Table of contents

Introduction

In modern information society, digital media plays a key role in shaping public opinion. However, alongside positive aspects such as easier, cheaper, and faster access to information, serious issues also arise related to the manipulation of the information space and the creation of manufactured consensus (Woolley, 2023). This phenomenon creates an illusion of unanimity on a specific issue, achieved through targeted use of technologies and coordinated activity in digital media. Instead of organic development of public opinion through free exchange of information and discussion, a manufactured consensus is imposed from outside, leading to a distorted view of the real situation. Manufactured consensus forms through various mechanisms, such as using bots and fake accounts to mimic mass support, artificially boosting activity metrics (likes, comments, reposts/shares), spreading disinformation, and selectively silencing opposing voices. The mechanism through which public consensus (more specifically, its subjective perception) affects an individual's decision-making relies on reflexive control theory (Lefebvre, 2015).

Manipulations of public opinion on digital media have become a serious threat: worldwide, political organizations and governments use social platforms to spread disinformation, undermine trust in institutions, and engage in other forms of information influence (Bradshaw & Howard, 2018). García-Orosa (2021) identified four waves in the development of digital democracy, each accompanied by new tools for communicating with potential voters and influencing them. The latest, fourth wave, at the time of publication, according to García-Orosa, includes elements such as disinformation, social media, bots, and astroturfing.

The beginning of this wave dates back to 2016, when interference by the Russian "troll factory" IRA (Internet Research Agency) was uncovered in the U.S. presidential election and the Brexit referendum (García-Orosa, 2021; Schoch et al., 2022). Among recent global examples of using coordinated media activity to forge manufactured consensus, the

U.S. presidential election in 2016 and the Anti-Vaccine Movement during the COVID-19 pandemic stand out (Woolley, 2023). Similar mechanisms are used by the Russian Federation to justify its military actions (Vasara, 2020). The danger of these activities is that they distort the true picture of public opinion, erode trust in information in the digital space, and contribute to societal polarization (Lazer et al., 2018). Although there are current reasons to identify a subsequent, fifth wave of digital democracy connected to AI use (Vaiqoh & Astuti Nurhaeni, 2024), these influence tools remain relevant and effective.

The creation of coordinated, inauthentic media activity aimed at manufacturing consensus in the modern information landscape is a complex process involving various actors with different goals. Recognizing these actors, understanding their motives, and developing effective methods and tools to identify them are crucial for countering manipulation and disinformation.

1. Literature review

The field of manipulation detection in digital media is rapidly expanding in academic research. The number of publications dedicated to this phenomenon is so large that a structured review becomes necessary. First, we identify the main types of information manipulation recognized by researchers. Then, we examine the features of different media platforms and discuss the reasons for choosing them for manipulation studies. Next, we review methods for studying various types of information manipulations, trace the development of these methodologies, and consider the main approaches to data collection, dataset construction, and their applications. At the end of the review, we explore the motivations behind information manipulation and summarize the key findings of our analysis.

1.1. Types of information manipulations

Academic literature discusses many forms of manipulative activities and their goals. The scope of information manipulations ranges from individuals creating fake accounts for messaging on dating websites to political campaigns that sway attitudes across entire countries. Therefore, without claiming to be comprehensive, we focus on selected types of information manipulations relevant to this study.

Astroturfing

This term first appeared in 1985 in a speech by U.S. Senator Lloyd Bentsen, who was flooded with identical letters as part of a fabricated letter-writing campaign. Today, it is used to describe disguising artificial public support as a grassroots initiative. Astroturfing is a type of coordinated disinformation campaign that creates the illusion of widespread support for certain initiatives or "bottom-up" narratives, as if they come from regular people. This distinguishes it from classical forms of disinformation, where false beliefs are imposed "from above" (through advertising or political propaganda). Besides its use in business, where astroturfing is considered a form of low-budget or "guerrilla" marketing

(Dimitrieska, 2025), it is increasingly employed for political aims. Keller et al. (2020) even introduced the term "political astroturfing" for cases where a group of controlled accounts pretends to be independent citizens and extensively broadcasts prepared messages.

Such campaigns may involve both software bots and genuine users (trolls). Research indicates that astroturfing can be detected through digital traces like message uniformity, synchronized actions, and content–behavior linkage patterns. Alieva et al. (2022), examining a campaign to discredit opposition politician Navalny on Russian-language Twitter, discovered that hundreds of accounts simultaneously promoted hashtags against him, acting as a unified entity. Schoch et al. (2022) also analyzed a campaign by the Russian IRA on Twitter. Their tweets were deliberately crafted and influenced media agendas until they were uncovered. Therefore, many studies focus on coordinated information operations. Methods for analyzing them range from content analysis (identifying which "agenda" is promoted) to graph analysis (how accounts are connected internally and to external websites).

Bots

Initially, this term was used to describe a program created to perform specific tasks in the digital space without direct human involvement. Software bots can automate routine operations, gather information, and interact with users. However, they can also mimic the behavior of real users and be used to manipulate the information environment and public opinion. Today, the term "bot" encompasses many types of automated accounts. Akhtar et al. (2023) identify 10 different types of bots based on their activity goals (for example, Spam bots, Follower bot, Astroturf bot, Trigger bot, etc.). Ellaky et al. (2023) identify 11 bot types across three categories: Benign, Neutral, and Malicious.

In addition, bots are categorized by implementation technology. Researchers differentiate between automatic software scripts (the most primitive bot type today), partially automated accounts (cyborgs) where a person occasionally intervenes, and bots with AI features (Ellaky & Benabbou, 2024). Cyborgs are the hardest to detect because they behave almost like regular users but activate scripts for mass actions at critical moments. For example, a type of cyborg has been documented that suddenly conducts mass messaging and posts very frequently to support a candidate, while remaining inactive at other times (Ellaky & Benabbou, 2024). Traditional bots, which were just simple software scripts, are

now easier to detect, but the rise and development of generative AI have greatly complicated their behavior and, consequently, their detection.

Trolls and "troll factories"

The term "troll" is sometimes used to describe real users (or groups of users) who intentionally provoke conflicts, insults, or disinformation, often pretending to be fictional characters (Marcondes et al., 2024). In the context of astroturfing, however, even real users are more frequently labeled as "bots." Troll research aims to expose their identities and tactics. One of the first major cases was the revealed network of the IRA "troll factory" from Saint Petersburg. Linvill and Warren (2020) detailed how these operators created thousands of profile accounts and posed as ordinary Americans from different political backgrounds to "engage in dialogue from both sides" and radicalize it. Trolls often work with bots to engage in more subtle psychological manipulation. Therefore, troll detection methods are also similar: researchers look for unusual stylistic features, atypical activity hours, and, most importantly, signs of network coordination (Marcondes et al., 2024). Trolls' goals may not only be political but also commercial – for example, groups producing fake reviews of competitors' products.

Fake followers

This type of manipulation focuses on creating an illusion of popularity. Since the number of followers is a key metric of online popularity, the "fake followers" service is highly sought after by both commercial accounts and individual social media users. Akhtar et al. (2023) rank the follower bot second in popularity after spam bots. Cresci et al. (2015) demonstrated that entire bot farms sell follower-boosting services and proposed an effective (as of 2015) method called "Digital DNA" for their detection. Key signs of fake followers include mass account creation simultaneously, minimal activity besides following, identical avatars or bios, and numeric or randomly generated names. In studies of manipulation, detecting inflated follower counts is crucial because it enhances other influence tactics. For example, politicians or brands with millions of followers – whether real or fake – appear more popular and attract genuine followers. In turn, fake popularity can be used to validate false messages, such as in astroturfing campaigns.

Disinformation content (fake news, conspiracy theories)

Along with methods for detecting actors (bots, trolls), a research direction is also developing to recognize manipulative messages and

narratives. In some studies, this task is treated as an application of NLP methods (Althubiti et al., 2022; Mundra et al., 2023). Research that combines content analysis with network analysis is also of interest. For example, Guarino et al. (2020) studied networks of Twitter accounts that shared links to specific propaganda websites. They found that such accounts form clusters around particular domains (for instance, a group of accounts that almost exclusively share links from one pseudo-news site). This allowed them to speak about *propaganda distribution networks*. Cross-platform diffusion is also noted, when the same false information is broadcast simultaneously through several channels – Twitter, Facebook, YouTube, etc. (Pote, 2024). Wilson and Starbird (2020) described specific cases where links to the same disinformation resource were coordinatedly posted on Twitter and Facebook, reinforcing each other's effects.

Other tactics

This includes everything that did not fall into the categories above but has been studied by scholars: *fake trends* (Kausar et al., 2022); *deepfake media manipulations* (Tong et al., 2024; Mints & Sidelov, 2022). The deepfake problem became especially relevant in 2025 due to the development of generative AI models (NanoBanana, Sora, Seedream, etc.) that enable the creation of realistic videos featuring specified characters. The availability of these services and the quality of the generated images and videos already suggest a dramatic increase in the impact of deepfakes on news media (Lundberg & Mozelius, 2025). Finally, in recent studies, researchers have paid attention to virtual influencers – accounts of synthetic personalities that first gain popularity through content created with generative AI and are later used for native promotion of needed ideas (Lee et al., 2025). At present, the main area of application of virtual influencers is marketing, but the boundary between marketing and manipulation is quite thin. Methodologies for studying this phenomenon are still being formed.

Thus, today this topic is a broad interdisciplinary field that covers diverse types of information manipulation in digital media, and the emergence of new types and tools of manipulation will guarantee the intensive development of this field in the future as well. Importantly, countermeasures tested against one type of threat are often applicable to others. For example, clustering methods developed for bot detection also allow successful detection of troll communities, and network

analysis of accounts that spread propaganda helps expose hidden operations regardless of the platform. A challenge in using academic results is that the manipulation landscape constantly changes under the influence of internal and external factors, necessitating continuous updating of observations and knowledge in this area.

1.2. Media platforms

Twitter

The vast majority of studies focus on Twitter, largely because, for a long time, this platform provided open access to data through an API, which many other social networks do not offer (Pote, 2024). Another reason why Twitter became the main platform for studying bots and the mechanisms of disinformation spread is that Twitter is widely used by politicians and other influential figures. Large information campaigns have repeatedly been observed on the platform.

Twitter has been studied, for example, in analyses of the spread of crisis events such as the COVID-19 pandemic (Schoch et al., 2022), and as a tool for influencing voting outcomes, for instance the U.S. elections in 2016 (Kollanyi et al., 2016) and 2020 (Ferrara et al., 2020), the Brexit referendum in 2016 (Howard & Kollanyi, 2017), and elections in other countries as well (Mints, 2025-a).

Due to the openness of the platform's data (before the change of ownership), many methods for bot detection on Twitter have been developed, along with benchmarks for comparing them. For example, the large, open TwiBot-20 dataset contains data on about 230,000 accounts and 33 million tweets for bot recognition tasks (Ellaky & Benabbou, 2024).

At the same time, researchers emphasize that focusing only on "Twitter bots" leaves gaps. Many astroturfing campaigns also rely on real (non-automated) accounts that act synchronously. Therefore, recent Twitter studies shift attention from individual accounts to the detection of coordinated account networks. Schoch et al. (2022) analyzed all exposed Twitter astroturfing campaigns and found that, on average, 74% of accounts in each campaign participated in joint posting (simultaneous tweets/retweets of the same messages), a behavior almost never observed among ordinary users. This indicates that Twitter campaigns leave "network traces" that can be used to detect them.

Facebook and Instagram

Despite Facebook's huge audience, there are fewer academic works on it, mainly due to restricted data access. Nevertheless, there are studies that examine covert manipulations on these platforms as well. For example, Schler & Bonchek-Dokow (2024) studied astroturfing on Facebook during various election campaigns and collected several million comments from politicians' pages. The researchers identified anomalous profiles on these platforms and proposed a method for detecting them using a combination of manual labeling and machine learning.

Among the most typical features of astroturfing on Facebook are unnatural, synchronized activity and the use of repeated comment templates over time. Using these patterns enabled training a classifier with an accuracy of about 92% (Schler & Bonchek-Dokow, 2024). Research also addresses Instagram, where campaigns of state trolling on the platform have been studied. For example, pro-government attacks on the Iranian diaspora on Instagram were analyzed (Schoch et al., 2022).

As noted in the cited sources, analyzing manipulations on Facebook/Instagram, unlike Twitter, requires labor-intensive web scraping (Pote, 2024), which has constrained their study. At the same time, the importance of these platforms is very high, and coordinated operations have been identified there as well. For instance, according to the Oxford Internet Institute project, cases of manipulative influence via Instagram, WhatsApp, Snapchat, and others were recorded in 12 of 48 examined countries (Bradshaw & Howard, 2018).

Reddit and forums

Platforms like Reddit have been studied less often. Reddit is a large news forum with a voting mechanism that, in theory, allows the hidden promotion of desired narratives (for example, through mass registration and coordinated voting by groups of bots or trolls). However, the rating mechanism (karma) makes it difficult for a large group of bots whose opinions would carry weight to appear quickly. Therefore, there is almost no data on large-scale influence operations on Reddit, although individual cases are known in which propagandistic political messages were posted in U.S. subreddits from accounts of unknown origin.

One bold experiment was conducted by Swiss researchers: they secretly launched AI-based bots in the *ChangeMyView* subreddit to test whether machines can influence participants' opinions. The experiment

caused a scandal, but it showed that the audience does not always distinguish a human from a program.

In academic literature on Reddit, studies also appear on detecting disinformation in topic communities. For example, in medical subreddits, bots were used to automatically filter harmful health myths (Sager et al., 2021). In general, Reddit remains a less studied field compared to Twitter or Facebook. This is partly because systematically collecting Reddit data is more difficult (external archives like Pushshift are needed) and partly because there are no scandals as high-profile as elections and bots on Twitter. Nevertheless, a number of studies show coordinated activity on forums, for example, by tracking repeated messages and links posted under different accounts, which can indicate hidden coordination (an analogue of astroturfing).

YouTube

The YouTube platform, as the world's largest video hosting service, is relatively understudied in the context of information manipulation, underscoring the scientific novelty of studies that focus on it. The main difficulty is the closed nature of recommendation algorithms and viewership data. Nevertheless, some researchers call YouTube a "powerful channel" for spreading disinformation worldwide. According to Serrano (2024), in 2022, more than 80 fact-checking organizations publicly accused the platform of enabling false content, including COVID-19 conspiracy theories, false election narratives, and other misinformation.

Among academic studies of propaganda spread through video, one can note the article by Allgaier (2019), which analyzed the ecosystem of climate videos on YouTube and found that algorithms recommended many videos denying climate change, contributing to a distorted picture. In another study (Calvo et al., 2022), a network analysis of YouTube recommendations was conducted, and a viewing graph was built, showing how videos with anti-vaccine disinformation cluster and direct viewers to each other. Thus, YouTube can draw users into information bubbles, but it remains unknown whether this is intentional on the platform's part or the result of users exploiting weaknesses in the recommendation algorithm.

It is also important to note that some sentiment analysis studies focus on comments under YouTube videos (Pradhan, 2021; Cunha et al., 2019). Some researchers point out the presence of "toxic" comments and suggest methods for analyzing them (Obadimu et al., 2019; Döring, 2020).