and 52% made use of unwelcome ways to reach an audience. Furthermore, the authors described that five specific spam campaigns that were controlling approximately 145,000 accounts, each of them with a different, unique spamming strategy, have been able to persist for months.

Sock puppets

Sockpuppeting, which refers to creating fake identities to interact with other users, particularly in online discussions, may easily bring to deceptions, manipulation of the public opinion, and the vandalization of online content. Sock puppets may be semi-automated (as well as fully automated) agents whose goal is to spread lies, misinformation, and falsehood, against the target on duty. The phenomenon has been studied in the literature from several points of view.

Online deception. Knowing the reason behind online deception is a necessary step to understand the sockpuppetry phenomena. One of the studies that moved the first steps in this direction is [139]. The authors created a Web-based survey and analyzed the answers of 257 users. According to the analysis: (i) most of the users think that online deception is widespread but only one-third of them engaged in one; (ii) online deception is performed by frequent (i.e., users who regularly visit the web), young, and competent users; and (iii), the identity "play" and the privacy concerns are among the most common motivation of engaging online deception.

Detecting sockpuppets in online communities. Several studies in the literature focused on detecting sockpuppets in OSNs/discussion communities. To make an example, in [140] authors proposed an algorithm that makes use of a combination of authorship-identification techniques and link analysis to detect sockpuppets on the web. They started by proposing a new social network model in which the nodes are represented by the users, and

two nodes are connected if these users have similar attitudes with respect to the topics with which they interact. After building such a graph: (i) the edges are pruned according to the impact of the writing styles of the node, and (ii), link-based community detection is performed. A similar graph has been built in [141]. The authors introduced a similar-orientation network where each node represents a user accounts and two nodes share an edge if they have similar sentiment orientations to most topics. In this case, the authors relied on a multiple random walk module to calibrate the weight of each edge and applied community detection algorithms to detect the sockpuppets gangs within the network.

Sockpuppets vs ordinary users. In a recent work [142], the authors studied sockpuppetry across nine discussion communities to highlight the differences between sockpuppets and ordinary users. In particular, the proposed work analyzed the linguistic traits of the sockpuppets, their posting behavior, and the social network structure. The study about the linguistic traits showed that sockpuppets tend to start fewer discussions if compared with normal users, they write posts that are usually shorter, they make extensive use of personal pronouns, such as "I", and they swear more. When one person controls more than one accounts, these accounts are more likely to interact on the same discussions at the same time. Furthermore, the authors offered a detailed taxonomy of sockpuppets' behavior in online discussions. For example, sockpuppets could vary in their deceptiveness (i.e., whether and how the sockpuppets pretend to be different users) or their supportiveness (i.e., whether the sockpuppets support the same arguments backed by accounts controlled by the same user).

Vandalization. The vandalization of online content became a serious concern over the past few years. One of the websites that took the brunt of it is Wikipedia, the world's largest crowd-sourced encyclopedia. On the Wikipedia platform, every user has the opportunity to write, edit, and leave comments to articles. The registration is optional and requires only a few personal information. This led to the creation of multiple identities by malicious users for several reasons, including block evasion and block stacking. In [143] the authors carried out a case study of sockpuppets detection in Wikipedia. In particular, they made use of machine learning techniques to solve authorship attributions problems and discover who is the real puppeteer behind the documents. Wikipedia as an experimental case has been used also by [144]. The authors explained how the currently adopted methods to detect sockpuppets (or multiple account identity deception in general), mainly based on profile data and text lexical features, are inefficient for the social media environment, given the large volumes of data involved. Because of this limitation, they introduced a new method of detection that relies on nonverbal behavior, thus being computationally efficient for the

social media environment.

Sockpuppetry on the news. The impact sockpuppeting had on society has been also highlighted by an interesting article of the New York Times, called "The Hand That Controls the Sock Puppet Could Get Slapped" [145]. The article tells the story of John Mackey, the CEO of Whole Foods Market, who intervened in anonymous online discussions to promote his supermarket chain's stock.

However, it is worth noting that, although sockpuppetry could be used with malicious purposes, it may not always be the case. Indeed, according to [142], some users could create different identities to participate in different discussions and enjoy activities in different spheres of interest, without ever pretending to be other users.

Political memes

Milner in 2013 defined memes as "multimodal artefacts remixed by countless participants, employing popular culture for public commentary" [146], while, according to Dictionary.com, a meme is "a humorous image, video, piece of text, etc. that is copied (often with slight variations) and spread rapidly by Internet users"[46]. Throughout history, memes ranged from harmless and nonsensical media to dangerous and controversial ones, a transition that eventually led to endless fights, discussions, and debates, with online social media as a stage. Although, for many, they may seem harmless and superficial, memes are changing the way business operations are conducted, by shaping the popular culture while playing an extremely active role in shaping the dynamics of modern society [147].

A meme is defined as a political meme when it depicts political figures. Political memes have been used in the past to reinforce an aspect of a candidate (or, conversely, to weaken or damage an opponent) enough to change the public opinion and have a severe impact on political elections. An example is given by Ted Cruz, an American politician that, during the 2016 primaries, has been declared (by a political meme) to be the Zodiac Killer (Axelrod), the pseudonym of a renowned serial killer operating in Northern California. The political meme went viral and, according to Public Policy Polling, 40% of the voters in Florida have been influenced by the meme, 10% of them had a serious thought that Ted could have been the Zodiac Killer[47].

This episode, along with many others, makes it clear how important it is

[46] `https://www.dictionary.com/browse/meme?s=t` (Last checked August 2020)

[47] `https://en.wikipedia.org/wiki/Ted_Cruz%E2%80%93Zodiac_Killer_meme` (Last checked August 2020)

to identify memes that can potentially affect people's public opinion. Their detection has been faced by some interesting articles in the literature, that are described below.

Detection. The detection of memes before their widespread diffusion and evolution became a priority, to understand the dynamics of the society, the future trends, and to plan any strategic action. For this reason, several research studies started facing this problem by providing interesting contributions. In [148], the authors offer a detailed quantitative analysis of the global news cycle and a study of the information propagation dynamics between mainstream and social media. They introduced a framework for tracking short phrases and identified memes exhibiting rich daily variation, showing how such an approach could represent coherently the news cycle (i.e., "the daily rhythms in the news media that have long been the subject of qualitative interpretation but have never been captured accurately enough to permit actual quantitative analysis"). In particular, 1.6 million mainstream media sites and blogs have been analyzed for three months, for a total of 90 million articles analyzed. Among the many interesting discoveries, they observed a lag of 2.5 hours between the peaks of attention about a phrase in the news social media compared to the same phrase in blogs. In [149], the authors introduced Truthy, a software system implemented as a website that allows to identify and analyze political memes in Twitter, helping at detecting astroturfing, smear campaigns, and other misinformation campaigns in the context of the United States political elections. Some of the memes identified through this framework are characterized by small diffusion networks, representing the perfect moment for the identification, to avoid any subsequent diffusion.

It is a common belief that Internet memes spread virally but the evidence is often in short supply. This need for investigating the epidemic dynamics of memes has been fulfilled by [150] that, in the proposed study, analyzed the temporal dynamics and the infectious properties of 150 famous Internet memes. By analyzing several mathematical models of epidemic spread contextualized in the Internet memes domains, the author discovered that: (i) traditional compartment models with constant parameters well describe the growth and the decline patterns of Internet memes but are not able to characterize short-lived bursts of Internet meme related activities; (ii) temporal dynamics of Internet memes are accurately summarized by log-normal distributions (for 70% of the 150 memes taken into account the probability of observing a log-normal model underlying the data distribution exceeded 90%); and (iii), the majority of famous Internet memes are mostly disseminated to homogeneous online communities and OSNs.

Trolls

Trolls represent Internet users that create and post offensive, divisive, provocative messages in online communities to provoke emotional responses and sow discord. Several forms of trolling can be found on social media, from the one in which trolls make statements about political debates to the ones in which trolls provoke and insult people, resulting in harmful and life-risky situations.

Among the many categories of trolls, it is possible to find the corporate, political, and special-interested sponsored trolls. Those accounts are usually employed by governments and organizations to opportunely manipulate public opinion or to start astroturfing campaigns. Another category is represented by trolls that, without any reason, rage against users, generating cyberbullying situations. Those kinds of trolls are usually found in social media, a habitat that several studies have defined much more dangerous than phone calls and messages when it comes to cyberbullying. According to the studies, victims of cyberbullying live several feelings (e.g., depression, anger, overwhelming, powerless, humiliation), and are almost twice as likely to attempt suicide[48].

Given these premises, the task of identifying troll accounts before they start generating chaos on social media is a priority, and the literature can boast several contributions that move the first steps in that direction.

Sentiment analysis. Since trolls represent users that usually post obscene, negative, and inflammatory comments on the Web, authors in [151] proposed a detection approach relying on the sentiment analysis of the messages. In particular, the authors decided to focus on three attributes of trolls: (i) repetitiveness, the trolls use to send a large number of messages; (ii) destructiveness, the trolls use to express negative sentiments to create chaos in the online discussion; and (iii), deceptiveness, the trolls messages may be deceptive to sow discord. Starting from a limited training set (i.e., 20 users), the binary classification through a Ranking SVM classifier reached a 60% generalized Received Operating Characteristic (ROC). However, during the experiment, the authors were relying on a sentiment analysis model trained on a dataset written in standard English, while the messages in the forums were often written in colloquial English (from Singapore). They improved the performance by applying domain adaptation techniques to the sentiment analysis, finally reaching a ROC of 78%.

Combined features. According to some studies in the literature, an effective solution to the challenge of detecting trolls in OSNs consists of the

[48] `https://www.warrington-worldwide.co.uk/2020/04/10/the-effects-of-internet-trolling/#:~:text=Some%20of%20the%20feelings%20that,the%20person%20disinterest%20in%20life` (Last checked August 2020)

integration of different classes of features. In [152], the authors proposed a holistic system, called TrollPacifier, that integrates the user-level features with the ones derived from the analysis of texts and the local social graph. Six groups of features have been identified in this study: (i) writing style-based features; (ii) sentiment-based features; (iii) behavior-based features; (iv) social interaction-based features; (v) linked media-based features; and (vi), publication time-based features, respectively. The introduced holistic classifier, applied on a dataset composed of 500 troll and 500 non-troll users, showed an accuracy of 95.5%.

Ranking. Some other studies in the literature faced the problem from a different perspective. Their goal is to introduce ranking methods on the OSN that will eventually cause automatic isolation of malicious accounts. For example, the authors in [153] proposed a mechanism to compute a trustworthiness ranking of users, aiming at demoting malicious users in the ranking, thus avoiding them gaining a high reputation in the network. The idea is to allow the propagation of both positive and negative opinions about the users, which will eventually influence their global trust score. The proposed method, called PolarityTrust, has been tested by carrying out different experiments: (i) by using a real-world dataset extracted from Slashdot.org; (ii) by relying on a set of randomly generated graphs; and (iii) by using a combination of a real-world dataset and generation techniques. Another example is given by [154] that, by exploiting natural language information of the text, defines the level of "trollness" of each post and classify the authors accordingly.

Cyberbullying. To fight the cyberbullying phenomenon that, in many cases, has led to disastrous consequences in the real life, authors in [155] proposed a supervised machine learning approach for detecting troll profiles in the Twitter OSN. The assumption made in this study is that the real person behind a troll profile will follow, with the personal profile, the troll profile, to stay updated on the activity surrounding the latter. The authors relied on four features: the text of the tweet, the time of publication of the tweet, its language, and its geo-position, respectively, and applied several supervised learning classifiers, including Random Forest, J48 (i.e., a type of decision tree classifier), K-Nearest Neighbor, and Sequential Minimal Optimization (SMO). The maximum accuracy, when applied to a dataset composed of 1900 tweets written by 19 users, has been reached by the SMO and the decision trees classifier, with 68.48% and 66.48%, respectively. The proposed methodology has been applied in practice in one elementary school in the city of Bilbao (Spain). The authors have been able to attribute the authorship of the offensive messages to the culprits that, frightened by possible consequences, voluntarily confessed and apologized.

Other machine learning approaches. Other studies in the literature exploit machine learning techniques to distinguish trolls from normal users. To make an example, in [156], the authors developed a machine learning model to predict whether a Twitter account is a Russian troll. Starting from a dataset of 170,000 accounts and relying on both behavioral and linguistic features, they demonstrated that it is possible to distinguish a troll with an AUC of 98.9% and a precision of 78.5%. The model has been subsequently applied in the wild, specifically to out-of-sample accounts, and led to the discovery of approximately 2.6% of mentions of top journalists occupied by Russian trolls. Furthermore, according to additional analysis, the author highlighted that these trolls are not merely software-controlled automated agents, and are able to manage their online identities in complex ways. In [157], instead, the authors proposed an interesting approach to analyze both user identities and their social roles in OSNs. In particular, they developed a new text distance metric (i.e., the time-sensitive semantic edit distance) useful to classify the social roles of trolls based on the traces (e.g., tweets) they leave on the social network. To develop and evaluate their method, the authors characterized the Russian trolls that attempted to manipulate public opinion during the 2016 United States presidential election, to understand their tactics based on their social roles and strategies. As a result, this study shows patterns in the similarities of tweets the Russian trolls left behind while posting online, providing useful insights into Russian troll activities both during and after the aforementioned election.

Lastly, in [158], the authors proposed an easy yet effective method to recognize opinion manipulation trolls on the Web. They assume that a user who is called "troll" by several people is likely to be one and classify them as such, reaching from 82 to 95% of accuracy. A subset of the authors, one year later [159], proposed a classifier able to distinguish "paid trolls" from "mentioned trolls". "Paid trolls" are defined as trolls that have been revealed from leaked reputation management contracts, while "mentioned trolls" are defined as trolls that have been called such by different people. The classifier was able to distinguish a paid troll by a mentioned troll with an accuracy of 81-82%.

1.5 Open Issues and Future Directions

The advent of Artificial Intelligence, defined "the new electricity" by Dr. Andrew Ng, VP & Chief Scientist of Baidu and Adjunct Professor at Stanford University, is completely revolutionizing the way we work, think, and live. Institutions, companies, universities, and people in general are heavily relying on the innovation provided by Artificial Intelligence to automate tasks and achieve results that were unthinkable until a few years ago. Unfor-

tunately, the undeniable benefits of this outstanding technology can also be exploited for malicious purposes: to create increasingly credible fake content or to automate fishing and disinformation campaigns on the web, to name a few. According to several experts, the support Artificial Intelligence may bring to the creation of misinformation campaigns, might further amplify a problem that is already difficult to manage. Indeed, spreading misinformation is way too easier than defending against it, and the support provided by Artificial Intelligence could even widen this gap.

One case in point is the experiment conducted by two data scientists from ZeroFOX, in which they implemented an artificial hacker able to compose and distribute more phishing tweets than humans, with a substantially better conversion rate. This experiment has shown how, with the support provided by Artificial Intelligence, it is possible to effectively reach a wider audience. Another fascinating – yet scary – innovative example, made possible thanks to Artificial Intelligence, is given by deepfakes.

Definitions

Deepfakes. Deepfakes are defined as synthetic media (e.g., video, audio, text) in which, by relying on AI algorithms, a person has been replaced with someone else in a way that makes the multimedia resource look authentic.

Although this technology is being primarily used to make "fun videos" (e.g., put the face of the actor Nicolas Cage literally everywhere), there are multiple ways to exploit it for malicious (and extremely dangerous) purposes. The technology brings people to believe that something is real when it is not and, if placed in the wrong hands, could potentially endanger the credibility of individuals and allow a faster spread of misinformation [2, 9]. With this technology, everyone can make anyone saying anything at any time, and the authoritativeness of information is inevitably compromised.

"If I do not see it I do not believe it", Thomas the Apostle said incredulously in the Gospel of John, when he refused to believe in the resurrection of Jesus. With the advent of deepfakes, unfortunately, seeing is no longer enough.

1.5.1 New Directions

A detailed analysis of the state of the art allowed us to understand the current open issues, the existing countermeasures, as well as their limitations when it comes to countering these new, AI-enabled threats. In the following, we recommend some valuable directions we think are worth investigating, to stop (or at least mitigate) the wave of disinformation and donate a new, reliable, truthful face to the Web.

- *Coordinated Inauthentic Behavior.* Coordinated Inauthentic Behavior is a term coined by Facebook and refers to a situation in which "groups of pages or people work together to mislead others about who they are or what they are doing"[49]. Valuable studies and OSNs, such as Facebook and Twitter, started focusing on the detection of coordinated inauthentic behavior, since devising techniques for spotting such behaviors is likely to provide better results when compared with the detection of individual malicious accounts. An interesting and potentially fruitful research direction thus consists of identifying suspicious coordination independently of the nature of individual accounts. However, in order to reach a valuable accuracy, several challenges have to be faced, including scalability problems of the group-based detectors and the intrinsic fuzziness of inauthentic coordination [2].

- *Proactiveness rather than reactiveness.* Given the quick propagation of low-quality information and the simplicity in automating the creation of fake accounts, it may be smart to push research towards proactive defenses rather than reactive ones [2, 160]. Indeed, it is very difficult to efficiently react once the propagation started.

- *Technology and policies.* At the current level of technology, machine learning makes it possible to create bots that emulate human behavior, think processes and strategies, in such a way as to be impossible to be distinguished from humans in several contexts [2]. The current machine learning techniques, once limited to responding to inputs from the users (e.g., in online reservations), are now capable of producing new synthetic content from scratch, using generative machine learning algorithms. The authenticity of these artfully created multimedia resources may be currently checked by digital forensics experts, who are able to detect the artifact by analyzing the lack of camera-induced imperfections, but the approach does not scale. [161]. Many researchers think that fighting fire with fire might be the ideal solution, i.e., respond to automatic systems for propagation and diffusion with automatic systems to prevent/stop it. However, according to the analysis in [161], technological solutions alone cannot address the challenges of fake content emulating human behavior, and there is the need of "developing public policy, legal, and normative frameworks for managing the malicious applications of technology in conjunction with efforts to refine it". Some of the technological defenses proposed in the aforementioned study are described in the following resource box.

[49] `https://about.fb.com/news/2018/12/inside-feed-coordinated-inauthentic-behavior/` (Last checked August 2020)

Resources

Technological defenses proposed by [161]:

- **Automate the process of digital forensics [162].** A detector may be able to detect imperfections on the multimedia content to categorize it as "fake" [163, 164]. However, even if an extremely accurate detector is introduced, the GAN technology would allow creating new content to avoid the detection by construction, thus placing a tie as the maximum aspiration for the defender.
- **Improve the provenance of human forms of digital communication.** The goal is to identify deepfakes as content that was digitally synthesized. In this way, every camera should be equipped with a tamper-proof cryptographic signing key, necessary to sign the picture it takes, and every generator would not be able to sign the produced fake content. Despite the obvious advantages, the creation and the distribution of the keys, as well as their authentication, make the solution logistically hard.
- **Total accountability.** If a public figure is concerned about fake videos and wants to protect herself, she could continuously record herself with a tamper-proof camera to demonstrate its estrangement from any facts created in a fake video concerning her. As the authors mention, in this case, the cure may be worse than the disease, since it would bring to striking privacy leakages.

- *Multimodal integration of features.* The studies in the literature dealing with the detection of low-quality information or malicious actors often rely on a feature extraction phase, in which several information are gathered to identify any suspicious pattern. Such information may be extracted from several sources: e.g., metadata of posts, the text of posts, profile information, the network graph, and possibly others. An interesting future direction consists of the research of efficient multimodal integrations of features to reach higher accuracy in shorter times [9].

Bibliography

[1] "How to escape your political bubble for a clearer view." `https://www.nytimes.com/2017/03/03/arts/the-battle-over-your-political-bubble.html`, (Last checked August 2020).

[2] S. Cresci, "A decade of social bot detection," *Communications of the ACM (forthcoming)*, 2020.

[3] A. Marwick and R. Lewis, "Media manipulation and disinformation online," *New York: Data & Society Research Institute*, 2017.

[4] J. Yan, "Bot, cyborg and automated turing test," in *The 2006 International Workshop on Security Protocols*, pp. 190–197, Springer, 2006.

[5] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Detecting automation of twitter accounts: Are you a human, bot, or cyborg?," *IEEE Transactions on Dependable and Secure Computing*, vol. 9, pp. 811–824, Nov 2012.

[6] E. Ferrara, "The history of digital spam," *Communications of the ACM*, vol. 62, no. 8, pp. 82–91, 2019.

[7] B. Waugh, M. Abdipanah, O. Hashemi, S. A. Rahman, and D. M. Cook, "The influence and deception of Twitter: The authenticity of the narrative and slacktivism in the Australian electoral process," in *The 14th Australian Information Warfare Conference (AIWC'13)*, 2013.

[8] J. Ratkiewicz, M. D. Conover, M. Meiss, B. Gonçalves, A. Flammini, and F. M. Menczer, "Detecting and tracking political abuse in social media," in *The 5th International AAAI Conference on Weblogs and Social Media (ICWSM'11)*, AAAI, 2011.

[9] G. Da San Martino, S. Cresci, A. Barrón-Cedeño, S. Yu, R. Di Pietro, and P. Nakov, "A survey on computational propaganda detection," in *The 29th International Joint Conference on Artificial Intelligence (IJCAI'20)*, 2020.

[10] N. Persily, "The 2016 US Election: Can democracy survive the Internet?," *Journal of democracy*, vol. 28, no. 2, pp. 63–76, 2017.

[11] M. Mazza, S. Cresci, M. Avvenuti, W. Quattrociocchi, and M. Tesconi, "Rtbust: exploiting temporal patterns for botnet detection on twitter," in *The 11th International Conference on Web Science (WebSci'19)*, pp. 183–192, ACM, 2019.

[12] A. M. Guess, M. Lerner, B. Lyons, J. M. Montgomery, B. Nyhan, J. Reifler, and N. Sircar, "A digital media literacy intervention increases discernment between mainstream and false news in the united states and india," *Proceedings of the National Academy of Sciences*, 2020.

[13] D. K. Flaherty, "The vaccine-autism connection: a public health crisis caused by unethical medical practices and fraudulent science," *Annals of Pharmacotherapy*, vol. 45, no. 10, pp. 1302–1304, 2011.

[14] C. A. Borella and D. Rossinelli, "Fake news, immigration, and opinion polarization," *SocioEconomic Challenges*, 2017.

[15] C. Garwood, *Flat earth: The history of an infamous idea*. Pan Macmillan, 2008.

[16] D. E. Allen and M. McAleer, "Fake news and indifference to scientific fact: President trump's confused tweets on global warming, climate change and weather," *Scientometrics*, vol. 117, no. 1, pp. 625–629, 2018.

[17] S. Van der Linden, A. Leiserowitz, S. Rosenthal, and E. Maibach, "Inoculating the public against misinformation about climate change," *Global Challenges*, vol. 1, no. 2, p. 1600008, 2017.

[18] M. Gabielkov, A. Ramachandran, A. Chaintreau, and A. Legout, "Social clicks: What and who gets read on twitter?," *ACM SIGMETRICS Performance Evaluation Review*, vol. 44, no. 1, pp. 179–192, 2016.

[19] R. M. Bond, C. J. Fariss, J. J. Jones, A. D. Kramer, C. Marlow, J. E. Settle, and J. H. Fowler, "A 61-million-person experiment in social influence and political mobilization," *Nature*, vol. 489, no. 7415, p. 295, 2012.

[20] C. A. Bail, B. Guay, E. Maloney, A. Combs, D. S. Hillygus, F. Merhout, D. Freelon, and A. Volfovsky, "Assessing the russian internet research agency's impact on the political attitudes and behaviors of american twitter users in late 2017," *Proceedings of the National Academy of Sciences*, vol. 117, no. 1, pp. 243–250, 2020.

[21] M. Rueda, "2012's biggest social media blunders in latam politics." https://abcnews.go.com/ABC_Univision/ABC_Univision/2012s-biggest-social-media-blunders-latin-american-politics/story?id=18063022, Last checked August 2020.

[22] T. Filer and R. Fredheim, "Popular with the robots: Accusation and automation in the argentine presidential elections, 2015," *International Journal of Politics, Culture, and Society*, vol. 30, no. 3, pp. 259–274, 2017.

[23] T. Peel, "The coalition's twitter fraud and deception." https://independentaustralia.net/politics/politics-display/the-coalitions-twitter-fraud-and-deception,5660, Last checked August 2020.

[24] E. Kusen and M. Strembeck, "An analysis of the twitter discussion on the 2016 austrian presidential elections," *arXiv preprint arXiv:1707.09939*, 2017.

[25] H. Ellyatt, "Us far-right activists, wikileaks and bots help amplify macron leaks." https://www.cnbc.com/2017/05/07/macron-email-leaks-far-right-wikileaks-twitter-bots.htm, Last checked August 2020.

[26] R. Brandom, "Emails leaked in 'massive hacking attack' on french presidential campaign." https://www.theverge.com/2017/5/5/15564532/macron-email-leak-russia-hacking-campaign-4chan, Last checked August 2020.

[27] S. Almasy, "Emmanuel macron's french presidential campaign hacked." https://edition.cnn.com/2017/05/05/europe/france-election-macron-hack-allegation/index.html, Last checked August 2020.

[28] E. Ferrara, "Disinformation and social bot operations in the run up to the 2017 french presidential election," *First Monday*, vol. 22, no. 8, 2017.

[29] C. Desigaud, P. N. Howard, S. Bradshaw, B. Kollanyi, and G. Bolsover, "Junk news and bots during the french presidential election: What are french voters sharing over twitter in round two?," tech. rep., COMPROP Data Memo, 2017.

[30] F. Brachten, S. Stieglitz, L. Hofeditz, K. Kloppenborg, and A. Reimann, "Strategies and influence of social bots in a 2017 german state election-a case study on twitter," *arXiv preprint arXiv:1710.07562*, 2017.

[31] K. Kupferschmidt, "Bot-hunters eye mischief in german election," *Science*, vol. 357, no. 6356, pp. 1081–1082, 2017.

[32] F. Morstatter, Y. Shao, A. Galstyan, and S. Karunasekera, "From alt-right to alt-rechts: Twitter analysis of the 2017 german federal election," in *Companion Proceedings of the The Web Conference 2018 (WWW Companion'18)*, pp. 621–628, IW3C2, 2018.

[33] T. R. Keller and U. Klinger, "Social bots in election campaigns: Theoretical, empirical, and methodological implications," *Political Communication*, vol. 36, no. 1, pp. 171–189, 2019.

[34] A. Applebaum, P. Pomerantsev, M. Smith, and C. Colliver, "'make germany great again': Kremlin, alt-right, and international influences in the 2017 german elections," *London School of Economics*, 2017.

[35] A. Vogt, "Hot or bot? italian professor casts doubt on politician's twitter popularity." `https://www.theguardian.com/world/2012/jul/22/bot-italian-politician-twitter-grillo`, Last checked August 2020.

[36] N. Squires, "Human or 'bot'? doubts over italian comic beppe grillo's twitter followers." `https://www.telegraph.co.uk/technology/twitter/9421072/Human-or-bot-Doubts-over-Italian-comic-Beppe-Grillos-Twitter-followers.html`, Last checked August 2020.

[37] C. Albanese, "Now bots are trying to help populists win italy's election." `https://www.bloomberg.com/news/articles/2018-02-19/now-bots-are-trying-to-help-populists-win-italy-s-election`, Last checked August 2020.

[38] DFRLab, "#electionwatch: Italy's self-made bots." `https://medium.com/dfrlab/electionwatch-italys-self-made-bots-200e2e268d0e`, Last checked August 2020.

[39] TheLocal, "Facebook shuts down more than 20 'fake news' pages in italy." `https://www.thelocal.it/20190513/facebook-shuts-down-more-than-20-fake-news-pages-in-italy`, Last checked August 2020.

[40] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race," in *The 26th International Conference on World Wide Web Companion (WWW'17 Companion)*, pp. 963–972, IW3C2, 2017.

[41] M. Orcutt, "Twitter mischief plagues mexico's election." `https://www.technologyreview.com/s/428286/twitter-mischief-plagues-mexicos-election/`, Last checked August 2020.

[42] M. Glowacki, V. Narayanan, S. Maynard, G. Hirsch, B. Kollanyi, L. Neudert, V. Barash, *et al.*, "News and political information consumption in mexico: Mapping the 2018 mexican presidential election on twitter and facebook," *The Computational Propaganda Project*, 2018.

[43] J. Robertson, M. Riley, and A. Willis, "How to hack an election." `https://www.bloomberg.com/features/2016-how-to-hack-an-election/`, Last checked August 2020.

[44] E. Gallagher, "Mexican botnet dirty wars: Bots are waging a dirty war in mexican social media." `https://www.youtube.com/watch?v=I3D3iIZGSt8`, Last checked August 2020.

[45] T. Guardian, "'i've had enough', says mexican attorney general in missing students gaffe." `https://www.theguardian.com/world/2014/nov/09/protests-flare-in-mexico-after-attorney-generals-enough-im-tired-remarks`, Last checked August 2020.

[46] P. Suárez-Serrato, M. E. Roberts, C. Davis, and F. Menczer, "On the influence of social bots in online protests," in *Social Informatics*, (Cham), pp. 269–278, Springer International Publishing, 2016.

[47] M. Stella, E. Ferrara, and M. De Domenico, "Bots sustain and inflate striking opposition in online social systems," *arXiv preprint arXiv:1802.07292*, 2018.

[48] D. Stukal, S. Sanovich, R. Bonneau, and J. A. Tucker, "Detecting bots on russian political twitter," *Big data*, vol. 5, no. 4, pp. 310–324, 2017.

[49] S. Zannettou, B. Bradlyn, E. De Cristofaro, G. Stringhini, and J. Blackburn, "Characterizing the use of images by state-sponsored troll accounts on twitter," *arXiv preprint arXiv:1901.05997*, 2019.

[50] S. Zannettou, T. Caulfield, W. Setzer, M. Sirivianos, G. Stringhini, and J. Blackburn, "Who let the trolls out?: Towards understanding state-sponsored trolls," in *Proceedings of the 10th ACM Conference on Web Science*, pp. 353–362, ACM, 2019.

[51] L. G. Stewart, A. Arif, and K. Starbird, "Examining trolls and polarization with a retweet network," in *Proc. ACM WSDM, Workshop on Misinformation and Misbehavior Mining on the Web*, 2018.

[52] S. Shane and V. Goel, "Fake russian facebook accounts bought $100,000 in political ads." https://www.nytimes.com/2017/09/06/technology/facebook-russian-political-ads.html, Last checked August 2020.

[53] R. Dutt, A. Deb, and E. Ferrara, ""senator, we sell ads": Analysis of the 2016 russian facebook ads campaign," in *International Conference on Intelligent Information Technologies*, pp. 151–168, Springer, 2018.

[54] E. Poyrazlar, "Turkey's leader bans his own twitter bot army." https://www.vocativ.com/world/turkey-world/turkeys-leader-nearly-banned-twitter-bot-army/, Last checked August 2020.

[55] S. Hegelich and D. Janetzko, "Are social bots on twitter political actors? empirical evidence from a ukrainian social botnet.," in *ICWSM*, pp. 579–582, 2016.

[56] P. N. Howard and B. Kollanyi, "Bots,#strongerin, and#brexit: computational propaganda during the uk-eu referendum," *Available at SSRN 2798311*, 2016.

[57] M. T. Bastos and D. Mercea, "The brexit botnet and user-generated hyperpartisan news," *Social Science Computer Review*, vol. 37, no. 1, pp. 38–54, 2019.

[58] C. Llewellyn, L. Cram, R. L. Hill, and A. Favero, "For whom the bell trolls: Shifting troll behaviour in the twitter brexit debate," *JCMS: Journal of Common Market Studies*, 2019.

[59] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of economic perspectives*, vol. 31, no. 2, pp. 211–36, 2017.

[60] A. Guess, B. Nyhan, and J. Reifler, "Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 us presidential campaign," *European Research Council*, vol. 9, 2018.

[61] A. Bessi and E. Ferrara, "Social bots distort the 2016 us presidential election online discussion," *First Monday*, vol. 21, no. 11-7, 2016.

[62] C. Shao, G. L. Ciampaglia, O. Varol, A. Flammini, and F. Menczer, "The spread of fake news by social bots," *arXiv preprint arXiv:1707.07592*, pp. 96–104, 2017.

[63] W. Samuel and H. Phil, "Bots unite to automate the presidential election." https://www.wired.com/2016/05/twitterbots-2/, Last checked August 2020.

[64] B. Ryan, "Nearly half of donald trump's twitter followers are fake accounts and bots." `https://www.newsweek.com/donald-trump-twitter-followers-fake-617873`, Last checked August 2020.

[65] A. Fourney, M. Z. Racz, G. Ranade, M. Mobius, and E. Horvitz, "Geographic and temporal trends in fake news consumption during the 2016 us presidential election," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 2071–2074, ACM, 2017.

[66] E. Mustafaraj and P. T. Metaxas, "From obscurity to prominence in minutes: Political speech and real-time search," 2010.

[67] A. Badawy, E. Ferrara, and K. Lerman, "Analyzing the digital traces of political manipulation: the 2016 russian interference twitter campaign," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 258–265, IEEE, 2018.

[68] A. Badawy, K. Lerman, and E. Ferrara, "Who falls for online political manipulation?," in *Companion Proceedings of The 2019 World Wide Web Conference*, pp. 162–168, ACM, 2019.

[69] M. Jensen, "Russian trolls and fake news: Information or identity logics?," *Journal of International Affairs*, vol. 71, no. 1.5, pp. 115–124, 2018.

[70] M. Forelle, P. Howard, A. Monroy-Hernández, and S. Savage, "Political bots and the manipulation of public opinion in venezuela," *arXiv preprint arXiv:1507.07109*, 2015.

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

[72] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "Dna-inspired online behavioral modeling and its application to spambot detection," *IEEE Intelligent Systems*, vol. 31, no. 5, pp. 58–64, 2016.

[73] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "Social fingerprinting: detection of spambot groups through dna-inspired behavioral modeling," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 4, pp. 561–576, 2017.

[74] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "Exploiting digital dna for the analysis of similarities in twitter be-

haviours," in *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 686–695, IEEE, 2017.

[75] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "Emergent properties, models, and laws of behavioral similarities within groups of twitter users," *Computer Communications*, vol. 150, pp. 47–61, 2020.

[76] A. Spangher, G. Ranade, B. Nushi, A. Fourney, and E. Horvitz, "Analysis of strategy and spread of russia-sponsored content in the us in 2017," *arXiv preprint arXiv:1810.10033*, 2018.

[77] S. Zannettou, T. Caulfield, E. De Cristofaro, M. Sirivianos, G. Stringhini, and J. Blackburn, "Disinformation warfare: Understanding state-sponsored trolls on twitter and their influence on the web," in *Companion Proceedings of The 2019 World Wide Web Conference*, pp. 218–226, ACM, 2019.

[78] B. S. Bello and R. Heckel, "Analyzing the behaviour of twitter bots in post brexit politics,"

[79] O. Solon, "Facebook's fake news: Mark zuckerberg rejects 'crazy idea' that it swayed voters." `https://www.theguardian.com/technology/2016/nov/10/facebook-fake-news-us-election-mark-zuckerberg-donald-trump`, Last checked August 2020.

[80] R. Max, "Donald trump won because of facebook." `https://nymag.com/intelligencer/2016/11/donald-trump-won-because-of-facebook.html`, Last checked August 2020.

[81] C. Dewey, "Facebook fake-news writer: 'i think donald trump is in the white house because of me'." `https://www.washingtonpost.com/news/the-intersect/wp/2016/11/17/facebook-fake-news-writer-i-think-donald-trump-is-in-the-white-house-because-of-me/`, Last checked August 2020.

[82] N. Mele, D. Lazer, M. Baum, N. Grinberg, L. Friedland, K. Joseph, W. Hobbs, and C. Mattsson, "Combating fake news: An agenda for research and action," *Di https://www. hks. harvard. edu/publications/combating-fake-news-agenda-research-and-action (Retrieved October 17, 2018)*, 2017.

[83] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian, "Novel visual and statistical image features for microblogs news verification," *IEEE transactions on multimedia*, vol. 19, no. 3, pp. 598–608, 2016.

[84] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th international conference on World wide web*, pp. 675–684, 2011.

[85] S. Vosoughi, M. Mohsenvand, and D. Roy, "Rumor gauge: Predicting the veracity of rumors on twitter," *ACM transactions on knowledge discovery from data (TKDD)*, vol. 11, no. 4, pp. 1–36, 2017.

[86] N. Hassan, F. Arslan, C. Li, and M. Tremayne, "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*, pp. 1803–1812, 2017.

[87] G. Karadzhov, P. Nakov, L. Màrquez, A. Barrón-Cedeño, and I. Koychev, "Fully automated fact checking using external sources," *arXiv preprint arXiv:1710.00341*, 2017.

[88] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[89] G. Pennycook and D. G. Rand, "Fighting misinformation on social media using crowdsourced judgments of news source quality," *Proceedings of the National Academy of Sciences*, vol. 116, no. 7, pp. 2521–2526, 2019.

[90] M. R. Pinto, Y. O. de Lima, C. E. Barbosa, and J. M. de Souza, "Towards fact-checking through crowdsourcing," in *2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pp. 494–499, IEEE, 2019.

[91] S. M. Mohammad, P. Sobhani, and S. Kiritchenko, "Stance and sentiment in tweets," *ACM Transactions on Internet Technology (TOIT)*, vol. 17, no. 3, pp. 1–23, 2017.

[92] Y. Yamaguchi, T. Takahashi, T. Amagasa, and H. Kitagawa, "Turank: Twitter user ranking based on user-tweet graph analysis," in *Web Information Systems Engineering – WISE 2010* (L. Chen, P. Triantafillou, and T. Suel, eds.), (Berlin, Heidelberg), pp. 240–253, Springer Berlin Heidelberg, 2010.

[93] B. Rath, W. Gao, J. Ma, and J. Srivastava, "From retweet to believability: Utilizing trust to identify rumor spreaders on twitter," in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pp. 179–186, 2017.

[94] J. Zhang, J. Tang, and J. Li, "Expert finding in a social network," in *International Conference on Database Systems for Advanced Applications*, pp. 1066–1069, Springer, 2007.

[95] A. Bozzon, M. Brambilla, S. Ceri, M. Silvestri, and G. Vesci, "Choosing the right crowd: expert finding in social networks," in *Proceedings of the 16th International Conference on Extending Database Technology*, pp. 637–648, 2013.

[96] R. M. Tripathy, A. Bagchi, and S. Mehta, "A study of rumor control strategies on social networks," in *Proceedings of the 19th ACM international conference on Information and knowledge management*, pp. 1817–1820, 2010.

[97] N. P. Nguyen, G. Yan, M. T. Thai, and S. Eidenbenz, "Containment of misinformation spread in online social networks," in *Proceedings of the 4th Annual ACM Web Science Conference*, pp. 213–222, 2012.

[98] T. Mitra and E. Gilbert, "Credbank: A large-scale social media corpus with associated credibility annotations," in *Ninth International AAAI Conference on Web and Social Media*, 2015.

[99] W. Y. Wang, "" liar, liar pants on fire": A new benchmark dataset for fake news detection," *arXiv preprint arXiv:1705.00648*, 2017.

[100] E. Tacchini, G. Ballarin, M. L. Della Vedova, S. Moret, and L. de Alfaro, "Some like it hoax: Automated fake news detection in social networks," *arXiv preprint arXiv:1704.07506*, 2017.

[101] G. C. Santia and J. R. Williams, "Buzzface: A news veracity dataset with facebook user commentary and egos," in *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[102] J. Golbeck, M. Mauriello, B. Auxier, K. H. Bhanushali, C. Bonk, M. A. Bouzaghrane, C. Buntain, R. Chanduka, P. Cheakalos, J. B. Everett, *et al.*, "Fake news vs satire: A dataset and analysis," in *Proceedings of the 10th ACM Conference on Web Science*, pp. 17–21, ACM, 2018.

[103] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "Fakenewsnet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media," *arXiv preprint arXiv:1809.01286*, 2018.

[104] A. Pathak and R. Srihari, "BREAKING! presenting fake news corpus for automated fact checking," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, (Florence, Italy), pp. 357–362, Association for Computational Linguistics, July 2019.

[105] F. K. A. Salem, R. Al Feel, S. Elbassuoni, M. Jaber, and M. Farah, "Fa-kes: A fake news dataset around the syrian war," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 13, pp. 573–582, 2019.

[106] F. Torabi Asr and M. Taboada, "Big data and quality data for fake news and misinformation detection," *Big Data & Society*, vol. 6, no. 1, p. 2053951719843310, 2019.

[107] N. Abokhodair, D. Yoo, and D. W. McDonald, "Dissecting a social botnet: Growth, content and influence in twitter," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pp. 839–851, ACM, 2015.

[108] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini, "Online human-bot interactions: Detection, estimation, and characterization," in *Eleventh international AAAI conference on web and social media*, 2017.

[109] R. J. Oentaryo, A. Murdopo, P. K. Prasetyo, and E.-P. Lim, "On profiling bots in social media," in *International Conference on Social Informatics*, pp. 92–109, Springer, 2016.

[110] N. Agarwal, S. Jabin, S. Z. Hussain, *et al.*, "Analyzing real and fake users in facebook network based on emotions," in *2019 11th International Conference on Communication Systems & Networks (COMSNETS)*, pp. 110–117, IEEE, 2019.

[111] R. Plutchik, "Emotions: A general psychoevolutionary theory," *Approaches to emotion*, vol. 1984, pp. 197–219, 1984.

[112] J. Echeverrìa, E. De Cristofaro, N. Kourtellis, I. Leontiadis, G. Stringhini, and S. Zhou, "Lobo: Evaluation of generalization deficiencies in twitter bot classifiers," in *The 34th Annual Computer Security Applications Conference (ACSAC'18)*, pp. 137–146, ACM, 2018.

[113] N. Chavoshi, H. Hamooni, and A. Mueen, "Identifying correlated bots in twitter," in *International Conference on Social Informatics*, pp. 14–21, Springer, 2016.

[114] A. Anwar and U. Yaqub, "Bot detection in twitter landscape using unsupervised learning," in *The 21st Annual International Conference on Digital Government Research*, pp. 329–330, 2020.

[115] C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer, "Botornot: A system to evaluate social bots," in *Proceedings of the 25th International Conference Companion on World Wide Web*, pp. 273–274, IW3C2, 2016.

[116] N. Chavoshi, H. Hamooni, and A. Mueen, "On-demand bot detection and archival system," in *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 183–187, IW3C2, 2017.

[117] V. Subrahmanian, A. Azaria, S. Durst, V. Kagan, A. Galstyan, K. Lerman, L. Zhu, E. Ferrara, A. Flammini, and F. Menczer, "The darpa twitter bot challenge," *Computer*, vol. 49, no. 6, pp. 38–46, 2016.

[118] F. Rangel and P. Rosso, "Overview of the 7th author profiling task at pan 2019: Bots and gender profiling in twitter," in *Proceedings of the CEUR Workshop, Lugano, Switzerland*, pp. 1–36, 2019.

[119] DFRLab, "#botspot,: Twelve ways to spot a bot." `https://medium.com/dfrlab/botspot-twelve-ways-to-spot-a-bot-aedc7d9c110c`, Last checked August 2020.

[120] M. Conti, R. Poovendran, and M. Secchiero, "Fakebook: Detecting fake profiles in on-line social networks," in *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pp. 1071–1078, IEEE Computer Society, 2012.

[121] Q. Cao, M. Sirivianos, X. Yang, and T. Pregueiro, "Aiding the detection of fake accounts in large scale social online services," in *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*, pp. 15–15, USENIX Association, 2012.

[122] M. La Morgia, A. Mei, S. Raponi, and J. Stefa, "Time-zone geolocation of crowds in the dark web," in *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)*, pp. 445–455, IEEE, 2018.

[123] M. La Morgia, A. Mei, E. Nemmi, S. Raponi, and J. Stefa, "Nationality and geolocation-based profiling in the dark (web)," *IEEE Transactions on Services Computing*, 2019.

[124] S. Gurajala, J. S. White, B. Hudson, and J. N. Matthews, "Fake twitter accounts: profile characteristics obtained using an activity-based pattern detection approach," in *Proceedings of the 2015 International Conference on Social Media & Society*, p. 9, ACM, 2015.

[125] D. Ramalingam and V. Chinnaiah, "Fake profile detection techniques in large-scale online social networks: A comprehensive review," *Computers & Electrical Engineering*, vol. 65, pp. 165–177, 2018.

[126] S. Adikari and K. Dutta, "Identifying fake profiles in linkedin.," in *PACIS*, p. 278, 2014.

[127] J. Haikarainen, "Astroturfing as a global phenomenon," 2014.

[128] F. B. Keller, D. Schoch, S. Stier, and J. Yang, "Political astroturfing on twitter: How to coordinate a disinformation campaign," *Political Communication*, vol. 37, no. 2, pp. 256–280, 2020.

[129] J. Zhang, D. Carpenter, and M. Ko, "Online astroturfing: A theoretical perspective," 2013.

[130] T. Chen, N. H. Alallaq, W. Niu, Y. Wang, X. Bai, J. Liu, Y. Xiang, T. Wu, and J. Liu, "A hidden astroturfing detection approach base on emotion analysis," in *International Conference on Knowledge Science, Engineering and Management*, pp. 55–66, Springer, 2017.

[131] S. Mahbub, E. Pardede, A. Kayes, and W. Rahayu, "Controlling astroturfing on the internet: a survey on detection techniques and research challenges," *International Journal of Web and Grid Services*, vol. 15, no. 2, pp. 139–158, 2019.

[132] A. H. Wang, "Detecting spam bots in online social networking sites: a machine learning approach," in *IFIP Annual Conference on Data and Applications Security and Privacy*, pp. 335–342, Springer, 2010.

[133] G. Stringhini, C. Kruegel, and G. Vigna, "Detecting spammers on social networks," in *Proceedings of the 26th annual computer security applications conference*, pp. 1–9, ACM, 2010.

[134] M. Singh, D. Bansal, and S. Sofat, "Who is who on twitter–spammer, fake or compromised account? a tool to reveal true identity in real-time," *Cybernetics and Systems*, vol. 49, no. 1, pp. 1–25, 2018.

[135] Z. Miller, B. Dickinson, W. Deitrick, W. Hu, and A. H. Wang, "Twitter spammer detection using data stream clustering," *Information Sciences*, vol. 260, pp. 64–73, 2014.

[136] C. Yang, R. C. Harkreader, and G. Gu, "Die free or live hard? empirical evaluation and new design for fighting evolving twitter spammers," in *International Workshop on Recent Advances in Intrusion Detection*, pp. 318–337, Springer, 2011.

[137] C. Grier, K. Thomas, V. Paxson, and M. Zhang, "@ spam: the underground on 140 characters or less," in *Proceedings of the 17th ACM conference on Computer and communications security*, pp. 27–37, ACM, 2010.

[138] K. Thomas, C. Grier, D. Song, and V. Paxson, "Suspended accounts in retrospect: an analysis of twitter spam," in *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pp. 243–258, ACM, 2011.

[139] A. Caspi and P. Gorsky, "Online deception: Prevalence, motivation, and emotion," *CyberPsychology & Behavior*, vol. 9, no. 1, pp. 54–59, 2006.

[140] Z. Bu, Z. Xia, and J. Wang, "A sock puppet detection algorithm on virtual spaces," *Knowledge-Based Systems*, vol. 37, pp. 366–377, 2013.

[141] D. Liu, Q. Wu, W. Han, and B. Zhou, "Sockpuppet gang detection on social media sites," *Frontiers of Computer Science*, vol. 10, no. 1, pp. 124–135, 2016.

[142] S. Kumar, J. Cheng, J. Leskovec, and V. Subrahmanian, "An army of me: Sockpuppets in online discussion communities," in *Proceedings of the 26th International Conference on World Wide Web*, pp. 857–866, IW3C2, 2017.

[143] T. Solorio, R. Hasan, and M. Mizan, "A case study of sockpuppet detection in wikipedia," in *Proceedings of the Workshop on Language Analysis in Social Media*, pp. 59–68, 2013.

[144] M. Tsikerdekis and S. Zeadally, "Multiple account identity deception detection in social media using nonverbal behavior," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 8, pp. 1311–1321, 2014.

[145] B. Stone and M. Richtel, "The hand that controls the sock puppet could get slapped," *New York Times*, 2007.

[146] R. M. Milner, "Media lingua franca: Fixity, novelty, and vernacular creativity in internet memes," *AoIR Selected Papers of Internet Research*, vol. 3, 2013.

[147] L. Shifman, *Memes in digital culture*. MIT press, 2014.

[148] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 497–506, ACM, 2009.

[149] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer, "Truthy: mapping the spread of astroturf in microblog streams," in *The 20th International Conference Companion on World Wide Web (WWW'11)*, pp. 249–252, ACM, 2011.

[150] C. Bauckhage, "Insights into internet memes," in *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[151] C. W. Seah, H. L. Chieu, K. M. A. Chai, L.-N. Teow, and L. W. Yeong, "Troll detection by domain-adapting sentiment analysis," in *2015 18th International Conference on Information Fusion (Fusion)*, pp. 792–799, IEEE, 2015.

[152] P. Fornacciari, M. Mordonini, A. Poggi, L. Sani, and M. Tomaiuolo, "A holistic system for troll detection on twitter," *Computers in Human Behavior*, vol. 89, pp. 258–268, 2018.

[153] F. J. Ortega, J. A. Troyano, F. L. Cruz, C. G. Vallejo, and F. EnríQuez, "Propagation of trust and distrust for the detection of trolls in a social network," *Computer Networks*, vol. 56, no. 12, pp. 2884–2895, 2012.

[154] E. Cambria, P. Chandra, A. Sharma, and A. Hussain, "Do not feel the trolls," *ISWC, Shanghai*, 2010.

[155] P. Galán-García, J. G. d. l. Puerta, C. L. Gómez, I. Santos, and P. G. Bringas, "Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying," *Logic Journal of the IGPL*, vol. 24, no. 1, pp. 42–53, 2016.

[156] J. Im, E. Chandrasekharan, J. Sargent, P. Lighthammer, T. Denby, A. Bhargava, L. Hemphill, D. Jurgens, and E. Gilbert, "Still out there: Modeling and identifying russian troll accounts on twitter," *arXiv preprint arXiv:1901.11162*, 2019.

[157] D. Kim, T. Graham, Z. Wan, and M.-A. Rizoiu, "Analysing user identity via time-sensitive semantic edit distance (t-sed): a case study of russian trolls on twitter," *Journal of Computational Social Science*, vol. 2, no. 2, pp. 331–351, 2019.

[158] T. Mihaylov, G. Georgiev, and P. Nakov, "Finding opinion manipulation trolls in news community forums," in *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pp. 310–314, 2015.

[159] T. Mihaylov and P. Nakov, "Hunting for troll comments in news community forums," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 399–405, 2016.

[160] S. Cresci, M. Petrocchi, A. Spognardi, and S. Tognazzi, "From reaction to proaction: Unexplored ways to the detection of evolving spambots," in *Companion Proceedings of the The Web Conference 2018 (WWW'18)*, pp. 1469–1470, 2018.

[161] D. Boneh, A. J. Grotto, P. McDaniel, and N. Papernot, "How relevant is the turing test in the age of sophisbots?," *IEEE Security & Privacy*, vol. 17, no. 6, pp. 64–71, 2019.

[162] S. Raponi, I. Ali, and G. Oligeri, "Sound of guns: Digital forensics of gun audio samples meets artificial intelligence," *arXiv preprint arXiv:2004.07948*, 2020.

[163] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking," *arXiv preprint arXiv:1806.02877*, 2018.

[164] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics: A large-scale video dataset for forgery detection in human faces," *arXiv preprint arXiv:1803.09179*, 2018.