Civilized truths, hateful lies? Incivility and hate speech in false information – evidence from fact-checked statements in the US

Hameleers, M.; van der Meer, T.; Vliegenthart, R.

DOI
10.1080/1369118X.2021.1874038

Publication date
2022

Document Version
Final published version

Published in
Information, Communication & Society

License
CC BY-NC-ND

Link to publication

Citation for published version (APA):
Hameleers, M., van der Meer, T., & Vliegenthart, R. (2022). Civilized truths, hateful lies? Incivility and hate speech in false information – evidence from fact-checked statements in the US. *Information, Communication & Society*, *25*(11), 1596-1613. https://doi.org/10.1080/1369118X.2021.1874038

INFORMATION, COMMUNICATION & SOCIETY
2022, VOL. 25, NO. 11, 1596–1613
https://doi.org/10.1080/1369118X.2021.1874038

Routledge
Taylor & Francis Group

OPEN ACCESS

Civilized truths, hateful lies? Incivility and hate speech in false information – evidence from fact-checked statements in the US

Michael Hameleers, Toni van der Meer and Rens Vliegenthart

Amsterdam School of Communication Research, University of Amsterdam, Amsterdam, the Netherlands

ABSTRACT

Digital information settings may not only offer an opportunity structure for democratic deliberation, but also facilitate the occurrence of negative phenomena – such as incivility, hate speech and false information. Even though extant literature has provided theoretical arguments for a discursive affinity between false or deceptive information and uncivil speech, we lack empirical evidence on whether and how false information and incivility converge. Against this backdrop, we rely on an extensive content analysis of fact-checked statements in the US (*N* = 894) to assess to what extent and how different forms of incivility are present in different degrees of false information. Our main findings illustrate that partisan attacks, negativity, and hate speech are most likely to occur in false information that deviates furthest from reality. These findings help us to dissect different degrees of untruthfulness based on their content features: Disinformation (goal-directed deception) may be distinguished from misinformation (unintentionally misleading content) based on the centrality of hostility, partisan attacks, and hate speech in the former.

ARTICLE HISTORY
Received 2 July 2020
Accepted 5 January 2021

KEYWORDS
Disinformation; hate speech; misinformation; (online) incivility; partisan attacks

Although digital information settings may promote political participation and deliberative debates among citizens (e.g., Habermas, 2006), online information settings also cultivate depersonalized spaces or echo chambers where incivility and communicative untruthfulness thrive and get amplified (e.g., Lowry et al., 2016). In this paper, we integrate research on two important threats of online news settings: The uncontrolled spread of false information on the one hand (e.g., Tandoc et al., 2018; Wardle, 2017) and hate speech and incivility on the other hand (e.g., Lowry et al., 2016). By assessing how incivility occurs across different types of false information in the US, we aim to offer important new insights into the relationship between uncivil speech and the dissemination of false or misleading information in today's fragmented information settings.

CONTACT Michael Hameleers m.hameleers@uva.nl Nieuwe Achtergracht 166, 1018 WV Amsterdam, the Netherlands

Supplemental data for this article can be accessed https://doi.org/10.1080/1369118X.2021.1874038

Misinformation is an umbrella-term for all information that is false and/or deceptive, or not based on relevant expert knowledge (Vraga & Bode, 2020) – it may also contain more extreme forms of politically motivated deception and intentionally misleading information (disinformation). Hate speech refers to any form of communication in which others are attacked, denigrated or intimidated on the basis of religion, ethnicity, gender, national origin or another group-based trait (e.g., Warner & Hirschberg, 2012). In this paper, we argue that incivility, out-group attacks, media critique and hate speech may be important content features that set more extreme forms of false information apart from less severe deviations from facticity. Attacking out-groups and expressing uncivil speech is unlikely to be accidental, and likely to be motivated by a political agenda – a defining feature of disinformation (Freelon & Wells, 2020). Although fact-checking platforms typically do not include the distinction between mis- and disinformation in their rating scales, we argue that politically motivated, partisan or ideological utterances in false information, such as hate speech and incivility, may be an indicator of disinformation.Disinformation may thus be distinguished from misinformation based on indicators of incivility that deliberately attribute negative qualities to out-groups.

Hate speech has oftentimes been studied in the context of (radical) right-wing populist politicians' communication (e.g., Van Spanje & De Vreese, 2014). These actors are also increasingly associated with the uncontrolled spread of disinformation (e.g., Bennett & Livingston, 2018; Marwick & Lewis, 2017). It can be argued that offensive language and hate speech strongly resonate with the politics of (partisan) disinformation. Specifically, in hate speech, a Manichean outlook between the in-group and out-groups is cultivated. These out-group descriptions are based on stereotypes and are founded on negative associations rather than empirical evidence or expert knowledge. In addition, hostile and damaging terms are used to describe the other, which lack an empirical basis. Communicators of hate speech may *deliberately* disseminate incorrect information related to out-groups to cultivate polarized divides in society and to create support for their radical right-wing issue positions.

To empirically assess the relationship between mis- and disinformation and incivility or hate speech, we rely on a manual content analysis of different types of false, compared to true, information flagged by independent fact-checkers in the US ($N = 894$) – supplemented by an exploratory qualitative content analysis that looks at incivility in general false information and intentionally or politically motivated disinformation. With this study, we aim to provide a comprehensive understanding of how hate speech may take shape in information settings characterized by post-factual relativism and disinformation (e.g., Van Aelst et al., 2017).

The discursive affinity between misinformation and incivility

Misinformation can be defined as any type of false or inaccurate information (e.g., Tandoc et al., 2018; Vraga & Bode, 2020; Wardle, 2017). In this paper, we contend that different degrees of false or untrue information can be mapped on a continuum that forms a 'space of untruthfulness', ranging from completely true to completely false information. We specifically distinguish between four main types of false information flagged by international examples of fact-checkers and conceptualized in empirical research

(Humprecht, 2018): (1) completely true information; (2) mostly true information; (3) mostly false information; and (4) completely false information.

There is growing consensus that misinformation research should take the political consequences and wider contextual backdrop in which it is spread into account (Bennett & Livingston, 2018). Most literature on misinformation points to the oftentimes partisan nature of false information (Pennycook & Rand, 2019) or the radical right-wing political agenda associated with the dissemination of false or misleading information (e.g., Bennett & Livingston, 2018; Marwick & Lewis, 2017). In line with this argument on the partisan nature and political agenda of false information, disinformation – an extreme or politically motivated form of misinformation – has been defined as false information that is disseminated intentionally, for example to influence electoral outcomes or to steer public opinion in a pre-defined way (Freelon & Wells, 2020; Tandoc et al., 2018; Wardle, 2017).

One particular agenda of disinformation may be to strengthen partisans' or issue publics' opposition to out-groups or minorities by communicating violent or hateful sentiments targeted at opposed groups (Bennett & Livingston, 2018; Marwick & Lewis, 2017). This illustrates that hate speech, partisan attacks and other forms of incivility could be seen as important indicators of disinformation. Hate speech and other deviations from civility are inherently un-factual: If facts and verified evidence implicitly depict certain groups in a negative way due to factual and/or observable traits, it should not be regarded as uncivil (hate) speech. Hence, negative speech associated with untruthfulness is likely to be goal-directed and intentional. In this paper, we therefore postulate that the interaction between communicative untruthfulness and hate speech corresponds to the politics of *disinformation*.

Expressing uncivil speech may imply that the truth is circumvented. Disseminating civil information about certain out-groups or merely describing the actions of politicians is likely to closely depict reality, without resorting to hostile or uncivil tones. However, when out-groups or partisan opponents are referred to in an uncivil way, (stereotypical) attributes are associated with these actors – a practice that deviates from empirical facts or truth-telling. Specifically, (negative) stereotyping creates an alternative reality in which people are categorized based on appearing similarities that may not exist in real-life, a categorization process driven by prejudice, but not facts (Katz & Braly, 1933). Hence, even though we cannot establish a strong causal relationship between false information and incivility based on content features alone, it can be argued that the process by which groups or individuals are categorized into stereotypical groups or assigned negative traits are likely to deviate from empirical facts or expert knowledge.

Based on this association – and the established relationship between disinformation and the negative stereotyping of out-groups in radical right-wing politics (Bennett & Livingston, 2018; Marwick & Lewis, 2017), we conceptualize the link between false information and incivility in two ways. First, the act of negatively stereotyping out-groups or devaluating (political) actors by name-calling, swearing, or attributing negative traits (Anderson et al., 2014; Chen & Lu, 2017; Coe et al., 2014; Papacharissi, 2004) is based on projection, prejudice, categorization and de-humanization – which does not have an empirical basis. Second, actors that express uncivil speech to targets have to resort to untruthfulness, as they are unlikely to ground this attack on expert knowledge or empirical evidence. A range of motives may predict lying and uncivility (i.e., to polarize or get attention), and these motives are typically associated with the politics of disinformation (Bennett & Livingston, 2018; Freelon & Wells, 2020). Since both incivility and untruthfulness may reinforce each other and are potentially confounded by the desire to grab attention, it is essential to start with exploring their conceptual link.

Different forms of incivility and hate speech

Online incivility can roughly be defined as any type of offensive statement that trespasses the ideal type of democratic communication (e.g., Waisbord, 2018) and deliberation (Anderson et al., 2014; Papacharissi, 2004). Online incivility is thus an umbrella term of which hate speech forms an important component. Incivility can be contrasted to civility – which refers to the extent to which discussion partners as seen as equals with legitimate opinions, also in the setting of disagreement (e.g., Habermas, 2006). Incivility, however, implies that discussion partners are not treated with respect, and can, among other things, take on the shape of name-calling, profanity, negative stereotyping, interpersonal disrespect, and (digital) shouting (Chen & Lu, 2017; Coe et al., 2014). Overall, online incivility includes the acts of online rudeness (Jamieson, 1997) and making outrageous claims at different actors or (partisan) groups (e.g., Papacharissi, 2004).

It can be argued that the aims and consequences of spreading false information and incivility align. Both the targeted spread of false information and the expression of uncivil speech may involve (des)identification processes in which groups are attacked and ascribed negative traits in a stereotypical way. Exposure to incivility, for example, is found to polarize issue-publics (Anderson et al., 2014) and trigger incivility in polarized discussions (Gervais, 2014). When agents of disinformation aim to polarize or enhance negative sentiments toward certain out-groups (Bennett & Livingston, 2018), incivility may be regarded as an important rhetorical divide to fuel antagonisms.

Just like we conceptualize mis- and disinformation as a continuum of untruthfulness anchored by completely true to completely false information, we argue that incivility can take on many different shapes and forms – trespassing the ideals of deliberative democracy to different degrees. We simplify the continuum of incivility for the sake of our empirical endeavor. As a baseline or control for more severe types of hostility and incivility, we look at a negativity bias, which we operationalize as a disproportionate emphasis on the negative aspects of events or phenomena (Van der Meer et al., 2019). Even though negativity on its own does not trespass the boundaries of freedom of speech, it can have negative consequences by cultivating a disproportionate negative worldview among the audience, or fostering cynicism and distrust (Van der Meer et al., 2019). Yet, we should note that negativity is part of the current media logic, and a key characteristic of news coverage and political communication. Emphasizing the negative side of issues or using a negative tone to evaluate evens or actors, is not regarded as uncivil speech. However, it can offer a context for hate speech, and elements of negativity can crystalize into more uncivil forms of speech.

A prominent type of incivility may consist of partisan attacks, which are especially relevant to consider in the setting of bipartisan U.S. politics. Partisan attacks can be understood as depicting the opposed party in hostile ways, for example by attributing (stereotypical) negative qualities to this party, blaming them for political failures, or any form of derogatory language associated with the opposed party (Gross & Johnson, 2016). Next to partisan attacks, we look at attacks on the media, a communication tactic that has mostly been associated with (right-wing) populist parties that blame the legacy media for being dishonest, or spreading Fake News (e.g., Egelhofer & Lecheler, 2019; Farhall et al., 2019). Again, such expressions may be harmful for deliberative democracy: Delegitimizing established knowledge and empirical facts may undermine trust in legacy journalism and decrease the public's common understanding of the same factual truths (e.g., Arendt, 1967; Van Aelst et al., 2017).

We regard hate speech as the most severe type of online incivility. Hate speech is banned in some countries, although countries as the US do not have a legal framework to regulate hate speech. However, in many European countries, such as the Netherlands and Belgium, there have been many hate speech prosecutions (Vrielink, 2016). In line with emerging consensus, we define hate speech as any statement that expresses an attack, abuse, intimidation and/or denigration of individuals and groups that are defined on the basis of an out-group they are said to be part of (e.g., Van Spanje & De Vreese, 2014; Walker, 1994; Warner & Hirschberg, 2012). Such abusive and hateful speech can be targeted at different individuals, who are, for example, defined on the basis of ethnicity, religion, gender or nationality (e.g., Walker, 1994). Hate speech can contain threatening language or explicitly incite or legitimize violence, but only in extreme cases (e.g., Davidson et al., 2017; Walker, 1994).

As first aim of this paper, we explore the relative salience of different forms of negativity, online incivility, media attacks, and hate speech in different forms of communicative untruthfulness. As there is little empirical evidence on the affinity between hate speech and online incivility on the one hand and different degrees of disinformation on the other hand, the first research question is explorative and aims to map the nature and salience of online incivility: In what ways are different degrees of incivility associated with different levels of untrue and false communication? (RQ1).

As the democratic ideals of facticity, rationality, balance and deliberation are typically circumvented in online incivility and hate speech (e.g., Anderson et al., 2014; Papacharissi, 2004; Waisbord, 2018), false and untrue information may be more likely to contain uncivil language and hate speech than truthful communication. Therefore, and in line with extant research that has postulated that the politics of disinformation might align with the expression of uncivil sentiments (e.g., Bennett & Livingston, 2018; Marwick & Lewis, 2017), we hypothesize that incivility is more likely to be expressed in untrue and false information than in truthful communication (H1).

We expect that types of false information that deviate more from facticity and honesty, have a higher likelihood to contain hate speech and online incivility. Hence, 'mostly false' news stories that contain some inaccurate claims (i.e., because a lack of information, or disagreement between expert sources) may be less prone to negatively depict out-groups and contain hostile sentiments than completely false disinformation – based on fabrication or manipulation. The most 'severe' category of misinformation distinguished in this paper has mostly been associated with (radical) right-wing issue positions and negative out-group depictions (Bennett & Livingston, 2018; Marwick & Lewis, 2017). In line with these theoretical expectations, the following hypothesis is postulated: Types of misinformation that deviate more from facticity have a higher likelihood to contain incivility than more truthful information (H2).

Some actors may be more likely to express hate speech and communicate in an untruthful manner than others. Radical right-wing or populist politicians are in particular associated with expressing hateful statements targeted at ethnic and religious minorities and immigrants (Van Spanje & De Vreese, 2014). At the same time, these actors have been associated with the communication of manipulative, dishonest and untrue communication (Marwick & Lewis, 2017; Waisbord, 2018). Based on the alleged link between the communication tactics of (radical) right-wing populist actors, hate speech and disinformation, we postulate the following hypothesis: Radical right-wing populist actors are more likely to express different degrees of incivility in untrue and false information than other actors (H3).

Method

Rather than classifying content as false information ourselves, we rely on the thorough, in-depth verification efforts of independent fact-checking platforms. We specifically use the databases of two US fact-checkers that are comparable in their approach and classification scheme: Politifact.org and Snopes.com. Both platforms are not affiliated with a specific partisan leaning (at least not explicitly so) and both distinguish between different levels of (un)truthfulness – ranging from completely false to completely or mostly true. We validate our analysis with an additional inductive qualitative analysis of (non-fact-checked) true statements and a sample in which the intention to deceive was explicated by fact-checkers.

Sample

Our sample frame was not restricted to specific (media) outlets, actors, topics or periods, although we restricted our content analysis to political statements. Our sampling procedure contained multiple stages. First of all, we structured our sample frame by five different degrees of untruthfulness used by the fact-checking platforms: (1) (completely) true; (2) partially/mostly true; (3) partially/mostly false; (4) (completely) false and (5) pants on Fire! (the final category only occurred in PolitiFact – which was merged with the fourth category). For every category, we randomly sampled 100 statements from each platform. The only inclusion criterion for the stories was that they had to deal with a political topic – which means that stories on food safety regulations and celebrity news, for example, were excluded. Such stories only contain a very small part of all the content verified by fact-checkers.

All verified stories were coded on the statement level. This means that we tracked or copied the original statement (e.g., Tweet by a politician) verified in the fact-checker's verdict and coded the content of the original statement or speech on uncivil language and hate speech.

Key variables: online incivility and hate speech

Our main variables aimed to map the presence of different components of online incivility – and hate speech more specifically. We discerned four main types of incivility: hostility and negativity in general, negative or hateful sentiments targeted at partisans or political opponents (partisan attacks), attacks on the (legacy) media, and hateful or inflammatory speech targeted at minority groups (hate speech). Coders coded the full statement that was verified by the fact-checker. The length of these statements differed. Sometimes, the fact-checker simply checked the veracity of one specific utterance of a politician (i.e., Are the Democrats indeed attacking Obamacare?). In other instances, a full Tweet, Facebook post, or news article was checked. If there were different statements checked by the fact-checker, for example as the news article made different claims, coders were instructed to complete the coding sheet for each individual statement. We will explicate the operationalization of the variables below (also see Appendix A for the coding sheet).

Negativity

With a single item, coders had to assess the overall negativity of the statement. Negativity, for example, revolves around the framing of issues as political failure, fiasco, disaster, crisis, frustration, collapse, flop, denial, rejection, neglect, default, deterioration, resignation, skepticism, threats, cynicism, defeatism or disappointment. Coders had to assess whether the statement was mostly negative, positive, balanced (equal negative and positive indications) or if the statement was neutral.

Partisan attacks

Next, coders assessed whether the statement contained a partisan attack on either the Republican or Democrat party, or members of these parties. In the coding instructions, it was explicated that such attacks go beyond negativity, and include a blame-attribution to the opposed parties (i.e., blaming the other party or opposed politicians for voter fraud, attributing negative qualities to out-party members). There were three categories: partisan attacks were absent, targeted at Democrats or party members or specific politicians with Liberal affiliation, or targeted at Republicans or party members or specific politicians with Conservative affiliation.

Media critique

This item was coded as a dichotomy (absent/present), and specifically asked coders to indicate whether the statement include attacks on the media, media criticism or any negative reference to the functioning of the press, journalists or the media. This could both refer to intentional and incidental inaccurate reporting.

Hate speech

This key variable was coded in three consecutive steps. First of all, coders had to assess whether the statement contained any negative statements connected to an out-group/minority group or a political actor/ideological group that is framed in a negative stereotypical way (i.e., based on gender, ethnicity, religion, race, sexual orientation, nationality, ideology or other group-level characteristics). If the answer to this question was ‘yes’, coders had to identify the specific out-group framed in a negative way, and code for the specific attribution made to this out-group (i.e., an indication of blame attribution, a statement on out-group inferiority, denying in-group membership). In the analyses, we regard the combination of the presence of negative statements that depicted an out-group in a negative way; derogatory language used to blame this out-group, frame it as inferior, or denying in-group membership as indicative of hate speech.

Contextual variables

Next to these main variables, we coded for the source of the statement, the outlet in which the original statement appeared, the topic of the statement, and the type of verification.

Inter-coder reliability and validity checks

The coding was performed by two native-speaking coders. After intensive coder training, which consisted of familiarization with the codebook, example coding, and refining the codebook with more detailed instructions based on the coders' feedback, a first round of inter-coder reliability testing was done (see Table 1 for an overview of reliability indices for the main variables). Both coders independently coded 45 statements (five statements for each category of (un)truthfulness distinguished). Even though most variables were coded with a sufficient reliability (Krippendorff's alpha > .63, agreement > 82%), all differences between coders were discussed until complete agreement was reached. Based on the discrepancies between coders, the codebook was revised by adding more detailed decision rules. After yet another round of coder training, the second inter-coder reliability sample of 45 statements was independently coded by the two coders. After this second round, the indices improved (see Table 1). Across the full ICR-sample, the following scores were achieved: Krippendorff's alpha > .68, agreement > 84%.

Analyses

To test our hypotheses, we ran logistic regression models in which the types of untruthfulness (the reference category were verified, truthful statements) were inserted as independent variables alongside contextual-level control variables. The different degrees of incivility were included as dependent variables (absent versus present). With these models, we estimated the extent to which different degrees of untruthfulness correspond to incivility on different levels – ranging from overall negativity to hate speech utterances. In our analyses, we control for the fact-checking platform, the topic of the statement, the type of refutation/verification made, the source of the (un)truthfulness, and the (media) platform on which the original statement was published.

Results

The salience of negativity, partisan attacks and hate speech in disinformation

Before testing the hypotheses, we present some descriptive statistics on the salience of misinformation and incivility. Table 2 presents an overview of the different proportions of negative, partisan attacks and hate-speech across the five types of (un)truthfulness distinguished in this paper. First of all, it can be observed that the majority of content verified by fact-checkers has a mostly negative tone. This negativity bias is most pronounced in content flagged as untrue but occurs across all categories. The proportions further demonstrate that the higher the deviation from facticity, the higher the proportion of partisan attacks on Democrats. For

Table 1. Results of two rounds of inter-coder reliability tests.

Variable	Source media speaker	Source statement	Source message	Topic	Rating	Modality	Type verification	Media critique	Negativity	Partisan attack	Hate speech	Out-group	Hate attribution	Hate language
Round I														
KALPHA	.79	.74	.63	.63	.89	.74	.75	1.000	.65	.81	.78	.75	.64	.75
% Agreement	98	84	82	76	91	89	84	100	80	93	93	89	93	87
Round II														
KALPHA	.78	.76	.68	.75	.89	.76	.78	1.000	.72	.78	.76	.76	.68	.76
%Agreement	98	86	84	85	91	89	86	100	85	92	92	90	95	90

Note: All inter-coder reliability tests were conducted on sample frames that are not part of the main study (both are from the same fact-checking platforms). The two rounds were conducted before coding the full sample frame, and there was a delay of about a month between the two rounds. KALPHA – Krippendorff's alpha.

Table 2. Negativity, incivility and hate speech in different types of (un)truthfulness.

	True	Mostly true	Mostly false	False	Pants on Fire!	Overall
Negative tone	52.2%	43.2%	62.4%	64.0%	57.8%	55.8%
Partisan attacks on Democrats	7.6%	9.0%	18.3%	26.9%	30.4%	17.1%
Partisan attack on Republicans	12.6%	7.5%	10.2%	9.1%	4.9%	9.3%
Media attacks	1.5%	3.0%	3.0%	4.1%	4.9%	3.1%
Hate speech	20.7%	14.6%	24.4%	32.0%	33.3%	24.0%
N	198	199	197	197	102	894

Note: Cell entries represent proportions of all fact-checked claims within the respective category of (un)truthfulness.

partisan attacks on Republicans, however, the pattern is reversed: The more information deviates from facticity, the *lower* the proportion of partisan attacks on Republicans. Compared to the other forms of uncivil speech distinguished, attacks on the media are a relatively rare event, and are most likely to occur in false information. Hate speech – the attribution of negative qualities to out-groups – occurs relatively frequently in all distinguished categories of (un)truthfulness. Overall, to answer RQ_1, negativity, partisan attacks and hate speech are relatively salient phenomena in the statements verified by the two fact-checking platforms.

Incivility across different types of verified (Un)truthfulness

As shown in Table 3, not all types of false information have a higher likelihood to contain uncivil speech compared to information verified as true. Hence, only completely false information has a significantly higher likelihood to contain negative speech compared to real information (marginally significant for partially false information).[1] The highest predicted probability for negativity was, for example, .75 when the statement was complete false, came from a politician about the topic domestic party politics and was verified on content and source level by platform Snopes. Attacks on Democrats are more likely to occur in mostly and completely false information, but the likelihood is not higher for partially false/mostly compared to true information. As an illustration, when the claim (about the function of the press) was completely false, came from a right-wing politician and was verified by PolitiFact, the predicted probability of attack on Democrats was .59. Hate speech is most likely to occur in completely false information, but the less severe deviations from facticity do not significantly contain more hate speech than verified information. For example, with a completely false claim from a political actor about domestic party politics, verified on content by Snopes, the predicted probability was the highest with .61 for hate speech. These findings offer partial support for H1: Information flagged as false has a higher likelihood to contain negativity, partisan attacks and hate speech than information rated as true. However, the less severe types of mis- and disinformation do not have a higher likelihood to contain uncivil speech, and partisan attacks are only more salient in disinformation when the Democrats are attacked.

The second hypothesis postulated the expectation that the relationships between false information and uncivil speech are most pronounced in the types of false information that contain the strongest deviations of truthfulness. The estimates depicted in Table 3 offer support for this expectation. More specifically, the associations between untruthful communication and negativity and hate speech are only significant for information that was rated as completely false. In addition, only mostly and completely false information

Table 3. Logistic regression models predicting likelihood of negative and uncivil language in (un)truthful communication.

	Negativity		Attack on Democrats		Attack on Republicans		Media critique		Hate speech	
	B (SE)	*OR* & 95% *CI*	B (SE)	*OR* & 95% *CI*	B (SE)	*OR* & 95% *CI*	B (SE)	*OR* & 95% *CI*	B (SE)	*OR* & 95% *CI*
(constant)	−.81 (.43)*	.07 [−1.54, .26]	−3.26 (−.11)*	.04 [−4.61, −2.26]	−4.75 (1.97)*	.01 [−6.82, −3.05]	−4.80 (2.98)	.01 [−18.79, −2.92]	−3.53 (.59)*	.03 [−4.67, −2.52]
FC (Snopes)	.29 (.15)	1.34 [−.04, .57]	−.26 (.26)	.77 [−.75, .26]	.07 (.30)	1.07 [−.47, .67]	−.37 (.51)	.69 [−1.43, .74]	.48 (.21)*	1.61 [.08, 1.01]
Topic	.01 (.01)	1.01 [−.02, .03]	.08 (.00)**	1.09 [.04, .13]	.05 (.02)*	1.06[.01, .11]	.16 (.12)	1.07 [.03, .60]	.06 (.01)**	1.06 [.04, .10]
Media source	−.01 (.00)	.99 [−.03, .02]	.01 (.02)	1.01 [−.02, .04]	.04 (.02)	1.04 [.01, .08]	−.00 (.03)	.99 [−.06, 06]	.04 (.01)*	1.04 [.02, .07]
RWP source	−.34 (.29)	.68 [−1.07, .20]	.62 (.32)†	1.87 [−.07, 1.29]	−.42 (2.09)	.66 [−8.77, 1.45]	.06 (.55)	1.07 [−1.05, 1.10]	−.07 (.35)	.94 [−.68, .71]
Mainstream politicians	.13 (.21)	1.14 [−.35, .47]	−.27 (.28)	.76 [−.82, .31]	1.58 (.30)**	4.87 [.64, 3.02]	−17.96 (.43)**	.01 [−18.62, −16.89]	.16 (.05)	1.18 [−.51, .76]
Journalist source	.11 (−.01)	1.11 [−.50, .85]	.47 (.38)	1.60 [−.28, 1.33]	.55 (3.10)	1.74 [−7.96, 1.80]	−.39 (4.24)	.68 [−18.22, 1.06]	.47 (.35)	1.61 [−.14, 1.24]
Citizen source	−.03 (.35)	.97 [.65, 1.46]	−.02 (.28)	.98 [−.54, .51]	.86 (.32) †	2.36 [−.03, 2.31]	−.13 (.52)	.88 [−1.40, .93]	−.22 (.28)	.80 [−.72, .56]
Verification type	.30 (.14)	1.35 [1.04, 1.76]	.01 (.01)	1.01 [−.37, .36]	.37 (.23)	1.45 [−.16, .81]	−.22 (.38)	.80 [−.96, .63]	.11 (.16)	1.12 [−.28, .43]
Mostly true	−.38 (.20)†	.68 [−.72, 0.01]	.32 (.37)	1.38 [−.53, 1.05]	−.46 (.38)	.63 [−1.31. .11]	.82 (2.99)	2.28 [−.62, 8.94]	−.34 (.31)	.71 [−1.02, .34]
Mostly false	.37 (.21)†	1.45 [−.07, .82]	.90 (.38)**	2.45 [.27, 1.69]	−.13 (.01)	.88 [−.78, .46]	.47 (2.41)	1.60 [−.86, 8.57]	.23 (.27)	1.26 [−.36, .72]
Completely false	.42 (.21)*	1.51 [.06, .83]	1.30 (.29)**	3.68[.76, 1.99]	−.45 (.32)	.64 [−1.09, .22]	.26 (2.46)	1.30 [−1.09, 8.39]	.72 (.27)**	2.05 [.15, 1.29]
Nagelkerke R^2	.053		.154		.112		.179		.112	
χ^2 (11)	35.97***		86.53***		47.44***		39.82***		69.62***	

SE = standard error; OR: odds ratio; df = degree of freedom. Two-tailed tests. bootstrapping was performed. Reference category of false information is true information. Unstandardized regression weights. SEs reported between parentheses. $^{\dagger}p < .010$, $^{*}p < 0.05$; $^{**}p < 0.01$; and $^{***}p < 0.001$. $N = 894$.

increases the likelihood that statements contain partisan attacks on the Democrats. Partially false information is thus not significantly associated with uncivil language, whereas more severe deviations from truthfulness are (the difference between partially false and completely false information is significant as well). This offers support for H2. The stronger the deviation from factual reality, the more likely negative speech, partisan attacks, and hate speech occur in verified statements.

Based on these findings, we find support for the main expectation for a discursive affinity between mis- and disinformation and negative, uncivil utterances. Even though we cannot completely rule out a selection effect of fact-checkers, we find nuanced differences between different types of false information – which does indicate that fact-free coverage is more likely to contain hate speech than truthful information.

Right-wing populists and uncivil speech in (Un)truthful communication

We expected that negativity, partisan attacks and hate speech in false information are most likely to come from the radical right or right-wing populist politicians (H3). To test this hypothesis, we ran binary logistic regression models in which the effects of a two-way interaction between right-wing populist sources and different degrees of untruthful information on negativity were estimated. The results first of all point to a non-significant negative interaction-effect between RWP sources and mostly false information on negativity ($B = -.05$, $SE = .54$, $p = .920$, 95% CI OR [.33, 2.71]). The same pattern was found for completely false information ($B = -.24$, $SE = .49$, $p = .632$, 95% CI OR [.30, 2.07]). Hence, contrary to the expectations postulated under H3, a negativity bias is not more likely to occur in false information disseminated by RWP politicians compared to other actors.

We see that partisan attacks mostly come from opposed political parties. As can be seen in Table 3, mainstream or Liberal political actors are most likely to attack the Republican party, and RWP actors (including Trump) are more likely than other actors to express negative sentiments to the Democrat party. However, the interaction effect between RWP sources and false information was non-significant for partisan attacks on the Democrats (mostly false: $B = -.05$, $SE = .68$, $p = .947$, 95% CI OR [.26, 3.59], completely false: $B = -.29$, $SE = .61$, $p = .628$, 95% CI [.23, 2.44]). This indicates that the partisan attacks of RWP actors are not more likely to occur in false compared to true information.

However, we do find a significant and positive two-way interaction effect between the presence of a RWP source and mostly false and completely false information on partisan attacks on the Republican party (Mostly false: $B = 18.87$, $SE = 6.77$, $p = .010$, 95% CI [−.20, 20.33], completely false: $B = 17.94$, $SE = 8.91$, $p = .010$, 95% CI [−.32, 19.54]). This means that RWP actors (compared to other sources) are more likely to attack the Republican Party in the context of false compared to true information. Although this pattern is in line with H3, it is surprising that this two-way interaction effect is positive and significant for partisan attacks on the Republican party, but not for attacks on the opposed Democrats – who are attacked across different degrees of false and true information in similar ways.

To assess whether this association is driven by the presence of a RWP source and not simply a reflection of partisan cleavages, we additionally estimated the interaction effects between mainstream political sources (most likely from the Democrat party) and false information on the occurrence of partisan attacks. Here, we find a significant negative two-way interaction effect between mostly false information and mainstream politicians on partisan attacks on the Republican Party ($B = -1.55$, $SE = .64$, $p = .016$, 95% CI OR [.06, .75]). This means that mainstream political actors are *less* likely to attack the opposed political party in partially false compared to true information. The interaction effect was non-significant for completely false information ($B = -.69$, $SE = .61$, $p = .264$, 95% CI OR [.15, 1.68]). We also found non-significant two-way interaction effects between mainstream politicians and false information on partisan attacks on the Democrats, although there is a marginally significant and positive two-way interaction effect for mostly false information ($B = 1.02$, $SE = .60$, $p = .089$, 95% CI OR [.89, 9.08]) – which does indicate that mainstream political actors are slightly more likely to attack their own party or politicians from their own party in partially false compared to completely true information.

Looking at the other indicators of uncivil speech – media critique and hate speech – we find no significant interaction effects between the presence of a RWP politician versus other actors and different degrees of false information on media critique and hate speech– which does not offer support for H3.

In sum, we found very limited support for H3: Although RWP actors are slightly more likely to attack Democrats, and mainstream politicians more likely to attack the Republican Party, these partisan attacks are not more likely to occur in information flagged as false, but rather point to a more general pattern of partisan divides in media coverage.

Exploratory qualitative assessment of the link between incivility and disinformation

To further explore the validity of the association between incivility and disinformation and its distinction to real information, we conducted a qualitative content analysis of false articles in which the intention to mislead was not identified ($n = 50$) and fact-checked articles in which the intention to deceive was explicated ($n = 50$) – manually coded by looking at the verification efforts and the political agenda underlying the dissemination of falsehoods (50 statements from each platform). We compared this to a sample of real articles that were not fact-checked ($n = 50$, randomly selected across news sources via LexisNexis). These real articles were matched on sources, topics and publication dates of the fact-checked statements. The content was coded on the statement level, and selective coding according to the principles of discourse analyses was conducted.

The overlap between the three categories of (mis)information is the centrality of negativity connected to partisan discussions. Yet, negativity was only explicitly connected to partisan attacks in false and intentionally deceptive content. In false statements, for example, political candidates were attributed blame for not taking care of the people: ‘Joe Biden tried to cut Social Security and Medicare for decades. Now Biden’s promising your benefits to illegal immigrants.’ These attributions could, however, not be identified as hate speech. Although we see some indicators of incivility (12%), swearing and out-group hostility were absent in this category. Hostility was more central (38%) in the content that can be classified as intentionally deceptive. To give an example, statements in this category spoke about ‘illegal aliens’ when referring to illegal immigration, associated political figures with Hitler, and wrongfully alleged Trump of making strong racist claims: ‘Africans Are Lazy, Good at Sex, Theft.’

Here, it should be noted that intentionally false information did not always contain hate speech or incivility directly, but also accused (opposed) politicians of expressing such uncivil comments – which can be an important disinformation tactic. In addition, the degree of falsity did not necessarily coincide with the severity of intentional deception. Specifically, partially false content was oftentimes based on the ‘cherry picking’ of some facts to make them reflect a political agenda or statement that delegitimized political opponents. The main conclusion of the qualitative validation, then, is that although negativity may be central across all types of information, misinformation contains more partisan attacks and blame attributions, whereas disinformation is indeed most likely to contain incivility and hate speech. We also see that the real (truthful) articles that were not verified by fact-checkers correspond strongly to the articles found to be true after verification: Hate speech and incivility were absent in both categories of real information, whereas negativity was present in both (although to a lesser extent than the false articles).

Discussion

It has been argued that today’s fragmented and digitized information settings create an opportunity structure for the spread of mis- and disinformation and hostile speech (e.g., Bennett & Livingston, 2018; Freelon & Wells, 2020; Van Aelst et al., 2017). In this study, we put the alleged association between uncivil speech and false information to an empirical test. Specifically, we rely on a content analysis of statements flagged as different degrees of (un)truthfulness by two independent fact-checking platforms in the US ($N = 894$), and coded for the presence of partisan attacks, media critique and hate speech across five categories of (un)truthfulness.

Our main findings offer support for the theoretically identified association between false information and incivility, especially when it comes to negativity and hate speech. Whereas negativity and hate speech can be found in all types of flagged content, it is most likely to occur when information is found to be completely false. These findings offer empirical evidence for a thesis that to date has mostly been based on theoretical arguments (Bennett & Livingston, 2018; Marwick & Lewis, 2017): Radical right-wing issue positions that negatively portray out-groups occur most in the categories of untruthfulness that deviate strongest from reality.

These findings have important theoretical implications. Although the distinction between ‘honest’ mistakes (misinformation) and deliberate deception (disinformation) has been regarded as central to the debate on misinformation (Karlova & Fisher, 2013), we know little about how we can distinguish both types of false information based on content features: How can we operationalize intentional deception when actors of disinformation try to format false or doctored information to match legacy journalism? The qualitative and qualitative findings of this paper suggest that hostility, hate speech and partisan attacks may be important content indicators to distinguish between mis- and disinformation. We hope that these findings offer a starting point for future research to more precisely and validly identify mis- and disinformation.

An implication of our findings is that when people expose themselves to false statements – for example by approaching alternative media platforms – they are also more likely to be exposed to hateful content and uncivil, negative speech, which points to the real-life democratic implications of disinformation. Guess et al. (2018) found that, in 2016, 25% of US citizens visited websites containing false information, and that these pages were mostly biased toward conservative views. This is also reflected in our analyses: We see that hostile attacks on the Democrats are more likely to occur in false compared to true information, whereas the degree of untruthfulness did not correspond to more hostile attacks on the Republican party. This could partially be due to the salience of Trump's false statements in our sample: He is known to refer to Republicans in hostile ways – often using delegitimizing labels that are not based on empirical knowledge (e.g., Farhall et al., 2019). This further supports the link between radical right-wing (populist) sentiments and disinformation (Bennett & Livingston, 2018). The presence of a right-wing populist leader in the US may partially explain why the link between partisan attacks on the Democrats was identified, whereas we did not see this for Republicans.

As exposure to online disinformation, online hostility and partisan cues may reinforce affective polarized divides (e.g., Bennett & Livingston, 2018), exposure to the most severe types of false information may cultivate incivility and hateful sentiments vis-à-vis different groups in society. Yet, we have to acknowledge the potential endogeneity of our findings. Although it reaches beyond the scope of this paper to prove causality, it could be argued that incivility motivates false information: If the aim is to attack opposed partisans or negatively stereotype out-groups, actors have to resort to lying as empirical evidence cannot support their claims. Alternatively, the aim of disinformation can be regarded as disrupting societal order by manipulating reality – expressing uncivil speech may be an important tool to fulfill this goal.

It should be noted that fact-checking platforms may be biased to include 'suspicious' claims – even if they are found to be true. Hence, fact-checkers are not likely to select statements that are clearly true and contain a bias in fact-checking statements that have a higher likelihood of being false than non-checked statements. The difference between categories would be stronger if we included more (baseline) true articles that were not suspicious and less likely to contain incivility. Future research may devote more attention to the selection biases of fact-checking platforms. As practical implication, it may be important that fact-checkers become more transparent about their inclusion criteria – emphasizing that they do not apply partisan lenses when flagging content as untrue.

Despite offering novel insights into the discursive link between disinformation and incivility and hate speech, this study has some limitations. First of all, we only included the statements of two fact-checkers in the context of the US. It remains to be tested how well our findings travel to the statements verified and flagged by other platforms in different countries – or statements that are not fact-checked and therefore less prone to false claims. Second, we base the classification of different degrees of false information on the classification scheme of fact-checkers – which includes a selection bias and verification bias that we cannot account for. Even though (computational) approaches that classify false information bottom-up still rely on verification and come with important validity issues, future research may arrive at a more comprehensive overview of the connection between civility and disinformation by classifying both phenomena inductively – and on larger corpora of statements. The intention behind the dissemination of false information is hard to identify based on content features of communication (which may hint at the intentions of the sender, but not directly include them). Hence, although our qualitative analyses looked at false statements in which a political agenda and motive for deception was revealed, future research should further explore how intentional deception can be separated from false content in general. Yet, we believe that the affinity between incivility and false information is an important starting point to reveal the hidden intentions of the communicator.

In addition, we did not explore the causal order of false information and incivility – future research may look into the specific motivations underlying incivility and the dissemination of false information to see how the underlying motivations align or diverge. Finally, we only looked at a sub-set of incivility indicators and may have overlooked some elements of online incivility in this paper. Future research may rely on a more extensive set of indicators of incivility to reveal nuances between the degrees of untruthfulness and uncivil speech.

Despite these limitations, we offer empirical evidence that confirms the alleged affinity between disinformation and incivility – highlighting the potentially harmful democratic implications of the uncontrolled spread of false information in today's digital media ecologies. To combat disinformation, it may not only be important to correct false information and stimulate critical media skills, but also to demonstrate the neutrality and impartiality of verification endeavors.

Note

1. We decided to report logistic instead of ordinal regression models because different degrees of falsehoods may be regarded as different types of untrue or false information instead of a one-dimensional ordering of falseness. Yet, ordinal regression models yield similar results as the patterns reported here.

Acknowledgements

This project was supported by a gift from Facebook's Content Policy Research Initiative. We are indebted to our coders and research assistants Nilou Yekta and Nirvi Maru. Next to their meticulous coding work, their thoughtful comments during the developmental stages of the codebook and validity checks have greatly improved the quality of our results.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This project was supported by a gift from Facebook's Content Policy Research Initiative.

Notes on contributors

Dr. Michael Hameleers (PhD, University of Amsterdam) is an Assistant Professor of Political Communication at the Amsterdam School of Communication Research (ASCoR). His research interests include (right-wing) populism, disinformation, and selective exposure [email: m.hameleers@uva.nl].

Dr. Toni van der Meer is an assistant professor at the Department of Corporate Communication of ASCoR, University of Amsterdam. His dissertation focused on the communicative interplay between the organization, news media, and the public in times of crisis. After his PhD, Toni has continued his line of research in crisis communication and published on negativity (bias) in the media, selective exposure and misinformation [email: G.L.A.vandermeer@uva.nl].

Rens Vliegenthart is a Professor in Communication Science (chair in Media and Organizations) in the department of Communication Science and at the Amsterdam School of Communications Research (ASCoR), University of Amsterdam. His research interests media-politics relations, media coverage of social movements and businesses, election campaigns, European integration, soccer hooliganism and time series analysis. In 2007, he completed his Ph.D. on the immigration debate in the Netherlands at the Vrije Universiteit Amsterdam (cum laude). In 2005–2006 he was a Fulbright visiting scholar in the department of Sociology, University of California, Irvine and in 2009–2010 a visiting associate professor in the department of Political Science and Public Management, University of Southern Denmark [email: r.vliegenthart@uva.nl].

ORCID

Michael Hameleers http://orcid.org/0000-0002-8038-5005

References

Anderson, A. A., Brossard, D., Scheufele, D. A., Xenos, M. A., & Ladwig, P. (2014). The "nasty effect:" online incivility and risk perceptions of emerging technologies. *Journal of Computer-Mediated Communication*, *19*(3), 373–387. https://doi.org/10.1111/jcc4.12009

Arendt, H. (1967, February 25). Truth and Politics. *The New Yorker.*

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

Chen, G. M., & Lu, S. (2017). Online political discourse: Exploring differences in effects of civil and uncivil disagreement in news website comments. *Journal of Broadcasting & Electronic Media*, *61* (1), 108–125. https://doi.org/10.1080/08838151.2016.1273922

Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, *64*(4), 658–679. https://doi.org/10.1111/jcom.12104

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

Farhall, K., Carson, A., Wright, S., Gibbons, A., & Lukamto, W. (2019). Political elites' use of fake news discourse across communications platforms. *International Journal of Communication*, *13*, 4353–4375. https://doi.org/1932–8036/20190005

Freelon, D., & Wells, C. (2020). Disinformation as political communication. *Political Communication*, *37*(2), 145–156. https://doi.org/10.1080/10584609.2020.1723755

Gervais, B. T. (2014). Following the news? Reception of uncivil partisan media and the use of incivility in political expression. *Political Communication*, *31*(4), 564–583. https://doi.org/10.1080/10584609.2013.852640

Gross, J. H., & Johnson, K. T. (2016). Twitter taunts and tirades: Negative campaigning in the age of Trump. *PS: Political Science & Politics*, *49*(4), 748–754. https://doi.org/10.1017/S1049096516001700

Guess, A., Nyhan, B., & Reifler, J. (2018). *Selective Exposure to disinformation: Evidence from the consumption of fake news during the 2016 US presidential campaign.* https://www.dartmouth.edu/~nyhan/fake-news-2016.pdfGoogle Scholar

Habermas, J. (2006). Political communication in media society: Does democracy still enjoy an epistemic dimension? The impact of normative theory on empirical research. *Communication Theory*, *16*(4), 411–426. https://doi.org/10.1111/j.1468-2885.2006.00280.x

Humprecht, E. (2018). Where 'fake news' flourishes: A comparison across four Western democracies. *Information, Communication & Society*, *22*(13), 1973-1988. https://doi.org/10.1080/1369118X.2018.1474241

Jamieson, K. (1997). *Civility in the House of Representatives*. APPC report 10. Retrieved March 28, 2011, from http://democrats.rules.house.gov/archives/hear01.html

Karlova, N. A., & Fisher, K. E. (2013). A social diffusion model of misinformation and disinformation for understanding human information behaviour. *Information Research*, *18*(1), paper 573.

Katz, D., & Braly, K. (1933). Racial stereotypes of one hundred college students. *The Journal of Abnormal and Social Psychology*, *28*(3), 280–290. https://doi.org/10.1037/h0074049

Lowry, P. B., Zhang, J., Wang, C., & Siponen, M. (2016). Why do adults engage in cyberbullying on social media? An integration of online disinhibition and deindividuation effects with the social structure and social learning model. *Information Systems Research*, *27*(4), 962–986. https://doi.org/10.1287/isre.2016.0671

Marwick, A., & Lewis, R. (2017). *Media manipulation and disinformation online*. Data and Society Research Institute. https://datasociety.net/output/media-manipulation-and-disinfo-online/

Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society*, *6*(2), 259–283. https://doi.org/10.1177/1461444804041444

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50. https://doi.org/10.1016/j.cognition.2018.06.011

Tandoc, E. C., Lim, Z. W., & Ling, R. (2018). Defining "fake news": A typology of scholarly definitions. *Digital Journalism*, *6*(2), 137–153. https://doi.org/10.1080/21670811.2017.1360143

Van Aelst, P., Strömbäck, J., Aalberg, T., Esser, F., de Vreese, C. H., 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

Van der Meer, T. G. L. A., Kroon, A. C., Verhoeven, P., & Jonkman, J. (2019). Mediatization and the disproportionate attention to negative news: The case of airplane crashes. *Journalism Studies*, *20*(6), 783–803. https://doi.org/10.1080/1461670X.2018.1423632

Van Spanje, J., & De Vreese, C. (2014). The way democracy works: The impact of hate speech prosecution of a politician on citizens' satisfaction with democratic performance. *International Journal of Public Opinion Research*, *26*(4), 501–516. https://doi.org/10.1093/ijpor/edt039

Vraga, E. K., & Bode, L. (2020). Defining misinformation and understanding its bounded nature: Using expertise and evidence for describing misinformation. *Political Communication*, *37*(1), 136–144. https://doi.org/10.1080/10584609.2020.1716500

Vrielink, J. (2016). Do we want more or fewer prosecutions of opinions: The Geert Wilders trial 2.0. *Netherlands Journal of Legal Philosophy*, *45*(2), 3–11. https://doi.org/10.5553/NJLP/.000053

Waisbord, S. (2018). The elective affinity between post-truth communication and populist politics. *Communication Research and Practice*, *4*(1), 17–34. https://doi.org/10.1080/22041451.2018.1428928

Walker, S. (1994). *Hate speech: The History of an American Controversy*. University of Nebraska Press.

Wardle, C. (2017). *Fake news. It's complicated*. First Draft. https://medium.com/1st-draft/fake-news-its-complicated

Warner, W., & Hirschberg, J. (2012). Detecting hate speech on the World Wide Web. In: *Workshop on Language in Social Media*. ACL, 19–26.