ISR Italian Sociological Review

University of Verona, Italy

ISSN 2239-8589 | Editor in Chief Paola Di Nicola
Managing Editor Debora Viviani

Tricked into Supporting: A Study on Computational Propaganda Persuasion Strategies

Valentina Nerino

How to cite
Nerino, V. (2021). Tricked into Supporting: A Study on Computational Propaganda Persuasion Strategies. [Italian Sociological Review, 10 (4S), 343-367]
Retrieved from [http://dx.doi.org/10.13136/isr.v11i4S.438]

[DOI: 10.13136/isr.v11i4S.438]

1. **Author information**
   Valentina Nerino
   Department of Sociology and Social Research, University of Trento, Trento, Italy

2. **Author e-mail address**
   Valentina Nerino
   E-mail: valentina.nerino@unitn.it

3. **Article accepted for publication**
   Date: January 2018

Tricked into Supporting: A Study on Computational Propaganda Persuasion Strategies

Valentina Nerino[*]

Corresponding author:
Valentina Nerino
E-mail: valentina.nerino@unitn.it

Abstract

The study reported in this paper aims to theoretically and empirically explore computational propaganda (CP) – a systematic process of political misinformation perpetrated on social networking platforms by automated agents with the aim of increasing support for specific political stances – focusing in particular on the factors determining its potential effectiveness. The claim maintained throughout this paper is that, among the possible factors determining this effectiveness, a pivotal one is represented by the design of CP messages themselves. Indeed, the hypothesis underlying this investigation is that the way CP content is created and presented is not casual, but deliberately designed to embed in it a set of persuasion strategies aimed at triggering a specific cognitive deliberation: considering misinformation as factual. Drawing from the Dual Process Theory of Cognition, the argument proposed is that info-cues contained in CP messages play a pivotal role in determining the likelihood of CP effectiveness. To test this hypothesis, a two-step analysis characterized by a mixed-method strategy has been implemented. To identify and collect CP messages, a machine learning algorithm able to perform bot-detection has been developed, while to analyze the content of those messages, a combination of qualitative and quantitative text analysis techniques has been employed. Lastly, preliminary results are presented and future work discussed.

Keywords: computational propaganda, bot-detection, heuristics.

[*] Department of Sociology and Social Research, University of Trento, Trento, Italy.

1. Introduction

The Internet is one of the most revolutionary inventions in the history of humankind. It has led to a paradigmatic shift in terms of data accessibility and communication power, transforming both information consumption and production. From this, a new category of online player has emerged: the prosumer – i.e., productive consumers, who are not only able to access an unprecedented amount of information, but can also now create their own content (Fuchs, 2018).

As a result, the Internet – and in particular Social Networking Platforms (SNPs) – held great promise for democracy. From their introduction, grass-roots movements and democracy advocates have been exploiting the accessibility and reach of Twitter, Facebook and other SNPs to voice their concerns and gain support in countries where freedom of speech was seriously limited (Bradshaw, Howard, 2018).

Following the example of these movements, other actors have started exploiting such platforms for their political goals. Yet, the messages they spread online are of a quite different nature: imbued with false or misleading information, they were specifically designed for *manipulating* social media users into supporting particular political claims, leaders or causes. Evidence of a systematic process of misinformation and manipulation occurring on SNPs has been provided by many scholars (for an overview, see Woolley, Howard, 2018), which demonstrates that those were not sporadic episodes limited to specific political events, but rather the manifestation of a pernicious communication strategy adopted by numerous political actors.

As pointed out by Woolley and Howard (2016), SNPs have been flooded with propaganda content, thus becoming a potential threat to the very democracy they were initially enhancing. By providing misleading information, deceptive descriptions of events, and deceitful characterization of opponents and outgroups, illegitimate[1] social media accounts depict a "reality" that is not factual[2], but that has the *potential* to generate real consequences (Bisbiglia, 2019; Rodny-Gumede, 2018; Siddiqui, Svrluga, 2016).

Motivated by this potential threat, in recent years numerous scholars have investigated this pernicious form of political indoctrination – referred to as *computational propaganda* (CP) (Woolley, Howard, 2016). From this massive research effort, a detailed account of CP manifestations has been compiled,

[1] Illegitimate accounts here are intended as social media profiles that profess an identity that is actually fictitious and specifically constructed to gain trust from other social media users.

[2] Where factual is intended as an event that has occurred in the real world (Bunge, 2003).

which provides an accurate depiction of tools, actors and events involved in the creation and circulation of CP. Without neglecting the importance of obtaining an accurate description of this phenomenon – which is a pivotal step toward the comprehension of its functioning – it could be argued that an equally important concern for researchers is to identify and assess CP effects, to understand the *actual impact* this phenomenon has on society.

As admitted by CP scholars themselves, the current literature lacks such contributions; consequently, very little is known about the actual impact these messages have on individuals and their actions (Woolley, Howard, 2018). However, to determine whether CP *does actually represent a threat to democracy*, researchers need to firstly conceptualize its effects, and then operationalize such concepts to make them empirically assessable: as trivial as it might seem, investigating extent and cause of an effect before having defined it is not a feasible task.

In an effort to provide some insights in this direction, this study develops a theoretical and empirical understanding of CP effects, with the aim of identifying and assessing potential factors responsible for the occurrence of these. The argument hereby presented is that CP can be considered *effective* when it manages to *persuade individuals* into considering misinformative content to be credible, trustworthy and, in the end, factual. As such, the research question underlying this investigation is: what increases the likelihood that CP content will be considered factual? Among the possible factors determining this likelihood, a pivotal one is the *design* of CP messages themselves. Indeed, the hypothesis underlying this investigation is that the way CP content is created and presented is not casual, but carefully crafted to deploy a set of *persuasion strategies* aimed at triggering a specific cognitive deliberation: considering misinformation as factual. Hence, the goal of this investigation: detecting and analyzing such strategies to better understand their underlying functioning.

Drawing from the Dual Process Theory of Cognition, the argument proposed is that *info-cues* contained in CP messages play a pivotal role in determining the likelihood of CP effectiveness. To test this hypothesis, a multidisciplinary approach has been adopted for formulating the theoretical framework informing the investigation, and designing and implementing the empirical analyses aimed at testing that framework. Accordingly, the paper is structured as follows. First, an overview of the literature on propaganda is presented and its limitations discussed. Second, the micro-approaches on cognition are introduced and a claim for their centrality in the study of CP effects is made. Drawing from these approaches, the theoretical framework underlying this study is presented, together with the mixed-method strategy developed for testing that framework. This strategy – encompassing computationally intensive and qualitative techniques – has been employed to:

1) identify CP accounts on Twitter; 2) analyze their messages to identify potential persuasion strategies embedded therein. Finally, results are presented and discussed, and further research development proposed.

Though preliminary, this paper presents a first attempt to investigate a complex phenomenon such as CP by means of a multidisciplinary approach, with the final aim of not only describing its functioning but also of identifying some of the factors potentially responsible for its manipulative power.

2. Literature review and theoretical framework

2.1 The case for a multidisciplinary approach

The phenomenon of propaganda and political manipulation has preoccupied scholars for centuries – perhaps millennia. In the realm of Social Science, the first attempt to conceptualize the phenomenon of propaganda dates back to the early twentieth century, with the seminal work of Lasswell (1927). In his essay "The Theory of Political Propaganda", he defines propaganda as "the management of collective attitudes by the manipulation of significant symbols" (Lasswell, 1927: 627). He maintains that individuals behave accordingly to attitudes, which are patterns of evaluation. The attitudes referring to an entire community – "collective attitudes" – are the targets of political propaganda, which Lasswell (1927: 628) describes as "the technique of using significant symbols" to alter collective attitudes. To achieve this alteration, propaganda uses a language of stimulus-response, which is doubly effective: it evokes the desired responses, while suppressing the undesired ones.

At the dawn of mass communication, Lasswell reflects on the role technology has in shaping political propaganda. Indeed, he maintains that the function of propaganda is in large measure "attributable to the social disorganization which has been precipitated by the rapid advent of technological changes" (Lasswell, 1927: 631). A combination of technological breakthroughs and increasing access to (basic) education caused a paradigmatic shift in the way people communicated. The advent of the popular press first and of the radio later resulted in a quicker and cheaper way of communicating and spreading information. This unprecedented abundance of information meant a greater accountability for rulers, since information about their decisions and actions could reach each and every ruled citizen in a very short amount of time. In Lasswell's terms, this accountability determined that "[m]ost of that which formerly could be done by violence and intimidation must now be done by argument and persuasion" (Lasswell, 1927: 631).

The paradigmatic shift represented by the surge of new technologies described by Lasswell evokes a contemporary scenario, where the Internet and

ICT developments have caused a drastic change in the way communication and information are conceived and delivered. In the light of the recent development of political communication strategies, Lasswell's argument seems frightfully accurate. Once heralded as the ultimate democratizing tool, the Internet and its web applications have been increasingly associated with "mass manipulation, vote suppression, and the propagation of false or misleading information" (Bradshaw, Howard, 2018: 16). By slowly substituting traditional information and communication sources, the Internet has increased its potential to be used as a tool for manipulating collective attitudes (Fuchs, 2018). Over the years, evidence of online manipulation and misinformation have been identified in numerous national contexts (e.g., Cresci et al., 2015; Cresci et al., 2017; Howard and Kollanyi, 2016; Jones, 2017; Kollanyi, Howard, Woolley, 2016; Pearce, 2013; Peel, 2013; Saka, 2014; Sanovich, Stukal, Tucker, 2018).

Experts have been particularly concerned with the role played by social networking platforms (SNPs) in this online process of collective misinformation, since they have found evidence that these platforms have been the vehicle through which CP – a new, technologically enhanced form of propaganda – has been circulating (Woolley, Howard, 2016).

Labelled in this way because of its automated nature, CP has been described as "the assemblage of social media platforms, autonomous agents, and big data tasked with the manipulation of public opinion" (Woolley, Howard, 2016: 4886). The term computational does not simply mean that the act of political indoctrination happens via computer, but rather that these political strategies are computationally enhanced, since they rely on ICT tools that are specifically designed and deployed as part of a political strategy (Woolley, Howard, 2018). Indeed, this new kind of propaganda relies on the creation of software programs that are capable of interacting with human users on SNPs to promote a particular political stance.

By combining the use of fake accounts[3] (characterized by various levels of automation[4]) and specifically designed content, political actors achieve two

[3] Social media accounts that do not refer to an existing and unique individual (but they want to appear as such), and are specifically created and managed for political indoctrination (Howard, Woolley, Calo 2018).

[4] These accounts have been classified on the basis of their level of automation – i.e., to what extent the actions they perform are determined by lines of code. The most automated tools are political bots, software programs created to perform simple, repetitive, typically text-based tasks that make them pass as human users. They are able to "rapidly deploy messages, interact with other users' content, and even affect or manipulate trending algorithms" (Woolley, Howard, 2018: 6). Cyborgs are accounts characterized by an intermediate level of automation. They use very sophisticated hybrid approaches that envisage an interchangeable combination of human and automated

major objectives: scalability and anonymity. Indeed, by automatically spreading partisan messages, political actors are able to dramatically increase their workload threshold – by making the dissemination process faster and less costly – while at the same time having the chance for their content to reach a potentially unlimited audience. Moreover, by using fake accounts, they remain unknown throughout this process of political indoctrination (Woolley, Howard, 2018).

Basing their argument on the Agenda Setting Theory (McCombs, Shaw, 1972), Woolley and Howard (2018) maintain that the manipulative power of CP messages derives from its ability (and success) to make certain political issues more prominent than others, thus shaping the political debate and influencing the electorate. They found evidence in support of these claims by analyzing the 2016 US presidential campaigns, where the systematic use of computational tools allowed some political actors to control the electoral agenda, thus directing the attention of both mass media and public towards specific issues.

Albeit acknowledging the importance of such pioneering studies, it is evident that their focus is on the SNP information environment rather than the effect of CP on individuals. Indeed, the authors maintain that "very little [is known] about the actual influence of highly automated accounts on individual political attitudes, aspirations, and behaviors … it is hard to demonstrate that any particular tweet, Facebook post, or other social media message has a direct effect on voters" (Woolley, Howard, 2018: 244).

If we limit our analysis to a macro-approach, Woolley and Howard are probably right, since using only macro theories and approaches to investigate effects of CP on individuals would be an extremely arduous task. Indeed, those (few) scholars investigating such effects do so by means of micro-perspectives, mostly drawing from cognitive sciences (e.g., Bail et al., 2018; Bail et al., 2020). The rationale behind the adoption of a micro-approach for studying manipulation has to be found in how manipulation itself is theoretically conceived. Though definitions may vary slightly, in this literature political manipulation is often referred to as an intended alteration of people's beliefs on a certain issue by means of specifically designed stimuli (Coxall, 2013). Although the effects of this alteration might be visible at the macro-level (e.g., changes in trending topics on SNPs, proliferation of certain junk news on various platforms, etc.), it seems reasonable to argue that the alteration of beliefs

activities (Grimme et al., 2017). This means that these accounts are not solely regulated by an algorithm, but they are characterized by human curation as well (Cresci et al., 2017). Trolls are the least automated accounts, since they are managed by actual human beings that hide their real identity by creating and impersonating fictitious personas over SNPs (Woolley, Howard, 2018).

originates from the micro-level, since whether a change in people's beliefs occurs is determined by how every *single individual* reacts to the stimuli she is exposed to. In other words, manipulation has to do with how individuals process information and – based on the result of that process – take decisions on the validity of that information. This is the reason why micro approaches and theories may provide meaningful insights for understanding the functioning of propaganda – including its effects.

Specifically, *Dual Process Theories of Cognition* seem particularly appropriate for the study of CP functioning, since they provide a compelling argument about how human cognition works and why in some cases it produces flawed deliberations. In their 1982 seminal work 'Judgement Under Uncertainty: Heuristics and Biases', Tversky et al. lay down the basis for a theory of human thinking that envisages the existence of two distinct – but interdependent – cognitive processes: System 1, which works in an associationist, "quick-and-dirty" fashion; System 2, which is characterized by a more deliberate, serial, rule-based way of processing information. While System 1 has the undoubted advantage of being faster and requiring less effort, it is also more likely to produce systematically biased deliberations, since it relies on heuristics, mental shortcuts specifically designed to reduce decision-making cognitive effort. Therefore, though efficient, this system has a drawback: judgements resulting from its adoption are sufficient to reach short-term goals, but they are subject to bias – since they rely on the uncritical application of available knowledge structures, rather than on an in-depth analysis of the information received (DiMaggio, 1997).

Such researches are particularly insightful for the study of CP, since heuristics may be the key to understand how it works and what its effects are. Indeed, as previously outlined, heuristics may lead to systematically biased decisions. The very fact that these "judgement errors" are systematic – and not random as the Rational Choice Theory maintains – means that they occur when certain mechanisms take place. Therefore, by learning how to trigger such mechanisms, it is possible to induce a biased cognitive reaction in a subject (Milosavljevic et al., 2012).

This insight is particularly compelling when applied to the investigation of CP, since it is able to explain how political actors deploying such communication tools may be successful in persuading their audience. Indeed, online political campaigners might exploit the cognitive flaw generated by heuristics to trigger systematically biased judgements on specific issues. To achieve this aim, political actors have to increase the chances of their audience (i.e., social media users) employing System 1 rather than System 2 when processing their messages, thereby prompting the resulting deliberation to favor the actors' message. The maintained hypothesis is that they manage to do so by designing political messages containing info-cues[5] able to trigger certain heuristics, thus increasing the likelihood of a biased line of reasoning (i.e., considering a CP message factual and legitimate). Therefore, the crucial point is to firstly *identify* such heuristics (and their related info-cues) and then be able to *test for their actual presence* in CP messages. Based on a detailed review of the existing literature, three pivotal heuristics have been identified.

2.2 Heuristics and info-cues

2.2.1 Availability

The availability heuristic represents the human tendency to rely on immediate examples that come to mind when evaluating specific topics, issues and concepts (Gilovich, Griffin, Kahneman, 2002). This mental shortcut operates on the notion that if something can be recalled, it must be important, or at least more important than alternative solutions which are not as readily recalled (Esgate, Groome, 2005). Therefore, events that are more familiar in memory are judged to be more frequent and, hence, more relevant, regardless of their actual likelihood (Ashcraft, 2006). As a result of such bias[6], individuals are likely to weigh their judgments according to more recent and prominent information, making new opinions biased toward that latest information they have been exposed to (Colman, 2015).

Because of its relation with news consumption, such a heuristic is particularly relevant for the phenomenon under investigation. Indeed, the prominence given by media to certain issues rather than others makes such issues easier to recall and, hence, more likely to be considered important by the audience (Steenbergen, Colombo, 2018). Therefore, if higher salience increases the likelihood for an individual to consider such issues as important, then it is legitimate to assume that the aim of political actors is to make their stances more prominent on all sort of media. While massively influencing traditional mass-media is an arduous task in democratic countries, manipulating the salience of an issue online is a much more achievable goal, especially if computational tools able to play algorithms and alter trend topics on SNPs are employed.

This is the reason why it can be expected that, among the info-cues deployed by computational propagandists, there will be some aimed at

[5] For an example of how info-cues trigger certain heuristics when individuals look for information online, please see the work by Sundar, Knobloch-Westerwick and Hastall (2007).

[6] Identified by Kahneman and Tversky in 1973, and lately denominated *familiarity bias* (Ashcraft, 2006).

triggering the availability heuristic. The hypothesis is that these unfold as terms that remind the SNP user of a specific topic/issue that in that moment is very prominent in main stream media because of a specific and emblematic event (e.g., addressing migration policies a few hours after an unprecedented number of boats have reached the European coasts).

2.2.2 Representativeness

The representativeness heuristic refers to the cognitive shortcut individuals employ to identify the correct categorization of an entity. By means of this heuristic, it is possible to assess similarity between objects and organizing them based on category prototypes. However, relying on representativeness when making judgements increases the chances of erroneous inferences, since the fact that something (or someone) is representative of a category does not directly translate to a greater likelihood of that something (or someone) behaving according to that categorization (Gilovich, Savitsky, 2002).

The representativeness heuristic has drawn the attention of numerous scholars because of its immediate link with stereotyping, a socio-psychological phenomenon intertwined with prejudice, discrimination and, eventually, social exclusion (Bodenhausen, 1990; Fiske, Neuberg, 1990; Gilovich, Griffin, Kahneman, 2002). Particularly effective and of immediate comprehension, stereotypes have been widely employed as a propaganda strategy through the centuries (Cole, 1998). For this reason, it is extremely likely that, also in the case of CP, stereotypes are used to induce and reinforce overgeneralized beliefs about specific social and ethnic groups. Thus, among the info cues employed by computational propagandists, it is reasonable to expect some triggering the representativeness heuristic, encouraging stereotypical beliefs. This means that terms referring to different topics/issues are used and combined within the same sentence, thus suggesting a link between their underlying concepts (e.g., the use of 'migrant' and 'security' in the same message).

2.2.3 Affect

This heuristic describes the importance that emotions have in guiding people's judgement and decisions. In particular, it refers to the fact that "representation of objects and events in people's minds are tagged to varying degrees of affect" (Slovic et al., 2007: 1335). When making judgements and taking decisions, individuals refer to "an 'affect pool' containing all the positive and negative tags consciously or unconsciously associated with the representations [of such objects and events]" (Slovic et al., 2007: 1335).

Therefore, when the affect heuristic is triggered, emotions generated by external stimuli will influence decisions regarding the action to undertake in regards to those stimuli. Triggering an emotional reaction in the audience is a widespread technique employed by communicators. The role of emotions in the art of persuasion has been extremely well known since antiquity. Most notably, the Greeks included pathos (i.e., emotional appeal) among the three modes of persuasion used in rhetorical speaking. Propaganda rhetoric is no different: emotions such as fear, indignation and anger are constantly triggered by particularly moving content, specifically designed to induce a desired reaction in the audience (Bussemer, 2005). For these reasons, it is sensible to expect that, also in the case of CP, content is designed to trigger powerful emotions among social media users, to trigger a gut-reaction in support of specific political stances. In operational terms, this means finding emotional language or attention-grabbing cues (e.g., use of specific punctuation, emoticons, capital letters) in CP messages.

Table 1. summarizes the argument advanced in this section, by providing an overview of the heuristics (and their respective info-cues) that are assumed to be a decisive factor in determining CP effectiveness.

TABLE 1. Heuristics and info-cues responsible for CP effectiveness.

Heuristics	**Info-cues**
Availability	Addressing a topic that is very prominent in mainstream media because of the occurrence of an emblematic event
Representativeness	Juxtaposing topics to imply a logical connection between their underlying concepts
Affect	Employing emotional language and attention-grabbing cues to trigger a gut-reaction

3. Methodology

To test the hypothesis of the existence of heuristic-based info-cues specifically designed to increase CP effectiveness, two subsequent analytical steps are needed: the first involves the *identification* and *collection* of CP messages, while the second concerns the *analysis* of content embedded in such messages.

Given the complex nature of the phenomenon under investigation, the identification of a suitable research design – encompassing both the micro and macro levels previously discussed – is of paramount importance. In fact, limiting the empirical investigation to a single methodological account could seriously compromise our understanding of the *overall* functioning of CP, including causes and extent of its effects. This is the reason why a mixed-method approach – not only encompassing both quantitative and qualitative methodologies but also adopting a multidisciplinary approach – is needed.

Accordingly, to implement the two analytical steps previously mentioned, a mixed-method strategy had been designed. This envisages: the development of a supervised machine learning (ML) algorithm able to detect social bots[7] for the identification and collection phase; the application of a combination of qualitative and quantitative text analysis techniques for the content-analysis phase.

3.1 Bot-detection: Identification of CP accounts and collection of their message

The aim of this analytical research stage is to develop a bot-detection strategy able to detect automated accounts on SNPs, to subsequently analyze the content they have been disseminating. The crucial issue in the identification of a successful bot-detection strategy relies on the detection features identified as the most predictive for such accounts. Based on the insights offered by the florid scholarship on bot-detection, features regarding account-level characteristics have been identified as the most predictive for automated social media profiles (Varol et.al., 2017; Woolley, Howard, 2018). Before discussing in detail the features used to detect social bots, some preliminary clarifications are needed to better contextualize the scope of this analysis.

3.1.1 Case Selection

Even though a plurality of SNPs is available online, only Twitter has the necessary features that make this empirical analysis feasible and meaningful. Indeed, Twitter is the social media where both official – i.e., from the communication staff of a party or candidate – and grass-rooted political communication occurs (Alonso-Muñoz, Marcos-García, Casero-Ripollés 2016). Moreover, it is one of the few SNPs providing such an "open" access to its data.

Given the interest in detecting bots that address political issues, it was necessary to identify a precise context in which CP could potentially occur. As the literature suggests, CP is deployed the most during major political events – e.g., elections, political crises and national-level demonstrations and protests (Woolley, Howard, 2018). For this reason, the *2019 European Elections* was a suitable event on which to perform this analysis. In terms of feasibility and meaningfulness, the scope of the investigation had to be limited to a single country, since recent research has shown how CP messages drastically differ in terms of national context (Woolley, Howard, 2018). Italy was considered to be a suitable choice because of the increasing use of CP by its political leaders (Nimmo, Pellegatta, 2018; Vogt, 2012) and because of the familiarity of the author with both language and political context.

[7] The automated accounts through which CP content is spread on SNPs.

3.1.2 Data Collection

Data collection was performed via the Twitter API, combining the use of both Real-Time and Search API. The retrieving criteria (i.e., the queries) employed to compile a dataset of suitable accounts on which to perform bot-detection were: geo-location – to capture tweets belonging (or pretending to belong) to the Italian Twitter-space; hashtags/keywords – to capture tweets related to the European Elections and its political campaigning. These criteria have been implemented into three distinct retrieving processes, outlined in Table 2. At the end of this data collection process, the dataset obtained contained a total of 3,254,257 tweets, posted by 237,572 unique accounts.

TABLE 2. Retrieving Processes Characteristics.

Criteria	***API***	***Time-frame***	***Queries***
Geo-location	Real-Time	April 1st – June 4th 2019	Tweets geotagged in Italy (i.e. tweet object coordinates = IT)
Keywords/Hashtags	Real-Time	April 1st – June 4th 2019	Tweets containing EU Elections institutional hashtags: #elezionieuropee2019 #elezioni2019 #elezioni #europa #ElezioniEuropee #EU2019 #Europee2019 #VoteEurope #stavoltavoto
Keywords/Hashtags	Search	7 days preceding the search	Tweets containing hashtags become popular during the electoral campaign: #il26VotoPD #pdnetwork #stavoltavotolega #domenicavotoLega #oggihovotatoLega #oggivotoLega #IoDomenicaNonVotoLega #iononvotolega #SeQuestoèunMinistro #SeNonCiFosseIlMoVimento #NoiconSalvini

3.1.3 Detection Methods

Based on the literature regarding bot-detection methods at the account-level, a supervised ML method has been developed for this analysis. The performance of this algorithm is highly dependent on the features the researcher identifies as those able to discern between bot and human accounts (Yang et al., 2019). Indeed, these features are not only those used to create the labelled dataset containing the Twitter accounts that will be used as a training set for the machine, but are also directly used in the ML training process itself – which underlines how crucial is their identification for the success of the whole detection process (Yang et al., 2019).

Drawing on the existing literature (e.g., Latah, 2020 Woolley, Howard, 2018; Varol et al., 2017), two sets of account-features have been identified as the most predictive of the automated nature of a Twitter account, namely *profile settings* and *account activity*. The former is directly related to the specific characteristics of a Twitter profile[8], while the latter regards the set of actions (e.g. tweeting, retweeting, liking, following, etc.) a social media user can perform over Twitter. Since social bots show peculiar behaviors in terms of frequency and magnitude of their actions, the distribution for each action performed by each individual account included in the dataset has been computed to detect *outlier accounts*. To provide an example, consider the "re-tweeting" action (i.e., re-posting of content). The information provided by the API includes the number of retweets performed by each account at the date of the data collection. Having a two-month timespan during which Twitter data were *continuously* collected, it is possible to obtain information about the number of re-tweets performed by each account at different points in time during those two months, and therefore to calculate the distribution[9] of retweets characterizing the activity of a given account in that timespan. Moreover, it is also possible to calculate the *proportion* of retweets in relation to other actions (e.g., tweeting an original content) – which is another indicator of automated action[10].

After having identified the key features to train the ML algorithm with, a labelled dataset of 1535 bot and human accounts – extracted from the original dataset – was created. This dataset contained the values of all the 74 features (Table 3.) extracted from each of the 1535 accounts, plus the label regarding their bot/human nature. The labelling process occurred through manual annotation, a common procedure for supervised ML techniques (Varol et al., 2017). To identify a suitable sample of political bots and human accounts to build the labelled dataset, a preliminary quantitative analysis of the identified

[8] E.g., length of and types of characters in screen-name and username, account age, settings regarding profile picture, background photo and theme, and bio-statement. For a complete overview, see Table 3.

[9] From each distribution, the following six statistics were computed and used as individual features: min, max, median, mean, standard deviation and skewness.

[10] The literature on computational propaganda provided evidence of the fact that: 1) political bots usually make a disproportionate use of retweets and 2) are more likely to follow a disproportionate number of accounts, if compared to the number of followers they manage to attract (Woolley, Howard, 2018).

features has been performed on the original dataset[11]. This means that each feature has been calculated at the account-level, for all accounts. Then these results have been aggregated, obtaining the distribution of each feature at the dataset-level. By analyzing the distribution of each feature, it was possible to select the accounts on which to perform the manual annotation – i.e., an inspection of each Twitter profile selected to decide on the "nature" of such account[12].

This selection was made on the basis of the position occupied by the accounts in each feature distribution since, according to the literature, accounts showing outlier behavior are to be considered the most suspicious ones (Varol et al., 2017). Therefore, for features with cardinal characteristics (i.e., for which a distribution could be computed), the twenty most "deviant" accounts – the "last" ten from each tail of the distribution – were selected for manual annotation. With the same logic, the "first" twenty accounts falling on the mean (or right above/below it) were selected, since there are very high chances that accounts behaving as the average are managed by humans (Varol et al., 2017). Conversely, for categorical features with dichotomic characteristics, twenty accounts from each category were randomly selected. This process has been done for each feature (74), with a total of 2342 accounts suitable for inspection, since several accounts repeatedly appeared among "outlier" and "median" groups. The manual annotation resulted in a labelled dataset containing 631 bot and 904 human accounts (807 accounts were excluded either because they were deactivated at the time of the inspection or because the inspection did not provide conclusive results). This dataset has been used to train the ML algorithm, which was based on a Random Forest classifier, considered to be one of the most robust in terms of overfitting problems (Sidana, 2017).

[11] To obtain meaningful results concerning twitter-activity features, the preliminary quantitative analysis was performed only on those accounts for which there were at least 10 tweets in the original dataset.

[12] The literature identifies some obvious "bot-flags", such as using a stock profile image or retweeting every message of another account within seconds (for a complete list see Varol et al., 2017). However, since a fixed set of rules that identifies bots with complete certainty does not exist, *Botometer* – a web-interface that can be used as a tool for bot-detection (https://botometer.iuni.iu.edu) – was consulted when manual inspection produced indecisive results. When even Botometer produced unclear results, the account was eliminated from further analysis.

TABLE 3. Bot-detection features.

Bot Detection Features	**Features**	**Description**
Profile settings	Screen-name length	Number of characters in the screen-name (if it has changed, the average screen-name length has been used)
	Number of digits in the screen-name	Number of digits in the screen-name (if it has changed, the average number of digits in all screen-names has been used)
	Number of changes in the screen-name	Number of times the screen-name has been changed
	Username length	Number of characters in the username (if it has changed, the average username length has been used)
	Number of changes in the username	Number of times the username has been changed
	Default profile (binary)	Whether the default profile settings were altered (0 = no alteration)
	Default profile picture (binary)	Whether the default profile picture was altered (0 = no alteration)
	Account age (days)	Account age (in days) at the time of the EU Election Day (May 26th 2019, 23:59:59)
	Profile description (AKA bio statement)	How many times the profile description was altered (0 = no alteration)
	Location (binary)	User-defined location for his/her account (0 = no location set)
Account activities	(*) Number of friends distribution	Number of accounts the account is following (i.e., followings or friends)
	(*) Number of followers distribution	Number of accounts the account is followed by (i.e., followers)
	(*) Number of statuses distribution (per day, per hour and total)	Number of statuses (i.e. tweets, retweets and quotes) issued by the user (per day, per hour and in total)
	(*) Number of favourites distribution	Number of statuses the user has liked (in the account's lifetime)
	(*) Number of retweets distribution (per day, per hour and total)	Number of retweets (including quotes) issued by the user (per day, per hour and in total)
	(*) Number of mentions distribution	Number of mentions contained in user's statuses
	Retweets/Statuses proportion (ratio – per day, per hour and total)	Total number of retweets (including quotes) divided by the total number of statuses (i.e. tweets, retweets and quotes). If ratio > 0.5, the user retweets more than he/she tweets
	Friends/overall relationships proportion (ratio)	Number of friends (mean) divided by number of relationships (i.e. number of friends (mean) + number of follower (mean). If ratio > 0.5 the user tends to have more friends than followers

*For each distribution, the following six statistics are computed and used as individual features: ***min, max, median, mean, std. deviation*** and ***skewness***.

3.1.4 Random Forest Classifier Results

Once the labelled dataset was finalized, it was used to train and test the ability of the Random Forest Classifier to discern between political bots and human accounts. Since the results of this training-test process can slightly vary

– given the fact that the data for each set is randomly selected – both training and test processes were repeated 10,000 times, and the results presented are the average of these 10,000 training-test processes.

The confusion matrix represented in Table 4. shows the performance of the trained classifier on the test-set data. In the upper-left and the bottom-right corners respectively, the True Negatives (TN) and the True Positives (TP) are shown. These values represent the accounts for which the predicted and the actual class (human or bot) matched. In other words, for how many cases the algorithm was able to provide a correct prediction on the nature of the account analyzed. The upper-right corner shows the number of False Positives (FP), namely those accounts that were predicted to be political bots by the algorithm but, in fact, were human-labelled accounts. Conversely, the bottom-left corner presents the False Negatives (FN), namely those accounts that the algorithm predicted to be humans, but were actually labelled as political-bots.

TABLE 4. Confusion Matrix (average of 10000 tests).

N=767	**Predicted human**	**Predicted bot**
Actual human	TN=442	FP= 10
Actual bot	FN= 12	TP=303

Table 5. shows different measurement of the Random Forest classifiers performance. Except for "accuracy" (which provides a score for the overall classification performance), each performance-measurement is calculated for both bot and human account classifications. *Recall* and *precision* scores are considerably high for both human and bot classification. This means that both the proportion of relevant instances that have been retrieved over the total amount of relevant instances (recall) and the proportion of relevant instances among the retrieved instances (precision) are extremely high (Powers, 2011). Likewise, the *F_1-score* – which is the harmonic mean of precision and recall – shows very positive results for both Human and Bot classifications. Lastly, the level of *accuracy* of the overall algorithm is also high (0.971), meaning that the model performs quite well at predicting whether a social media account is either a human user or a bot. In conclusion, the overall performance of the classification algorithm is considered satisfactory.

TABLE 5. Random Forest Classifier Performance (average of 10000 tests).

	Precision	**Recall**	***F_1-score***	***Accuracy***
Human	0.974	0.977	0.975	0.971
Bot	0.968	0.962	0.965	

3.2 Text Analysis

Once the robustness of the detection algorithm was tested, it was deployed on the original corpus of accounts (N = 237572) to obtain a dataset of alleged bots. Data-cleaning was then performed on the resultant dataset to exclude those bot-accounts that, though characterized by an automated nature, were not actually disseminating political content on Twitter – as in the case of social bots employed for commercial purposes. The result of this data-cleaning process was a dataset of alleged political bots, on which both qualitative and quantitative text analysis was performed to test the hypothesis of the existence of heuristic-based info-cues embedded in CP messages.

3.2.1 Qualitative text analysis

To identify the presence of info-cues embedded in CP messages disseminated by political bots, a qualitative text analysis was performed. For this analysis to be feasible, a selection of tweets from each account belonging to the political-bots dataset was performed. The tweets selected had to fulfil two requirements: containing more than five words, and containing at least one of the buzzwords related to the 2019 European Election Campaign, identified by the Twitter platform as *trending topics* of that electoral event.

Table 6. provides a systematized overview of those words on the basis of the themes they were addressing, namely *partisan characteristics* (i.e., specific names of parties/leaders or more general indication of ideology/position on the political spectrum), *policy topics* (i.e., migration, the EU, the economic crisis and religious diversity), and the *2019 EU electoral event* itself.

TABLE 6. Overview of the buzzwords used to select the accounts on which to perform the qualitative text analysis.

Theme	Sub-theme	Buzzwords
Partisan Characteristics	Parties names	Lega; Fratelli d'Italia; Forza Italia; M5S/Movimento Cinque Stelle; PD/Partito Democratico; LeU/Liberi e Uguali
	Political Leaders name	Salvini; Meloni; Berlusconi; Di Maio; Di Battista; Zingaretti; Renzi; Grasso; Boldrini
	Ideological characteristics	Destra, Sinistra; Sovranisti/Sovranismo; Populisti/Populismo; Estremisti/Estremismo; Fascisti; Comunisti; Europeisti; Euroscettici
Policy Topics	Migration	Immigrazione; Immigrati/Migranti; Clandestini; Esseri/Diritti Umani; Sicurezza; Invasione; Quote
	The EU	Parlamento; Consiglio; Commissione; Europa; Unione; Bruxelles; Strasburgo; Burocrati
	The Economic Crisis of the Eurozone	Crisi; Lavoro; Disoccupazione/Disoccupati; Povertà/Poveri; Euro
	Religious Diversity	Papa; Bergoglio; Francesco; Cristiani; Ebrei; Musulmani; Islam
2019 EU Electoral Event	N.A.	Elezioni; Europee; Voto; Votare; Urne; Elettori/Elettrici

On these bases, 1132 tweets were selected and manually analyzed. Each tweet was checked for the three different kinds of info-cues previously identified. To check for info-cues triggering the availability heuristic, the following procedure was adopted: firstly, the main topic addressed by the tweet was identified and coded; subsequently the date on which the tweet was posted was determined using the tweet time stamp; lastly, the socio-political issues addressed by the news occurring in that week were examined to check whether the issue addressed by the tweet matched with those most prominent on mainstream media. To determine the presence of info-cues triggering the representativeness heuristic, the text of each tweet was analyzed to assess whether it addressed more than one topic. If so, the text was searched for an underlying (causal) link between two (or more) issues. To check for info-cues triggering the affect heuristic, each tweet was searched for the presence of emotional language, emoticon, attention-grabbing punctuation and capital letters.

3.2.2 Qualitative text analysis results

The results of the qualitative text analysis seem to corroborate the hypothesis of the existence of a set of info-cues embedded in CP messages and specifically designed to trigger the three heuristics identified (Availability, Representativeness and Affect). Indeed, more than 68 percent of the examined tweets addressed a topic that was prominent in the mainstream media if not on the same date of the publication of the tweet, at least in the same week. Less than 50 percent of the tweets examined addressed more than one topic but, among those which did, a large majority (74 percent) tried to established a causal link between the topics. The most prominent link established was the one between migration and security, followed by that between Islam and violence. Finally, emotional language was found in more than 80 percent of the sample examined. The most frequently used words conveying an emotional attachment were of a negative nature, with "vergogna(ti)"[13] and "buffone/i"[14] among the most used. Capital letters and exaggerated punctuation (in particular exclamation marks) were used in half of the cases, while emoticons were only in 18 percent of them. Interestingly, a third of the tweets examined made use of irony to attract users' attention and advance their argument.

Therefore, the qualitative text analysis performed has substantiated the hypothesis of the existence of persuasion strategies embedded in CP messages, that employ info-cues able to trigger a set of heuristics with the aim of inducing a favorable (biased) cognitive reaction. Given qualitative text analysis limited

[13] "Shame on you" (*Italian*).

[14] "Buffoon(s)" (*Italian*)

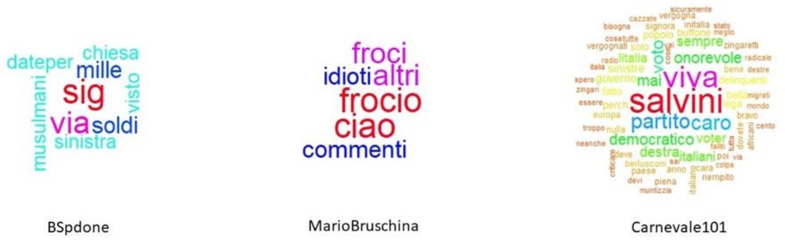
scope in terms of quantity of data analyzed, a quantitative text analysis using Latent Dirichlet Allocation (LDA) techniques has been implemented to assess the existence of any latent topic neglected by the qualitative analysis.

3.2.3 Quantitative text analysis using LDA

Quantitative text analysis has been employed as a tool to expand the analysis performed by means of qualitative methods. Given the interest in exploring possible overlooked topics, LDA has been identified as the most suitable technique for this aim, given its widespread use for discovering latent topics in textual documents (Blei, Ng, Jordan, 2003).

To perform this quantitative analysis, both 'TM' and 'LDA' R-packages have been employed. By means of the former, the text of the tweets has been extracted and all stop words and emoticons have been detected and excluded from the analysis. Subsequently, a word-frequency analysis has been performed to find the most frequent words in tweets' text and its results have been graphically plotted in a word-cloud (Figure1.).

FIGURE 1. Example of word-cloud related to BOT tweets.



By means of the latter package, a topic modelling analysis based on an LDA algorithm has been performed. This analysis was firstly run on the entire collection of tweets (using different values of the parameter *k*[15] based on the possible interpretations of the themes addressed by the most frequent words previously identified) and then performed on each sub-set of tweets generated by the same user.

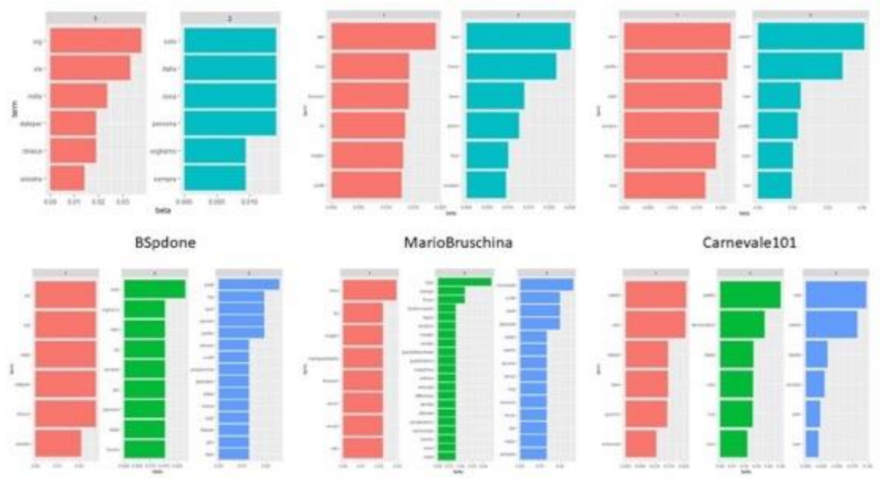
3.2.4 Quantitative text analysis results

The quantitative analysis based on an LDA algorithm provided only partial insights. Indeed, in the case of the analysis performed on the whole collection

[15] Value referring to the number of hidden topics to be discovered in the collection of documents under investigation.

of tweets, no compelling results have been revealed – i.e., no latent topic encompassing the whole corpora has been identified. Conversely, the analysis performed on the sub-set of tweets generated by the same account provided some interesting insights. Indeed, this analysis showed that every automated account had its own 'vocabulary', and that accounts differ from one another on topics addressed and levels of 'language sophistication' (i.e., variety of words characterizing their tweets). An illustrative example is represented by the three accounts identifiable as 'BSpdone', 'MarioBruschina' and 'Carnevale101'. As represented in the graphs below (Figure 2.), the former two were characterized by minimal vocabulary diversity and limited number of topics addressed, thus making their persuasion strategies potentially less effective and more detectable, while the latter showed a higher level of language sophistication and topic diversity, characteristics that increase the likelihood of passing as a credible human account.

FIGURE 2. Results of LDA analysis on BOT tweets (first row: k =2; second row: k=3).



4. Conclusions and Discussion

The aim of the study addressed by this paper was to provide a theoretical understanding and an operative definition of CP effects, and to subsequently identify potential factors responsible for their occurrence. The claims underlying the entire investigation were that 1) CP can be considered effective when it manages to persuade individuals to consider misinformative content as legitimate; 2) Among the possible factors determining this effectiveness, a pivotal one is represented by the design of CP messages themselves, that are created to include a set of info-cues aimed to trigger three different heuristics: Availability, Representativeness and Affect. The deployment of such heuristics when processing CP messages increases the likelihood for social media users to follow a biased line of reasoning that induces them to consider misinformation as factual.

To test this hypothesis, a two-step analysis characterized by a mixed-method strategy has been implemented. Firstly, to identify and collect CP messages, a bot-detection strategy based on a ML algorithm has been developed. Subsequently, to analyze the content disseminated by those accounts identified as alleged bots, a combination of qualitative and quantitative text analysis techniques has been employed. The results obtained – especially in the case of the qualitative analyses – support the hypothesis of the existence of a heuristic-based set of info-cues embedded in CP messages. However, further investigations are needed to: identify other (possibly non-textual) info-cues able to trigger the proposed set of heuristics or other relevant ones neglected in this study; identify other factors potentially increasing CP effectiveness (e.g., social media user characteristics); test for the actual impact these factors have on social media users' opinions regarding the truthfulness of a piece of information.

Though characterized by some limitations, these results offer interesting insights into CP effectiveness and they may represent a point of departure for further investigations in this direction. More generally, this paper highlights the potential benefits of employing a multidisciplinary approach when investigating a complex, multidimensional phenomenon such as CP. Indeed, addressing phenomena that encompass both macro and micro levels solely from a rigid, one-dimensional perspective might seriously limit our understanding of their *overall* functioning. Consequently, basing our empirical investigations on a single method of analysis might lead to "incomplete" results – in the sense that they address only a specific portion of the overall issue. In this sense, a mixed-method strategy provides researchers with a set of tools for exploring different aspects of the same phenomenon, thus allowing them to reach a more comprehensive understanding of its function.

The main issue characterizing multidisciplinary approaches employing mixed-method strategies is represented by the difficulty of developing diverse technical and analytical skills necessary to apply such methods. However, this "learning cost" is considered not only advisable, but possibly indispensable to investigate complex phenomena such as CP. This paper represents an example in this sense, since it investigates a complex phenomenon such as CP by means of a multidisciplinary approach, with the final aim of not only describing its function but also of identifying some of the factors (i.e., the design of the CP messages themselves) potentially responsible for its manipulative power.

References

Alonso-Muñoz, L., Marcos-García, S., Casero-Ripollés, A. (2016), Political leaders in (inter) action. Twitter as a strategic communication tool in electoral campaigns, *Trípodos*, (39), 71–90.

Ashcraft, M.H. (2006), *Cognition* (4th ed.), Upper Saddle River, Pearson Education Inc.

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. F., Lee, J., Mann, M., Merhout, F., Volfovsky, A. (2018), Exposure to opposing views on social media can increase political polarization, *Proceedings of the National Academy of Sciences*, 115(37), 9216–9221.

Bail, C. A., Guay, B., Maloney, E., Combs, A., Hillygus, D. S., Merhout, F., Freelon, D., Volfovsky, A. (2020), 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*, 117(1), 243–250.

Bisbiglia, V. (2019, May 9), *Casal Bruciato, Corcolle, Tor Sapienza: quando la rabbia dei quartieri contro gli stranieri si basa su notizie false o gonfiate*, Il Fatto Quotidiano, https://www.ilfattoquotidiano.it/2019/05/09/casal-bruciato-corcolle-tor-sapienzaquando-larabbia-dei-quartieri-contro-gli-stranieri-si-basa-su-notizie-false-o-gonfiate/5164228/

Blei, D.M., Ng, A.Y., Jordan, M.I. (2003), Latent Dirichlet Allocation, *Journal of machine Learning research*, 3, 993–1022.

Bodenhausen, G.V. (1990), Stereotypes as judgmental heuristics: Evidence of circadian variations in discrimination, *Psychological Science*, 1(5), 319–322.

Bradshaw, S., Howard, P.N. (2018, January 29), *Why does junk news spread so quickly across social media? Algorithms, advertising and exposure in public life*, Knight Foundation, https://kf-site-production.s3.amazonaws.com/media_elements/files/000/000/142/original/Topos_KF_White-Paper_Howard_V1_ado.pdf

Bunge, M. (2003), *Philosophical dictionary*, New York, Prometheus Books.

Bussemer, T. (2005), *Propaganda. Konzepte und Theorien,* Wiesbaden, VS Verlag für Sozialwissenschaften | Springer Fachmedien Wiesbaden GmbH

Cole, R. (ed.) (1998), *Encyclopaedia of Propaganda*, Abingdon-on-Thames, Routledge.

Colman, A.M. (2015), *A dictionary of Psychology* (4th ed.), Oxford, Oxford University Press.

Coxall, M. (2013), *Human Manipulation. A Handbook*, Granada, Cornelio Books.

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

Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., Tesconi, M. (2017), The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race, in Barrett, R. (ed.), *WWW'17 Companion: Proceedings of the 26th International Conference on World Wide Web Companion, April 3–7, 2017, Perth Convention and Exhibition Centre, Perth, Australia* (pp. 963-972), International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland.

DiMaggio, P. (1997), Culture and Cognition, *Annual Review of Sociology*, 23(1), 263–287.

Esgate, A., Groome, D. (2005), *An Introduction to Applied Cognitive Psychology*, New York, Psychology Press.

Fiske, S.T., Neuberg, S.L. (1990), A continuum of impression formation, from category-based to individuating processes: Influences of Information and Motivation on Attention and Interpretation, *Advances in experimental social psychology*, 23, 1–74.

Fuchs, C. (2018), Propaganda 2.0: Herman and Chomsky's Propaganda Model in the Age of the Internet, Big Data and Social Media, In J. Pedro-Carañana, D. Broudy, J. Klaehn (ed.), *The Propaganda Model Today: Filtering Perception and Awareness*, pp. 71–92, London, University of Westminster Press.

Gilovich, T., Griffin, D., Kahneman, D. (eds.) (2002), *Heuristics and biases: The psychology of intuitive judgment,* Cambridge, Cambridge University Press.

Gilovich, T., Savitsky, K. (2002), Like goes with like: The role of representativeness in erroneous and pseudo-scientific beliefs, In T. Gilovich, D. Griffin, D. Kahneman (ed.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 617–624), Cambridge, Cambridge University Press. https://doi.org/10.1017/CBO9780511808098.036

Grimme, C., Preuss, M., Adam, L., Trautmann, H. (2017), Social bots: Human-like by means of human control? *Big Data*, 5(4), 279–293. https://doi.org/10.1089/big.2017.0044

Howard, P.N., Kollanyi, B. (2016, June 20), *Bots, #Strongerin, and #Brexit: Computational Propaganda during the UK-EU Referendum*, SSRN, http://dx.doi.org/10.2139/ssrn.2798311

Howard, P.N., Woolley, S., Calo, R. (2018), Algorithms, bots, and political communication in the US 2016 election: The challenge of automated political communication for election law and administration, *Journal of Information Technology & Politics,* 15(2), 81–93. https://www.tandfonline.com/doi/full/10.1080/19331681.2018.1448735

Jones, M.O. (2017, June 7), *Hacking, bots and information wars in the Qatar spat*, The Washington Post, https://www.washingtonpost.com/news/monkey-cage/wp/2017/06/07/hacking-bots-and-information-wars-in-the-qatar-spat/

Kollanyi, B., Howard, P.N., Woolley S.C. (2016, November 17), *Bots and Automation over Twitter during the Third U.S. Presidential Debate*, Data Memo OII (Oxford Internet Institute), https://demtech.oii.ox.ac.uk/research/posts/bots-and-automation-over-twitter-during-the-third-u-s-presidential-debate/

Lasswell, H.D. (1927), The theory of political propaganda, *American Political Science Review*, 21(3), 627–631. https://doi.org/10.2307/1945515

Latah, M. (2020), Detection of malicious social bots: A survey and a refined taxonomy, *Expert Systems with Applications*, 151, 113383.

McCombs, M.E., Shaw, D.L. (1972), The agenda-setting function of mass media, *Public opinion quarterly*, 36(2), 176–187.

Milosavljevic, M., Navalpakkam, V., Koch, C., Rangel, A. (2012), Relative visual saliency differences induce sizable bias in consumer choice, *Journal of Consumer Psychology*, 22(1), 67-74.

Nimmo, B., Pellegatta, A. (2018, January 25) *ElectionWatch: Italy's Self-Made Bots. How the Lega's followers automate themselves*, Medium, https://medium.com/dfrlab/electionwatch-italys-self-made-bots-200e2e268d0e

Pearce, K. (2013, March 10), *Cyberfuckery in Azerbaijan*, Katy Pearce – Adventures in Research, http://www.katypearce.net/cyberfuckery-in-azerbaijan/

Peel, T. (2013, August 26), *The Coalition's Twitter fraud and deception*, Independent Australia, https://independentaustralia.net/politics/politics-display/the-coalitions-twitter-fraud-and-deception,5660

Powers, D.M.W. (2011), Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation, *Journal of Machine Learning Technologies*, 2 (1), 37–63.

Rodny-Gumede, Y. (2018), Fake It till You Make It: The Role, Impact and Consequences of Fake News, In B. Mutsvairo, B. Karam (ed.), *Perspectives on Political Communication in Africa,* pp. 203–219, London, Palgrave Macmillan.

Saka, E. (2014), AK Party's social media strategy: Controlling the uncontrollable, *Turkish Review*, 4(4), 418–423.

Sanovich, S., Stukal, D., Tucker, J.A. (2018), Turning the virtual tables: Government strategies for addressing online opposition with an application to Russia, *Comparative Politics*, 50(3), 435–454. https://doi.org/10.5129/001041518822704890

Sidana, M. (2017, February 28), *Intro to types of classification algorithms in Machine Learning*, Medium, https://medium.com/@Mandysidana/machine-learning-types-of-classification-9497bd4f2e14

Siddiqui, F., Svrluga, S. (2016, December 5), *NC man told police he went to DC pizzeria with gun to investigate conspiracy theory*, The Washington Post, https://www.washingtonpost.com/news/local/wp/2016/12/04/d-c-police-respond-to-report-of-a-man-with-a-gun-at-comet-ping-pong-restaurant/

Slovic, P., Finucane, M. L., Peters, E., MacGregor, D. G. (2007), The affect heuristic, *European journal of operational research*, 177(3), 1333–1352.

Steenbergen, M.R., Colombo, C. (2018), Heuristics in political behavior, In A. Mintz, L.G. Terris (ed.), *The Oxford Handbook of Behavioral Political Science*, Oxford Handbooks Online. https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780190634131.001.0001/oxfordhb-9780190634131

Sundar, S.S., Knobloch-Westerwick, S., Hastall, M.R. (2007), News cues: Information scent and cognitive heuristics, *Journal of the American society for information science and technology*, 58(3), 366–378.

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., Flammini, A. (2017), Online human-bot interactions: Detection, estimation, and characterization, in D. Ruths (ed.), *Proceedings of the of the 11th International AAAI Conference on Web and Social Media, 11 (1), May 15–18, 2017, Hyatt Regency Montreal, Montreal, Quebec, Canada* (pp.280–289), AAAI Press.

Vogt, A. (2012, July 22), *Hot or bot? Italian professor casts doubt on politician's Twitter popularity*, Guardian, http://www.theguardian.com/world/2012/jul/22/bot-italian-politician-twitter-grillo

Woolley, S.C., Howard, P.N. (2016), Automation, algorithms, and politics. Political communication, computational propaganda, and autonomous agents – Introduction, *International Journal of Communication*, 10 (9), 4882–4890.

Woolley, S.C., Howard, P.N. (2018), *Computational Propaganda: Political Parties, Politicians and Political Manipulation on Social Media*, Oxford: Oxford University Press.

Yang, K.C., Varol, O., Davis, C.A., Ferrara, E., Flammini, A., Menczer, F. (2019), Arming the public with artificial intelligence to counter social bots, *Human Behavior and Emerging Technologies*, 1(1), 48–61.