Fueling Toxicity?

Studying Deceitful Opinion Leaders and Behavioral Changes of Their Followers during the Dutch '21 Elections

Puck Guldemond, Andreu Casas, and Mariken van der Velden

Accepted for Publication at *Politics & Governance*

Abstract

The spread of deceiving content on social media platforms is a growing concern amongst scholars, policymakers, and the public at large. We examine the extent to which influential users (i.e. *Deceitful Opinion Leaders*) on Twitter engage in the spread of different types of deceiving content. Thereby overcoming the compartmentalized state of the field. We introduce a theoretical concept and approach that puts these deceitful opinion leaders at the center, instead of the content they spread. Moreover, our study contributes to the understanding of the effects that these deceiving messages have on other Twitter users. For 5,574 users and 731,371 unique messages, we apply computational methods to study changes in messaging behavior after they started following a set of eight Dutch deceitful opinion leaders on Twitter during the Dutch 2021 election campaign. The results show that users clearly apply more uncivil language, become more affectively polarized and talk more about politics after following a deceitful opinion leader. Our results thereby underline that this small group of deceitful opinion leaders change the norms of conversation on these platforms. Hence, this accentuates the need for future research to study the literary concept of deceitful opinion leaders.

Keywords: Computational Communication Science; Disinformation; Opinion Leaders; Social Media

Introduction

Social media (SM) platforms play a key role in our daily lives. People increasingly use SM to interact with friends and family, voice their opinions, consume news, and engage in politics (Popan et al., 2019; Spohr, 2017; Weeks et al., 2017). However, some information on SM is misleading, i.e. untrue, partly true, and potentially purposefully deceitful. This has been studied by scholars using the concepts fake news (Egelhofer & Lecheler, 2019; Guess et al., 2019), disinformation (Bennett & Livingston, 2018; McKay & Tenove, 2020), rumors (DiFonzo & Bordia, 2007; Friggeri et al., 2014), or conspiracies (Douglas et al, 2019; Sunstein & Vermeule, 2009), amongst other concepts. Academics have raised concerns, stating that deceitful content endangers democracy and society at large (Groshek & Koc-Michalska, 2017; Lazer et al., 2018; McKay & Tenove, 2020). For example, deceitful content has led to online discussions between SM users in which uncivil language is common, especially when these discussions are about polarizing political topics (Weeks & Gil de Zúñiga, 2021). This tone of voice, in turn, fuels toxicity on SM platforms (Kim et al., 2021). That is, uncivil language spills over to other SM users, which affects their attitudes towards those who are addressed in these messages, potentially leading to polarization. Uncivil messages are those that are insulting, harassing, very dismissive towards others, racist or against a minority group, misogynist, or cursing words, enabling a toxic sphere (Davidson et al., 2017; Theocharis et al., 2016). To remedy toxicity on their platforms, keeping them a healthy place for public debate, SM companies often remove users who spread deceitful content. This removal has fueled a societal debate about whether or not these actions are justified, because such regulatory measures stand in contrast to the claims that SM would provide more equal opportunities for free expression of political views than traditional media (Balkin, 2017). Hence, SM companies as private actors are engaged in regulating the "practical conditions of speech" in the digital space (Balkin, 2017). Yet, *are those who spread deceitful content harmful to others?* Currently, the empirical evidence on if, and to what extent, deceitful content harms other platform users is scarce. Therefore, we need a systematic study on disseminating a variety of types of deceitful content (e.g. fake news, conspiracies, rumors and disinformation), and the effect thereof on other SM users.

We argue that the current state of the field aiming to understand the negative consequences of deceitful content is limited in three ways. First, previous research has been very compartmentalized. Scholars have studied different types of deceitful content in isolation (Weeks & Gil de Zúñiga, 2021). We, however, claim that when disseminating deceitful information, SM accounts spreading deceitful content often do not stick to just one type of deceitful information: they spread a variety of deceitful content throughout. Anecdotally, the now purged SM accounts of far-right radio show host Alex Jones shows that he engages in conspiracies, as well as rumors and misleading information (Berr, 2019; Coaston, 2018; EenVandaag, 2018; Haselton, 2019; Paul, 2019). In our empirical analysis, we assess the validity of our claim that these kinds of salient accounts engage in the spread of different types of deceiving information. Thereby, this study meets and expands the work of Weeks and Gil de Zúñiga (2021) who call for research that goes beyond the mere distinction between different types of deceitful information. Furthermore, we build upon the work of Chadwick and Stanyer (2022) who theoretically argue for the need to have an overarching framework bridging the myriad of studies addressing deceitful content. We theorize and empirically demonstrate how various types of deceitful content are addressed, allowing us to gauge the harm of this content for other users, and thereby for democracy and society at large. Second, existing research is focused on the type of content spread, rather than on the SM accounts disseminating this information. If we aim to better understand the effect these salient accounts have on other users, and also to some extent, whether the removal of accounts disseminating deceitful content is justified, we argue that not the content, but the SM accounts should be at the center of analysis. We are interested in the negative effects that all these types of deceitful content

have on SM users, not just a particular type of deceitful content. Building upon the two-step flow of communication theory (Katz & Lazarsfeld, 1955), we introduce a theoretical concept that examines these salient accounts fueling SM platforms with toxicity by posting deceitful content, terming them *Deceitful Opinion Leaders* (DOL). Third, while there is plenty of existing knowledge about the overall prevalence and dissemination of deceitful content, we know little about the effects that DOLs have on their followers and other users on SM platforms. Scholars suggest that exposure to deceitful messages can have harmful consequences such as adopting more uncivil behavior and lead to increasing levels of affective polarization (Iyengar et al., 2012, 2019; Popan et al., 2019; Theocharis et al., 2016; Yarchi et al., 2020). We test whether exposure to deceitful information actually has these malicious effects.

To empirically assess the type of deceitful content DOLs spread as well as its effect on other SM users, we use an innovative research design that allows us to study the type of content DOLs disseminate and if this fuels toxicity on Twitter. Twitter is known as a key mainstream platform that allows us to collect the data needed to test our hypotheses. For a two month period (March 2, 2021 till May 4, 2021), we tracked eight Dutch DOLs (Maurice de Hond, Lange Frans, Sietske Bergsma, Robert Jensen, Blackbox News, Wierd Duk, Cafe Weltschmerz and Isa Kriens) and their followers. These DOLs are not an exhaustive, nor a representative list of DOLs in the Dutch Twitter sphere. Yet, they are well known for engaging in the dissemination of deceitful information (e.g. see Showredactie, 2020), and thereby a most likely case to test our approach and theoretical concept. All DOLs have accounts with a high number of followers (i.e. more than 11.000), showing that these DOLs voice opinions that are valued and accepted by others. Moreover, DOLs often spread deceitful content about highly polarized and political issues. This results in an (online) public space fueled with toxicity and deceitful content (Bergmann, 2020). The collected messages of these DOLs allow us to assess the validity of our claim that DOLs engage in the spread of different types of deceiving content. For each day in the period under investigation, we monitored for each DOL if they had new followers (N = 32.245). Subsequently, for each of these new followers, we collect the tweets they posted before and after they started following a DOL. Our analysis is two-fold. First, we look at the tweets posted by the DOLs and use content analysis to corroborate that they indeed engage in a wide variety of deceitful content, such as rumors and disinformation. Then, we look at the tweets sent by the new followers before and after, and use computational methods to test the extent to which they become more politically engaged, and post more uncivil and affectively polarized messages after following a DOL.

We show that, after starting to follow one of the eight DOLs in our sample, these users did increase their number of political, uncivil and affectively polarized tweets. The effects are statistically significant and of substantial magnitude. We observe stronger longer-term (30-days) than shorter-term (15-days) effects, although after two weeks their behavior starts reverting back to levels similar to those prior to following the DOL. We also observe stronger effects for those who started following more than one DOL. Our results thereby underline that while there is a small group of DOLs, they do have a substantial effect on how other SM users behave on these SM platforms. To keep SM platforms a healthy forum for public debate, SM companies regulate what can be posted. Fueled by fear that the dissemination of deceitful information distorts a healthy public debate, and thereby is detrimental to society, SM accounts engaging in this behavior are purged. Our results, however, demonstrate that following a DOL has a gate-way effect: not only are SM users adopting their norms of conversation (i.e. using more uncivil language) but they also introduce their SM followers to a view of politics that these followers feel more comfortable to engage in. This sheds an important light on the question of how to regulate SM platforms so that they can maintain fostering public debate without endangering the democratic process of deliberation.

Deceitful Opinion Leaders on Social Media and their Effects

Over the last decades, the media environment has changed drastically, into a high-choice media environment (van Aelst et al., 2017). This has affected the communication flow from the media to the masses. Many people receive news via SM through one of their online connections (Weeks et al., 2017). Hence, these connections function as a mediator between media to the mass public. This process was first explained by Katz and Lazarsfeld (1955) as the two-step flow of communication theory, which acknowledges this process of person-to-person influence and calls these mediators opinion leaders. Those are people that are held in high esteem and whose opinions are valued and accepted by others (Bergström & Belfrage, 2018; Choi, 2015; Katz & Lazarsfeld, 1955). In the early days of mass media, opinion leaders received information from the media and shared that information with their network via (offline) personal interactions. In the digital age, this process is similar, but takes place in an online environment: SM users seek out certain individual SM accounts for guidance and information (Choi, 2015). The information that SM users are exposed to depends on the opinions, interest and behavior of their online connections (Bergström & Belfrage, 2018). These opinion leaders inform and thereby potentially shape attitudes of less active recipients (Bergström & Belfrage, 2018; Carlson, 2019). Yet, they do not necessarily need to be message carriers for the greater good. In recent years, we have witnessed opinion leaders that deliberately spread information that is untrue or deceiving, such as Alex Jones or Lange Frans in the Dutch context. Influential accounts that engage in this behavior, we coin as *Deceitful Opinion Leaders (DOLs)*. DOLs are defined as SM users (a) with a large number of followers, (b) who engage in the production and dissemination of at least one type of deceitful content to their audiences.

Why do people follow DOLs and what is the effect thereof? Previous research demonstrates that most people are not engaged with politics, but do enjoy following entertaining content. As a by-product of seeking entertainment, politically inattentive individuals are exposed to information about political and societal issues (Baum, 2002). Social networks like Twitter provide increasing opportunities for people to be exposed to political content, even when using Twitter for different purposes, such as entertainment (Kim et al., 2013). DOLs typically post highly entertaining and engaging content, such as sarcastic or cynical comments. Hence, people are in part likely to follow them for entertainment value. A side-effect of following DOLs is that their followers are *incidentally exposed* (Bergström & Belfrage, 2018; Kim et al., 2013; Weeks et al., 2017) to political content – i.e. when DOLs tweet about societal, controversial and political issues, their followers (and the followers of their followers via sharing patterns) see this content. The same dynamic holds for exposure to misleading information (Lazer et al., 2018; Stroud, 2008). We argue that DOLs have a key role in the information others receive, resulting in a high influence on what DOL followers talk about (Zaller, 1992). That is, the topics of conversation – i.e. the deceitful information about societal and political topics – likely spillover to the DOL followers, leading to the following hypothesis:

H1: After following deceitful opinion leaders, users will tweet more about politics than they did before following them.

Next to *what* DOLs talk about, *how* they speak about political topics is also likely to be carried over to their followers. According to Weeks and Gil de Zúñiga (2021), online political interactions are often uncivil. The highly emotional nature of SM platforms provide a 'perfect storm' for the spread of deceiving and misleading content (Weeks & Gil de Zúñiga, 2021). DOLs often use inflammatory and uncivil rhetoric when discussing political topics or when referring to politicians (see Table 3 for an example). Due to anonymity, the threshold for uncivil behavior is lowered on SM platforms (Groshek & Koc-Michalska, 2017; Theocharis et al., 2016). In a SM environment, people

tend to say and do things that they would not necessarily do when being in the offline world (Suler, 2006). Therefore, SM platforms facilitate this uncivil behavior online (Groshek & Koc-Michalska, 2017). This, in turn, results in the usage of more uncivil language, posing threats, hard criticism and showing anger and hatred on SM platforms online, creating an online sphere rife with uncivil behavior (Suler, 2006; Theocharis et al., 2016). Impolite and uncivil discourse on SM platforms has a poisonous and polarizing effect. When people are exposed to incivility, they are more likely to use incivility in their own comments and messages (Gervais, 2015; Theocharis et al., 2016). This implies that those following DOLs, who are expected to use uncivil and inflammatory language, are more likely to mimic their rhetorical style; leading to the following hypothesis:

H2: After following deceitful opinion leaders, users will utilize more uncivil language.

Uncivil behavior on SM platforms reduces openness towards outgroups as uncivil discourse has poisonous and polarizing effects (Groshek & Koc-Michalska, 2017; Theocharis et al., 2016). As mentioned above, DOLs often talk about political topics or politicians in an uncivil manner. They use an 'us versus them' rhetoric when referring to the political elite. By doing so, they create an in-group (DOLs and their followers) and an out-group (the political elite and their followers). Based on the social identity theory (Tajfel et al., 1979), scholars have theorized and demonstrated that belonging to an in-group with a strong social identity, leads to the disliking and disfavoring of out-groups (Iyengar et al., 2012, 2019; Harteveld, 2021). Online, this results in SM users following more like-minded accounts that fit within their in-group. This implies that once SM users follow a DOL, they are likely to be immersed into an online community of like-minded people, forming online homogenous networks (Shu et al., 2017; Barbéra 2015; Barbéra et al. 2015). These homogeneous social networks reduce tolerance for alternative worldviews and amplify affective polarization, resulting in division and animosity between different parties, individuals or groups that hold opposite views on (political) topics (Iyengar et al., 2012, 2019; Lazer et al., 2018; Yarchi et al., 2020). Assuming that SM users following and mimicking a DOL can be seen as a united front (i.e. in-group), they are likely to view others as an out-group who they oppose. Thereby, they likely contribute to rising hostility towards other societal groups. By following DOLs, we expect users to become less tolerant, hence more polarized, towards outgroups with different opinions and ideas. Therefore, we expect the following:

H3: After following deceitful opinion leaders, users will become more affectively polarized.

Data & Methods

We collected the following data to assess the extent to which DOLs engage in the dissemination of different kinds of deceitful content, as well as to test our three hypotheses about the effects that they have on the behavior of their followers. First, we selected a set of DOLs to study. Then, we tracked their SM behavior to explore the types of deceitful content they posted. In addition, we needed to track the SM behavior of their followers. Ideally, for a clear identification, we wanted to track and study their behavior before versus after they started following a given DOL.

For a two month period during the 2021 Dutch elections (March 2 through May 4, 2021), we studied the Twitter behavior of a convenience sample of eight well-known Dutch DOLs (see Table 1 for a detailed list), and those ordinary users who started following them during the period of analysis. Although these DOLs are not a representative nor comprehensive sample of all DOLs, they are among the most visible ones in the Dutch twittersphere, and they are very suitable to conduct a proof of concept analysis to validate the theoretical concept and expectations put forward in this paper. Future research should address the conditions under which the findings presented here extend to a larger and

more comprehensive sample of DOLs. Despite this limitation, we believe the approach and analysis presented here contribute to build a better understanding of the actions of these types of opinion leaders, and how they shape conversations on SM platforms.

Table 1: List of the eight Deceitful Opinion Leaders we study.

Name	**Twitter handle**	**Num. Followers March 2, 2021**	**Num. of Followers May 5, 2021**	**Num. of New Followers Analyzed (H1-H3)**
Maurice de Hond	@mauricedehond	118,237	127,404 (+7.7%)	2,558
Wierd Duk	@wierdduk	84,617	90,403 (+6.8%)	1,101
Lange Frans	@langefrans	70,744	72,021 (+1.8%)	235
Robert Jensen	@robertjensen	52,686	56,462 (+7.1%)	588
Sietske Bergsma	@sbergsma	31,224	35,165 (+12.6%)	321
Café Weltschmerz	@cafeweltschmerz	17,062	17,754 (+4%)	32
Blck Bx	@blckbxnews	16,141	22,474 (+39.2%)	382
Isa Kriens	@isakriens	11,658	12,931 (+10.9%)	632
Total	-	-	**32,245** (**13,337** unique)	**5,574**

The first day (March 2, 2021), we pulled the list of followers for each of these DOLs. We only include followers that have sent at least one tweet before, to enable a comparison before and after these Twitter users started to follow a DOL. Then, everyday (until May 4, 2021) we pulled the following additional information: the messages sent by the DOLs that day, the list of users who started following a given DOL that day, (up to) the last 3,200 messages sent by these new followers (to gather information about their posting behavior prior to following the particular DOL), and the messages posted that day by the new followers detected in previous days (to gather information about their posting behavior after they started following a particular DOL).

We use the collected data for two main purposes. First, we manually code the messages posted by the DOLs themselves for whether they contain fake news, disinformation, conspiracy and/or rumors (non-mutually exclusive categories). The goal is to assess our claim that these DOLs engage in the dissemination of different types of deceitful content. As shown in Table 2, we rely on existing and validated definitions when coding for these four types of deceitful messages (see Appendix B for the codebook). For each DOL, 10 tweets were coded by two authors, resulting in 80 annotated tweets in total, leading to intercoder reliability values using Krippendorf's alpha of 0.99 for fake news, 0.98 for disinformation, 0.97 for conspiracy, and 0.93 for rumors.

Table 2: Definitions of deceitful content used for coding.

Fake news	Has a journalistic format, but is low in facticity (Egelhofer & Lecheler, 2019)

Disinformation	False information that is purposely spread to deceive people, seeking to amplify social divisions and distrust (Bennett & Livingston, 2018; McKay & Tenove 2020)
Conspiracy	Efforts to explain events, practices or secret plots that consist of two or more powerful actors, acting in secret for their own benefit and working towards a malevolent or unlawful goal against the common good (Douglas et al., 2019; Sunstein & Vermeule, 2009).
Rumor	Circulating information whose veracity status is yet to be verified at the time of spreading (DiFonzo & Bordia, 2007; Friggeri et al., 2014)

Then, in order to test potential behavioral changes, we count the number of political (H1), uncivil (H2), and affectively polarized (H3) tweets that new followers posted during the days before, versus the days after they started following the first DOL in our sample. We use two time windows for this before/after analysis, 15 and 30 days, to assess the robustness of the findings to this subjective cut-off. We collected data from 13,377 unique new followers for the DOLs in our sample. For a clear identification, when testing our hypotheses we will restrict our sample to: (a) users who started following one of the DOLs after March 2, 2021 (for the previous followers we do not know exactly the date they started following the DOL), (b) users for which we have collected their messages for the entire before and after time windows, (c) users who did not stop following the followed DOL during data collection (a total of 3,451 users started following one of the eight DOLs under analysis, but stopped following them before the end of data collection). Our final analytical sample includes a total of 5,574 followers (see Table 1) who sent a total of 731,371 tweets during the 30 days prior/after combined.

To count the number of political, uncivil, and affectively polarized tweets, we trained three machine-learning classifiers. First, we annotated 2.896, 5.242 and 855 for whether they were uncivil, political and affectively polarized, respectively (binary categories). Table 4 provides an overview of the annotated messages per classifier. Messages were coded as uncivil if they were insulting, harassing, very dismissive towards others, racist or against a minority group, misogynist, or when they consisted of cursing words (Davidson et al., 2017; Theocharis et al., 2016). Messages were coded as political if (a) a political party or organization was mentioned, and/or (b) if messages touched on relevant policy issues. Finally, messages were coded as being affectively polarized if users showed dislike towards an opposing group (by naming them, tagging them or mentioning them), such as a politician, political party, or societal group (e.g. conservatives/liberals, immigrants) (see Appendix C for the codebook). 100 Tweets were coded by two authors, leading to intercoder reliability values using Krippendorf's alpha of 0.86 for political tweets, 0.85 for uncivil language, and 0.87 for affectively polarizing language.

Since uncivil, political, and affectively polarized tweets are rare, in order to have as many true positives in our annotated set as possible, we used random sampling as well as active learning when selecting the cases to be annotated (Miller et al., 2020). Hence, the number of true positives in our annotated dataset are not really a reflection of the prevalence of these quantities in the overall dataset. Table 3 shows examples of the types of messages coded as political, uncivil and affectively polarized.

Table 3: Examples of political, uncivil and affectively polarized tweets (translated from Dutch).

Political tweet	According to the left-wing opposition parties, the deal does not go far enough, while the PVV believes that the cabinet has caved in.
Uncivil tweet	@DDStandard she's ugly. she's stupid .. she's not adding anything. just a hopeless nigger who also tries to shout something .. and nobody listens. she will never become someone like Pim Fortuyn..Sylvana cannot even stand in his shadow
Affective polarized tweet	RT I didn't think much of the left wing voters, but voting for fucking Sigrid Al Qaq-Kaag is like selling your soul to Europe...

Then we used the full corpus of annotated data to fine-tune three times the same transformer model (the Dutch version of BERT (de Vries et al., 2019) – *bert-base-dutch-cased*), one for each of the three (political, uncivil, affectively polarized tweets) classifiers. Deep transformer models such as BERT have been shown to improve machine text classification in many domains, including political and communication science (Terechschenko et al., 2020). In each case, we used 20% of the annotated data to create a completely untouched validation set. Then, we split 70/30 the remaining data into a train and test set. We used the train set to estimate model fit and update the model weights at each training iteration, and the test split the assess out-of-sample performance and to decide when to stop training the model further. We stopped the training when the test loss did not improve for 3 complete iterations. We trained each model three times, using a different train/test split each time (three-fold cross-validation). Finally, we assessed out-of-sample accuracy on the untouched validation set (which remained constant across the three folds).

In Table 4 we report the performance of each model based on this three-fold cross-validation conducted on the validation set. The uncivil and political classifiers perform very well: overall accuracy, as well as precision and recall are very high; and precision and recall are very similar, indicating that in the rare cases in which a classifier makes the wrong prediction, it is equally likely to miss-classify messages that are (v. aren't) uncivil/political. The performance of the affective polarization classifier is slightly lower – high accuracy (83%) but slightly lower levels of precision (65%) and recall (71%) – but the classifier is highly balanced (similar levels of precision and recall). We have no reason to believe that there is any systematic error for any of the classifiers. So any remaining noise would mean that we are conducting conservative tests of our hypotheses.

Finally, we use these classifiers to predict whether the rest of the unlabeled messages posted by the new 5,574 DOL followers are political, uncivil and affectively polarized; and to count the number of political/uncivil/polarizing tweets sent the 30 days before and the 30 days after starting to follow the first DOL in our sample.

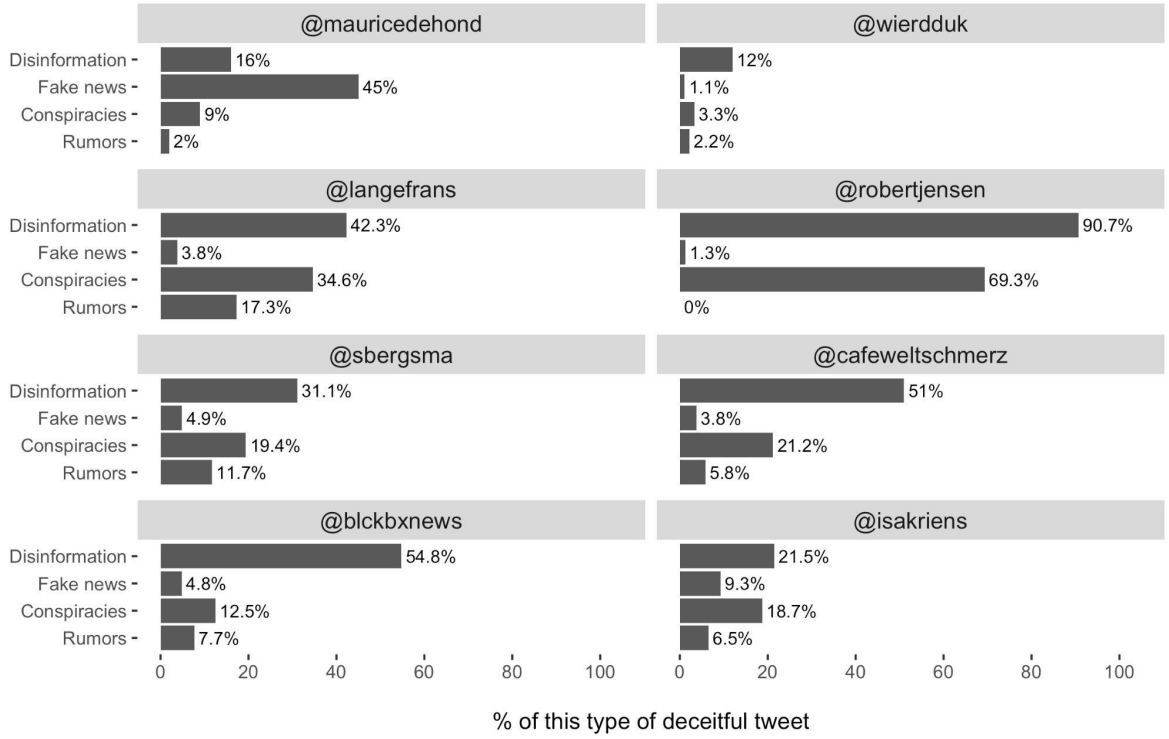
Table 4: 3-fold cross-validated performance of three BERT classifiers predicting binary outcomes: political, uncivil and affectively polarized tweets.

Classifier	N annotated	True Positives	Accuracy	Precision	Recall
Political	5.242	59%	86%	94%	86%
Uncivil	2.896	39%	86%	83%	80%
Aff. Pol.	855	35%	83%	65%	71%

Results

We begin by assessing whether DOLs indeed engaged in the dissemination of many types of deceitful content (e.g. fake news, disinformation, conspiracies, and rumors). We then move to testing our hypotheses regarding the behavior of new followers.

Figure 1: The percentage of tweets sent by the Deceitful Opinion Leaders under analysis that contain different types of deceitful content.

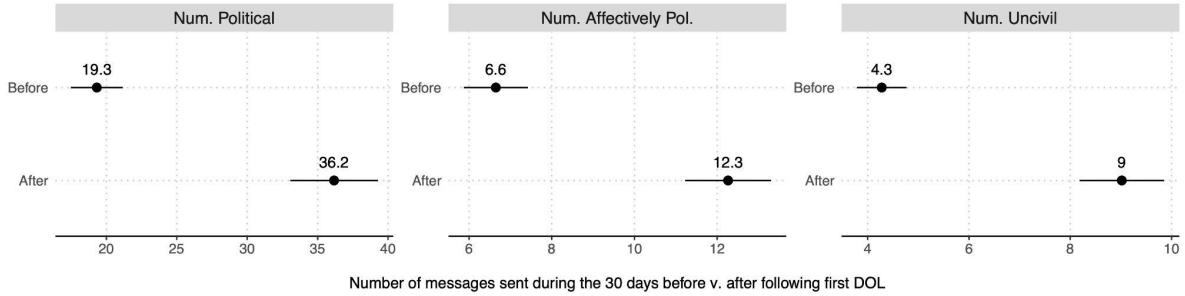


In Figure 1 we study the distribution of the deceitful content that was spread by each DOL during the period of analysis. In line with our theoretical framework, the figure illustrates that all DOLs engage (to some extent) in the dissemination of all types of deceitful content under scrutiny, from fake news to conspiracies and rumors. For example, with the exception of @robertjensen, the remaining DOLs posted at least one message containing each of the deceitful typologies under study. Although sometimes they have a clearly "preferred" deceitful category (e.g. 45% of @mauricedehond's tweets spread fake news, and 54.8% of @blckbxnews' messages promoted disinformation), they also engage in the spread of other kinds of deceitful content quite often (e.g. 16% and 9% of @mauricedehond's tweets contained disinformation and conspiracies, respectively; and 12.5% and 7.7% of @blckbxnews' messages had conspiracies and rumors in them). These results align with our argument that the main goal of these actors is to inject toxicity into online environments, and that each type of deceitful content is simply one of many tools in the toolbelt of DOLs. In addition, the results emphasize that a user-centric (rather than, or in combination to, a content-centric and compartmentalized) analysis is needed to have a more clear understanding of the spread of deceitful content on SM and their effects.

To test hypotheses 1, 2 and 3, we turn to the set of new followers for which we had collected enough information to explore a potential change in behavior after following the first DOL in our sample (N = 5,574). In Figure 2 we show the average number of political, uncivil, and affectively polarized tweets these users sent during the 30 days before (vs. after) following the first DOL. We see

starck differences across the board. The users were more politically engaged (sending 36.2 political tweets in the 30 days after, vs. 19.3 political tweets in the 30 days prior), more uncivil (9 vs. 4.3 uncivil tweets) and affectively polarized (12.3 vs. 6.6 polarizing tweets).

Figure 2. Average number of political, uncivil, and affectively polarized tweets (plus 95% Confidence Interval), sent during the 30 days before (vs. after) following the first Deceitful Opinion Leader (DOL).



Given that we collected the data during an election period, we wanted to control for whether a user started following a DOL before the election day (as users may have been more likely to discuss politics during the *after* time window). Hence, we created the variable *Campaign Post Days,* which accounts for the number of *post* 15/30 days that overlapped with the electoral campaign (so the number of days between the day a user started following the first DOL and election day, March 17). This variable is 0 for those who started following a deceitful opinion leader after March 17. As specified in Model 1, for a more clear test of our hypotheses we use linear models predicting the difference ($Ypost - Ypre$) for three outcomes of interest (number of uncivil, affectively polarized, and political tweets) as a function of the mentioned control variable *Campaign Post Days*. For each of these linear models, the intercept parameter (α) provides information about the average difference in messaging behavior between the post and pre difference after accounting for the control variable.

Model 1: Model specification used to test Hypotheses 1, 2, and 3

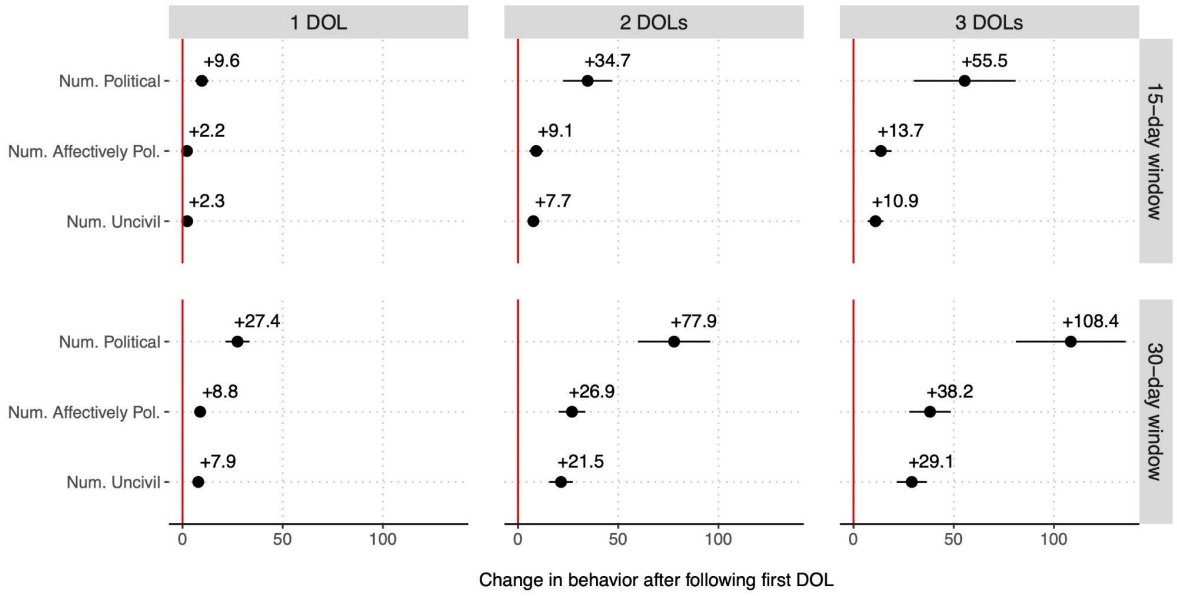
$$(Ypost - Ypre) = \alpha + \beta\_CampaigPosDays + \varepsilon$$

In Figure 3 we report the α coefficient for several linear models. For each of the four outcomes of interest, we ran six models with the same specification (i.e. Model 1), where we varied the time window to calculate the post/pre periods (15 and 30-day windows), and the number of DOLs the user followed within the 15/30 days after following the first opinion leader. In the first column (*1 DOL*) we include all the users in our sample (N = 5,574/3,891), and in the other columns we estimate the models using only those users who followed at least a second DOL (*2 DOLs*) within the next 15/30 days (N = 1,336/1,014), and at least a third DOL (*3 DOLs)* (N = 555/421); i.e. each analysis includes the number of unique Twitter users that meet the criteria. These variations allow us to disentangle differential effects across time (whether we observe stronger effects when comparing 15 vs. 30 days), and also across different levels of engagement (e.g. users who decided to follow more than one of the DOLs in our sample).

We find strong support for our three hypotheses. Across the board, we see an increase in the number of political, uncivil and affectively polarized tweets. All estimates presented in Figure 3 are statistically significant at the conventional 0.05 level. We observe the mildest effects among those who only followed 1 of the DOLs in our sample. But even among those, we observe a substantial change in behavior, particularly when we compare the behavior during the 30 days after (vs. before)

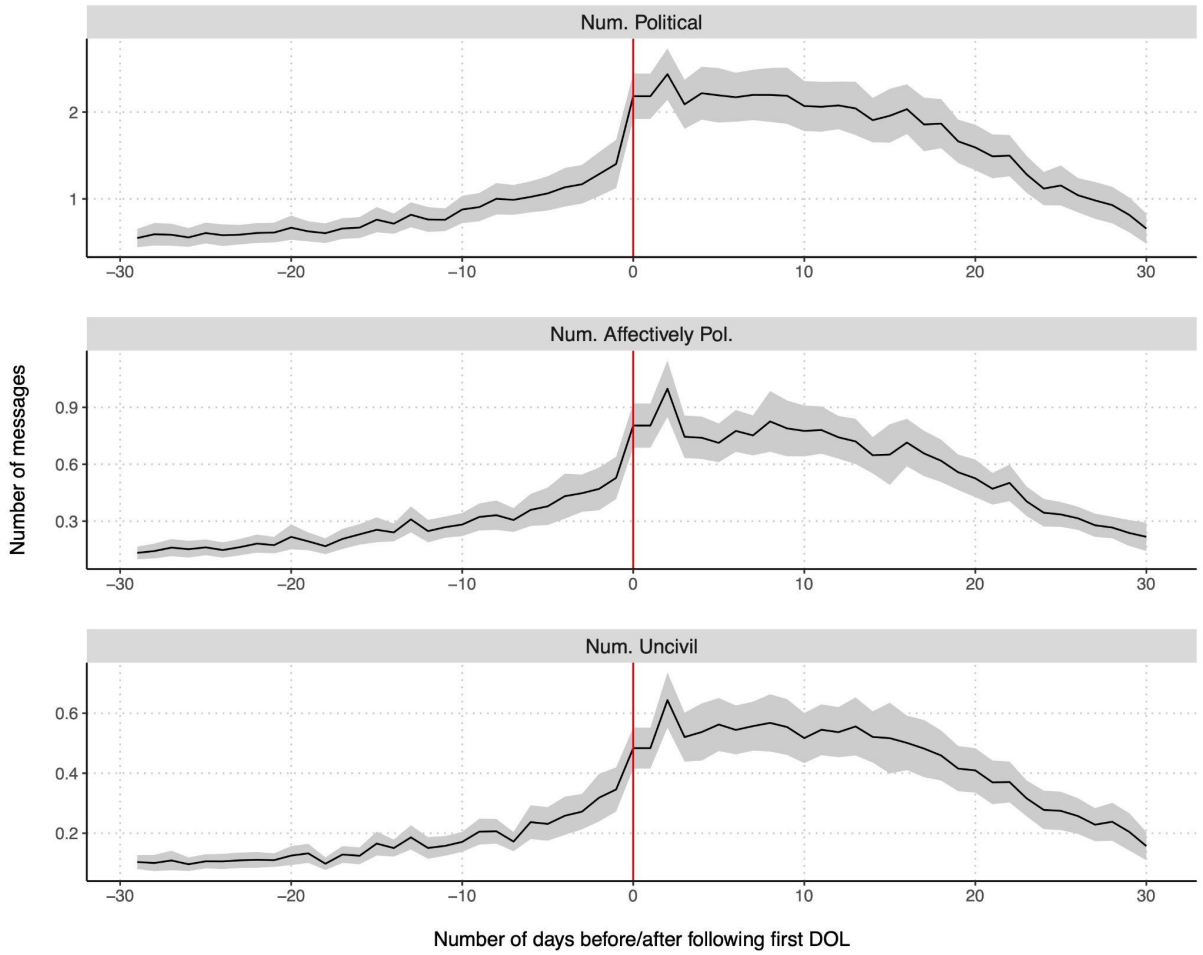
following the DOL. On average, these users sent 27.4 more political messages, 8.8 more affectively polarized messages, and 7.9 more uncivil ones. We observe the strongest effects among those who followed a second and a third DOL during the 30 days after following the first DOL in our sample (N = 421). On average, they radically sent more political (+108.4), affectively polarized (+38.2), and uncivil (+29.1) messages. These findings are not driven by the new followers of one particular DOL, but are a reflection of a general pattern observed across the followers of the different DOLs in our sample (see Appendix D). In addition, this change in behavior cannot be simply explained by these users retweeting messages originally posted by the DOLs they started following (see Appendix A). On average (95% confidence intervals included), only 0.8% (0.6-1%) of the political tweets, 0.5% (0.3-0.7%) of the uncivil messages, and 0.5% (0.3-0.7%) of the affectively polarized tweets they sent during the 30 days after following the first DOLs are retweets of that DOL.

Figure 3. Coefficients (+95% confidence intervals) from linear models estimating a change in behavior after following one, two, and three Deceitful Opinion Leaders (DOL).



For a more detailed picture, in Figure 4 we explore the functional form of these effects. The figure shows the average number of political, uncivil and affectively polarized tweets (+95% confidence intervals) the users in our sample sent each of the 30 days before and after following the first DOL. Figure 4 is standardized, so the exact date of day 0 differs across users, depending on when they started following the DOL. We observed a slight upper trend right before they started following the DOL. This indicates that at least some users already started shaping their behavior prior to day 0. This could be because they may have already been exposed to some tweets from these DOLs via retweets from their networks, or some factor motivated them to change their behavior and potentially to seek for these kinds of opinion leaders. Then, we observe a clear jump at the moment the users started following the first DOL. The number of overall tweets and the uncivil, affectively polarized, and political ones remained high for about 15-20 days. After that period, the behavior of the users gradually reverted back to their levels of activity prior to following the DOL. The patterns described in Figure 4 clearly point to these DOLs playing a crucial role in the radicalization of online environments. Independently of what motivated these users to start following these DOLs – whether it was a very intentional decision, or as a result of incidental exposure via retweets from one's networks – we observe starck and substantive changes in behavior that contribute to increase the levels of toxicity and incivility on the SM platform.

Figure 4. Average number of political, uncivil and affectively polarized tweets sent by followers of DOLs, during the 30 days before and after following the first DOL (N = 3,891).



Conclusion & Discussion

This paper tackles three shortcomings of existing literature studying the dissemination of deceitful content. First, existing literature is very compartmentalized, as it mostly focuses on one type of deceitful content (e.g. conspiracies, fake news or misleading information). We show that in reality, salient social media (SM) accounts engage in the spread of all sorts of deceitful content throughout. Each type of deceitful content that they disseminate is just one of many tools in their toolbelt. Second, we lack an overarching approach that puts these influential SM users at the center, instead of the content that they spread. We do so, by putting forward a new theoretical concept: *Deceitful Opinion Leaders* (DOLs). Third, this study contributes to the understanding of the individual-level effects that these types of deceiving messages have on other SM users. We show that after following a DOL on Twitter, significant behavioral changes start to occur amongst their followers: users send more political, uncivil and affectively polarized messages. For example, on average the analyzed users sent around 28 political tweets, 8 uncivil tweets and 9 affectively polarized tweets more during the 30 days after following a DOL, compared to the 30 days prior. These behavioral changes seem to gradually revert back to their levels of activity prior to following the DOL. Although at the individual level these behavioral changes do not last long, at the aggregate level these effects have a substantive

impact: DOLs gather new followers every day, meaning that these behavioral effects are constantly occurring, having a longer-lasting effect on the behavior and norms of conversation on Twitter.

Although this paper adds important results to existing literature, it is not without limitations. This paper provides a first aim in studying the effects of DOLs on SM platforms. There are other influential DOLs who were not included in this research. Moreover, all of the DOLs in this study are Dutch. Hence, this study only focuses on the Dutch SM landscape. Furthermore, this study only considers Twitter, while DOLs are active on many platforms. To assess the generalizability of the effects that DOLs have on other SM users, future studies should aim to address additional factors that influence these findings, such as platform affordances and the level of radicalization of a platform. We expect the work presented here to inspire future work focusing on a more comprehensive and representative sample of DOLs from different contexts on different platforms, in order to provide further insights into the conditions under which these opinion leaders shape our online environments. Despite these limitations, this research finds valid and important results that show significant individual-level effects from following DOLs who engage in the spread of deceitful content online. Even though research finds that only a small proportion of SM users spread deceitful content *per se* (e.g. Guess et al. 2019), the spread of deceitful content via SM leads to substantial effects on other users in the platform.

The results of this study provide a first look into the distribution of the spread of deceitful content by DOLs, and the individual-level effects that DOLs have on their followers. Importantly, this study adds to the empirical evidence of the effects of deceitful content on SM users. The findings of this study add to existing literary knowledge of the consequences of deceitful content in online environments. In addition, the results of this study provide empirical evidence to the societal debate whether or not these influential SM users should be removed in order to maintain a healthy forum for public debate. Removing DOLs from Twitter would reduce toxicity on the platform. However, doing so might have negative effects if DOLs move on to other platforms to spread their deceiving content. This might result in higher levels of radicalization and polarization. Especially on Telegram which is known to have a high number of users that support conspiracy theories. Furthermore, these findings underline that a small group of DOLs change the norms of conversation on SM platforms. Hence, this accentuates the need for future research to study the literary concept of DOLs.

References

Balkin, J. M. (2017). Free speech in the algorithmic society: big data, private governance, and new school speech regulation. *UCDL Rev*., 51, 1149.

Barberá, P. (2015). Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data. *Political Analysis*, *23*(1), 76–91. https://doi.org/10.1093/pan/mpu011

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber? *Psychological Science*, *26*(10), 1531–1542. https://doi.org/10.1177/0956797615594620

Baum, M. A. (2002). Sex, Lies, and War: How Soft News Brings Foreign Policy to the Inattentive Public. *American Political Science Review*, *96*(1), 91–109. https://doi.org/10.1017/S0003055402004252

Bennett, W. L., & Livingston, S. (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication*, *33*(2), 122–139. https://doi.org/10.1177/0267323118760317

Bergmann, E. (2020). Populism and the politics of misinformation. *Safundi*, *21*(3), 251–265. https://doi.org/10.1080/17533171.2020.1783086

Bergström, A., & Jervelycke Belfrage, M. (2018). News in Social Media: Incidental consumption and the role of opinion leaders. *Digital Journalism*, *6*(5), 583–598. https://doi.org/10.1080/21670811.2018.1423625

Berr, J. (2019, June 25). *Conspiracy Theorist Alex Jones Evades Crackdown On Social Media Sites*. Forbes. Retrieved May 2, 2022, from https://www.forbes.com/sites/jonathanberr/2019/06/24/conspiracy-theorist-alex-jones-evades-crackdown-on-social-media-sites/?sh=1809755a4f87

Carlson, T. N. (2019). Through the Grapevine: Informational Consequences of Interpersonal Political Communication. *American Political Science Review*, *113*(2), 325–339. https://doi.org/10.1017/S000305541900008X

Chadwick, A., & Stanyer, J. (2022). Deception as a Bridging Concept in the Study of Disinformation, Misinformation, and Misperceptions: Toward a Holistic Framework. *Communication Theory*, *32*(1), 1-24. doi:10.1093/ct/qtab019

Choi, S. (2015). The Two-Step Flow of Communication in Twitter-Based Public Forums. *Social Science Computer Review*, *33*(6), 696–711. https://doi.org/10.1177/0894439314556599

Coaston, J. (2018, August 6). *Alex Jones banned from YouTube, Facebook, and Apple, explained*. Vox. Retrieved May 2, 2022, from https://www.vox.com/2018/8/6/17655658/alex-jones-facebook-youtube-conspiracy-theories

Davidson, T., Macy, M., Warmsley, D., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Arxiv*, *(2017 03 11)*.

DiFonzo, N., & Bordia, P. (2007). Rumor, Gossip and Urban Legends. *Diogenes*, *54*(1), 19–35. https://doi.org/10.1177/0392192107073433

Douglas, K. M., Uscinski, J. E., Sutton, R. M., Cichocka, A., Nefes, T., Ang, C. S., & Deravi, F. (2019). Understanding Conspiracy Theories. *Political Psychology*, *40*(S1), 3–35. https://doi.org/10.1111/pops.12568

EenVandaag. (2018, August 8). *Complotdenker Alex Jones verbannen van bijna alle sociale media*. Retrieved May 2, 2022, from https://eenvandaag.avrotros.nl/item/complotdenker-alex-jones-verbannen-van-bijna-alle-sociale-media/

Egelhofer, J. L., & Lecheler, S. (2019). Fake news as a two-dimensional phenomenon: A framework and research agenda. *Annals of the International Communication Association*, *43*(2), 97–116. https://doi.org/10.1080/23808985.2019.1602782

Friggeri, A., Adamic, L., Eckles, D., & Cheng, J. (2014, May). Rumor cascades. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 8, No. 1).

Gervais, B. T. (2015). Incivility Online: Affective and Behavioral Reactions to Uncivil Political Posts in a Web-based Experiment. *Journal of Information Technology & Politics*, *12*(2), 167–185. https://doi.org/10.1080/19331681.2014.997416

Groshek, J., & Koc-Michalska, K. (2017). Helping populism win? Social media use, filter bubbles, and support for populist presidential candidates in the 2016 US election campaign. *Information, Communication & Society*, *20*(9), 1389–1407. https://doi.org/10.1080/1369118X.2017.1329334

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, *5*(1), eaau4586. https://doi.org/10.1126/sciadv.aau4586

Harteveld, E. (2021). Fragmented foes: Affective polarization in the multiparty context of the Netherlands. *Electoral Studies*, *71*, 102332. https://doi.org/10.1016/j.electstud.2021.102332

Haselton, T. (2019, May 3). *Alex Jones was banned from Facebook, but an hour later he was back on Facebook livestreaming*. CNBC. Retrieved May 2, 2022, from https://www.cnbc.com/2019/05/02/alex-jones-banned-from-facebook-but-hes-already-back.html

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The Origins and Consequences of Affective Polarization in the United States. *Annual Review of Political Science*, *22*(1), 129–146. https://doi.org/10.1146/annurev-polisci-051117-073034

Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, Not Ideology. *Public Opinion Quarterly*, *76*(3), 405–431. https://doi.org/10.1093/poq/nfs038

Katz, E., Lazarsfeld, P. F., Columbia University. Bureau of Applied Social Research, & Lazarsfeld, P. F. (1955). *Personal influence: the part played by people in the flow of mass communications* (Ser. Foundations of communications research). Free Press.

Kim, J. W., Guess, A., Nyhan, B., & Reifler, J. (2021). The Distorting Prism of Social Media: How Self-Selection and Exposure to Incivility Fuel Online Comment Toxicity. *Journal of Communication*, *71*(6), 922–946. https://doi.org/10.1093/joc/jqab034

Kim, Y., Chen, H.-T., & Gil de Zúñiga, H. (2013). Stumbling upon news on the Internet: Effects of incidental news exposure and relative entertainment use on political engagement. *Computers in Human Behavior*, *29*(6), 2607–2614. https://doi.org/10.1016/j.chb.2013.06.005

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

McKay, S., & Tenove, C. (2021). Disinformation as a Threat to Deliberative Democracy. *Political Research Quarterly*, *74*(3), 703–717. https://doi.org/10.1177/1065912920938143

Miller, B., Linder, F., & Mebane, W. R. (2020). Active Learning Approaches for Labeling Text: Review and Assessment of the Performance of Active Learning Approaches. *Political Analysis*, *28*(4), 532–551. https://doi.org/10.1017/pan.2020.4

Paul, K. (2019, May 3). *Facebook bans Alex Jones, other extremist figures*. U.S. Retrieved May 2, 2022, from https://www.reuters.com/article/us-facebook-extremists-usa-idUSKCN1S82D7

Popan, J. R., Coursey, L., Acosta, J., & Kenworthy, J. (2019). Testing the effects of incivility during internet political discussion on perceptions of rational argument and evaluations of a political outgroup. *Computers in Human Behavior*, *96*, 123–132. https://doi.org/10.1016/j.chb.2019.02.017

Showredactie. (2020, October 21). *YouTube verdedigt verwijderen account Lange Frans: 'Richtlijnen meermaals geschonden.'* AD.nl. Retrieved May 2, 2022, from https://www.ad.nl/show/youtube-verdedigt-verwijderen-account-lange-frans-richtlijnen-meermaals-geschonden~a5606d7d/

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter*, *19*(1), 22–36. https://doi.org/10.1145/3137597.3137600

Spohr, D. (2017). Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Business Information Review*, *34*(3), 150–160. https://doi.org/10.1177/0266382117722446

Stroud, N. J. (2008). Media Use and Political Predispositions: Revisiting the Concept of Selective Exposure. *Political Behavior*, *30*(3), 341–366. https://doi.org/10.1007/s11109-007-9050-9

Suler, J. (2004). The Online Disinhibition Effect. *CyberPsychology & Behavior*, *7*(3), 321–326. https://doi.org/10.1089/1094931041291295

Sunstein, C. R., & Vermeule, A. (2009). Conspiracy Theories: Causes and Cures*. *Journal of Political Philosophy*, *17*(2), 202–227. https://doi.org/10.1111/j.1467-9760.2008.00325.x

Tajfel, H., Turner, J. C., Austin, W. G., & Worchel, S. (1979). An integrative theory of intergroup conflict. *Organizational identity: A reader*, *56*(65), 9780203505984-16.

Terechshenko, Z. and Linder, F. and Padmakumar, V. and Liu, F. and Nagler, J. and Tucker, J.A. and Bonneau, R. (2020). A Comparison of Methods in Political Science Text Classification: Transfer Learning Language Models for Politics. *Available at SSRN*: http://dx.doi.org/10.2139/ssrn.3724644

Theocharis, Y., Barberá, P., Fazekas, Z., Popa, S. A., & Parnet, O. (2016). A Bad Workman Blames His Tweets: The Consequences of Citizens' Uncivil Twitter Use When Interacting With Party Candidates: Incivility in Interactions With Candidates on Twitter. *Journal of Communication*, *66*(6), 1007–1031. https://doi.org/10.1111/jcom.12259

Van Aelst, P., Strömbäck, J., Aalberg, T., Esser, F., de Vreese, C., Matthes, J., Hopmann, D., Salgado, S., Hubé, N., Stępińska, A., Papathanassopoulos, S., Berganza, R., Legnante, G., Reinemann, C., Sheafer, T., & Stanyer, J. (2017). Political communication in a high-choice media environment: A challenge for democracy? *Annals of the International Communication Association*, *41*(1), 3–27. https://doi.org/10.1080/23808985.2017.1288551

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

de Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., and Nissim, M. (2019). BERTje: A Dutch BERT Model.. *arXiv*. https://doi.org/10.48550/ARXIV.1912.09582

Weeks, B. E., & Gil de Zúñiga, H. (2021). What's next? Six observations for the future of political misinformation research. *American Behavioral Scientist*, *65*(2), 277-289. doi:10.1177/0002764219878236

Weeks, B. E., Lane, D. S., Kim, D. H., Lee, S. S., & Kwak, N. (2017). Incidental Exposure, Selective Exposure, and Political Information Sharing: Integrating Online Exposure Patterns and Expression on Social Media. *Journal of Computer-Mediated Communication*, *22*(6), 363–379. https://doi.org/10.1111/jcc4.12199

Yarchi, M., Baden, C., & Kligler-Vilenchik, N. (2021). Political Polarization on the Digital Sphere: A Cross-platform, Over-time Analysis of Interactional, Positional, and Affective Polarization on Social Media. *Political Communication*, *38*(1–2), 98–139. https://doi.org/10.1080/10584609.2020.1785067

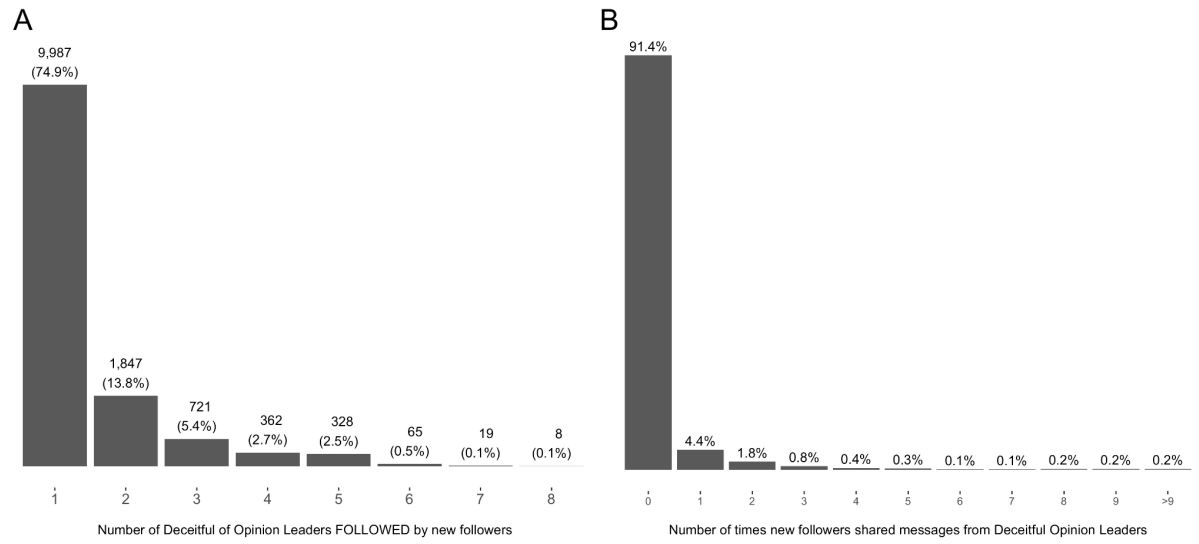
Zaller J. (1992). *The nature and origins of mass opinion*. Cambridge University Press.

Appendix A. Number of DOLs followed and number of DOL tweets shared.

To provide some further descriptives on the sample of followers we study, in Figure A.1 we show the number of DOLs followed by all new followers (N = 13,337) and the number of times these followers shared messages of DOLs. Figure A.1.A displays how many DOLs the new followers started to follow during data collection. It indicates that 74,9% followed one DOL, and 25,1% followed two or more DOLs. In other words, during a time period of 2 months, almost one third of new followers started to follow more than one DOL. This finding supports existing literature, as it indicates that DOLs and their followers form homogeneous social networks online (Barbéra 2015; Shu et al., 2017).

Figure A.1.B displays the percentage of followers that shared messages from a DOL. The Figure shows that the majority (91,4%) of users in this dataset shared 0 tweets of DOLs. The other 8,6% that did share messages of a DOL, mostly only shared 1 message (4,4%). These findings align with other work studying the dissemination of deceitful content. For example, Guess et al. (2019) found that 8,5% of the respondents in their study shared fake news articles to their Facebook friends. This finding illustrates that, in line with previous literature (Guess et al., 2019; Vosoughi et al., 2018), only a small proportion of SM users contribute to the dissemination of misleading content online.

Figure A.1. Number of DOLs followed by new followers (A), as well as the number of times they shared a message from a DOL (B).



Appendix B. Codebook for Types of Disinformation.

Fake news

- Has a journalistic format, but is low in facticity (Egelhofer & Lecheler, 2019)
- Everything that looks like a reel news article, but comes from sources that aren't (always) factual. Such as the DOL's themselves.
- Open the links in tweets to see the articles.
- When a tweet is fake news, it can not be disinformation as well.

Disinformation

- False information that is purposely spread to deceive people, seeking to amplify social divisions and distrust (Bennett & Livingston, 2018; McKay & Tenove 2020)
- Similar to fake news, but does not always have a link to an article with a journalistic format.
- Sometimes an article of a legitimate news source is linked, but spinned or interpreted in a way that makes the informatie not factual.

Conspiracy

- Efforts to explain events, practices or secret plots that consist of two or more powerful actors, acting in secret for their own benefit and working towards a malevolent or unlawful goal against the common good (Douglas et al., 2019; Sunstein & Vermeule, 2009).
- Everything that creates an us against them feeling, or hints at the elite being corrupt or misleiding etc.

Rumor

- Circulating information whose veracity status is yet to be verified at the time of spreading (DiFonzo & Bordia, 2007; Friggeri et al., 2014)
- Whenever tweets contain frases as: if they will do X then …, if X is true then …, they are saying that X will happen …, etc.

Appendix C. Codebook for Political, Uncivil and Affective Polarization Content.

Political variable

- Binary variable (0 = not politics | 1 = politics
- Use a broad understanding of what's political: not only when political parties/institutions are mentioned, but also when topics that may have political implications, be picked up by media/politicians, eventually be part of a policy change, etc.
- If text field is empty, add NAs to all coding columns/variables
- If a tweet is from a political leader but doesn't say anything about the political party or about politics in general, code it as 0. For instance, if Jesse Klaver tweets what he had for breakfast this morning. If the party is mentioned, code it as political. So, for instance if Jesse Klaver tweets: *This morning I had a lot of veggies for breakfast #GroenLinks*
- Code foreign languages via translate

Uncivil language

- Binary variable (0 = no | 1 = yes)
- Insulting, harassing, cursing words, racist, misogynist, very dismissive towards others, against a (minority) group: immigrants, LGBTIQ community, women
- "Vote them out" → not uncivil, but is negative
- Uncivil → Think about if you are Twitter or Facebook would you ban this tweet?
- Opinions that are factual/neutral aren't uncivil for example: *The U.S. Capitol Was Attacked by a Nation of Islam Follow and "GOP Leader" Kevin McCarthy Hasn't Said a Word About It. It's Time for Him to Resign. https://t.co/wSCWUcAsj9*

Affectively polarized

- It has to show dislike towards the opposing group, it has to be negative feelings towards the outgroup (affectively polarized)
- Check the nicknames for politicians, ideologic groups and newspapers GeenStijl uses for political groups
- Needs to mention a political party or person or group or left/right (ideology)
- When tweets are about politicians in general, but it's not clearly directed at a political person, a political group etc, we code it as NA. For example: *Aan @Tjeenk Willink wat zijn wij Nederlanders zielig om te denken aan Rutte.wat nu als ze hem van zijn fiets rijden in Den Haag en als gevolg hiervan overlijd hij ,heeft Nederland een fucking probleem????gewoon zielig.*
- When a tweet is directed at a political person and uncivil we code it as AP for example: *@hugodejonge @GGD_RR Stumpert, je houdt geen 1.5m afstand. Leugenaar.*
- When groups are posed against each other and politicians in general or by name, or political parties are mentioned

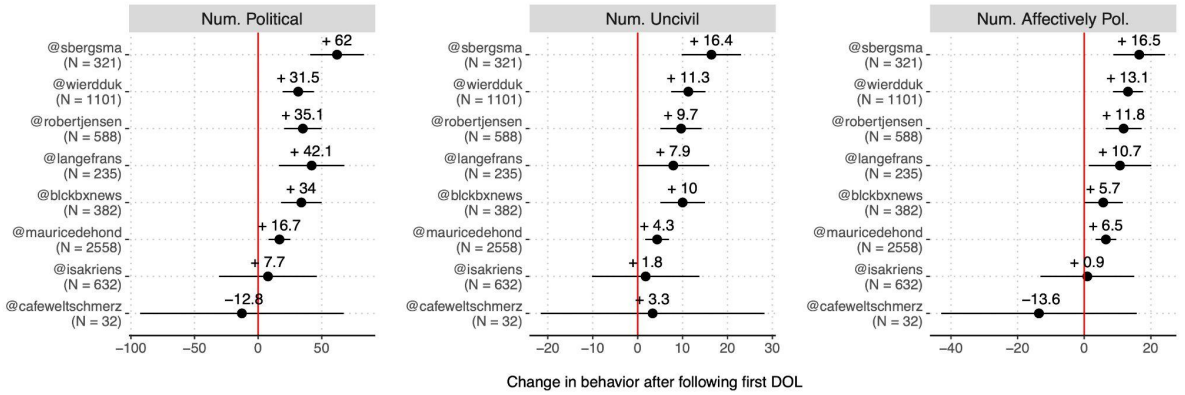
Misleading content

- Binary variable (but one tweet can be multiple types of misleading content) (0 = no | 1 = yes)
- We do not code videos, therefore we use NA when a tweet links to a video. Or we code it based on the text.
- If a tweet has been deleted, we code it as NA

Appendix D. Models with DOL-fixed effects and heterogenous effects by DOL.

In this Appendix we provide further details about the change in behavior of new DOL followers after following the first DOL in our sample. In particular, we are interested in disentangling whether the effects observed in the paper are mainly driven by following a particular DOL, or whether these are patterns consistent among the new followers of all the eight DOLs in our sample.

Figure D.1. Coefficients (+95% confidence intervals) from linear models estimating a change in behavior after following the first Deceitful Opinion Leader (DOL). Same model specification than the models reported in Figure3, but with the inclusion of DOL-fixed effects.



In Figure D.1 we report the results for models similar to those in Figure 3, but with the difference that in these new models we include DOL-fixed effects. In the figure we report the results for the fixed effects, which indicate the difference between the messages sent in the 30 days after (vs. before) following each particular DOL.

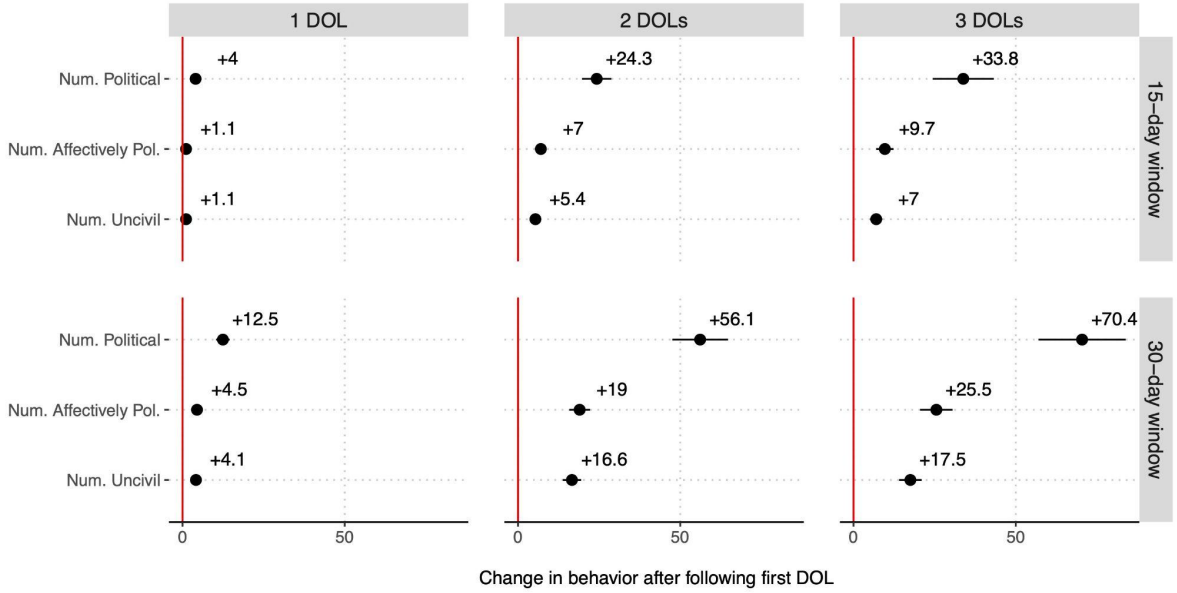
We observe positive, significant and substantive effects for the users in our sample who decided to follow most of these DOLs, with two minor exceptions: we find positive but not significant results for the 632 users who started following @isakriens, and we find clear null results for the 32 users in our sample who started following @cafeweltschmerz. The followers of the remaining 6 DOLs clearly increased the volume of tweets in general during the 30 days after they started following them, as well as the number of political, uncivil and affectively polarizing tweets. Although we observe some mild variations (for example, the 321 that started following @sbergsma are the ones who changed their behavior more substantially), overall we find very similar results. In sum, these findings indicate that the results reported in the paper are not driven by those who started following a particular DOL, but are a reflection of a general pattern that happened across those who followed the different DOLs in our sample.

Appendix E. Analysis after excluding outliers.

In this Appendix we replicate Figures 3 and 4 of the paper, where we explore changes in behavior among new DOL followers, but in this case we assess the robustness of the original findings by excluding potential outliers in the data.

In Figure E.1, which replicates Figure 3, we used Cooks Distance (Cook, 1997) to remove any potential outlier from the regression models. For each of the models, we first calculated the same regression reported in Figure 3, then we calculated the Cooks Distance for each of the regression observations, and we finally estimated a new regression model after excluding the observations with a Cooks Distance that was four times larger than the average distance. In Figure E.1 we observe the size of the effects to be smaller once we remove these potential outliers. For example, when using a 15-day window to explore changes in behavior after following the first DOL (top-left panel), we find that users tweeted 4 more political tweets, 1.1 more uncivil messages, and 1.1 more affectively polarized tweets during the 15 days after following the DOL (compared to the previous 15 days), rather than 9.6, 2.3, and 2.2 respectively, as reported in Figure 3. Nevertheless, we still observe effects of a substantive magnitude, particularly if we look at the effects for longer time periods (e.g. 30-day window) and when following more than one DOL; and all the statistical findings reported in Figure 3 hold when excluding these potential outliers.

Figure E.1. Replication of the model results reported in Figure 3, after excluding regression outliers (those with a Cooks Distance 4 times larger than the average).



In Figure E.2 we replicate Figure 4 of the paper after removing from the dataset those users considered as outliers when estimating the four regressions reported in the bottom-left panel in Figure E.1 (so those considered as outliers when estimating a change in the political, uncivil, and affectively polarized messages; 30 days after following the first DOL). Note that a given user may be considered an outlier for one of the models (e.g. change in number of political tweets) but not the others, and so information from a given user may be excluded when calculating the moving averages from some panels in Figure E.2, but not the others. Similar to the findings in Figure E.1, the overall volume of tweets reported in Figure E.2 are slightly lower than those reported in Figure 4. Nevertheless, we observe the same exact patterns once these potential outliers are removed.

Figure E.2. Average number of political, uncivil and affectively polarized tweets sent by followers of DOLs, during the 30 days before and after following the first DOL. Potential outliers have been removed from the data.