Listen to What They Say: Better Understand and Detect Online Misinformation with User Feedback

Hubert Etienne and Onur Çelebi

Abstract. Social media users who report content are key allies in managing online misinformation; however, no research has yet been conducted to understand the trends underlying their activity. We suggest an original approach to studying misinformation: examining it from the reporting users' perspective both at the content level and comparatively across regions and platforms. We propose the first classification of reported content, resulting from the review of 8,975 items reported on Facebook and Instagram in France, the UK, and the US. This allows us to observe meaningful distinctions regarding misinformation propagation between countries and platforms, as it significantly varies in volume, type, and topic. Our review completes existing typologies of manipulation techniques serving disinformation campaigns with a novel one, "the excuse of casualness," which presents a concrete challenge for algorithmic detection, thus confirming the need to better classify user reports as a key signal. We then identify four reporter profiles, from which we derive four reasons for inaccurate reporting that are capable of explaining 55% of the inaccuracy in misinformation reporting. Finally, we demonstrate that a simple classifier trained on a small dataset with a combination of basic reporting signals can identify these inaccuracy types, thus improving the quality of the reporting signal.

1 Introduction

Social media users who report content (hereafter abbreviated "reporters") are key allies in the management of online misinformation. By posting comments expressing disbelief and providing fact-checking materials, they constitute the first line of defense against the potential virality of a hoax on social media and reduce the impact of false news on people's beliefs. By reporting on false claims and encouraging others to do so, reporters also provide moderators with relevant signals that support misinformation detection. However, hundreds of millions of reports were submitted by Facebook and Instagram users in June 2020 under the "false news" category alone, many of which are not directly relevant to fact-checkers. Filtering user reports therefore constitutes a key challenge for the moderation of social media platforms, but the difficulty of

doi:10.54501/jots.v1i5.106

doing so has often led experts to consider them as complex signals that are too noisy to efficiently support algorithmic detection and prioritization of relevant content for fact-checking. This difficulty, in addition to the limited access to reporting databases, explains why reporters are still absent from the growing literature on misinformation, which instead focuses on the propagators of hoaxes. A second gap in the literature relates to the lack of web-data-based comparative analysis between countries and platforms, although such research is necessary to develop an in-depth understanding of misinformation. We suggest an original approach to fill these gaps: leveraging mixed methods to examine misinformation from the reporters' perspective, both at the content level and comparatively across regions and platforms.

This paper aims to demonstrate the relevance of such an approach for improving the understanding of misinformation on social media and developing better moderation methods. Its contribution to this objective is threefold. We start by presenting the first research methodology to classify user reports, resulting from the human review of a dataset of 8,975 content items reported on Facebook and Instagram in France, the UK, and the US in June 2020, and we give a detailed description of this methodology to enable its use in future studies.

We then leverage qualitative and quantitative methods to analyze this original dataset, and from the results we obtain two key findings. The qualitative review of our dataset allows us to confirm and complete previous categorizations of information manipulation techniques serving "disinformation"—the intentional spreading of false news, whereas "misinformation" remains neutral vis-à-vis the users' intentions—with a novel technique, which seems to have emerged on Instagram in the US. This new manipulation technique presents significantly more challenges for algorithmic detection and human moderation, therefore confirming the importance of better filtering of user reports to improve the accuracy of misinformation detection models. The quantitative analysis of the dataset allows us to observe meaningful distinctions between countries, thus tempering the discourse on a global "infodemic" (WHO 2020). For example, our analysis suggests that COVID-19 misinformation did not impact France as strongly as it did the US, in terms of volume of false news in circulation, and that the key moderation issues vary by country: while misinformation appears to be a key issue in the US for both Facebook and Instagram, it seems to be a minor issue for Instagram in France, which instead faces a spamming issue.

Finally, we use statistical and computational methods to classify user reports, on the basis of which we present two important results. First, our classification allows user reports to be distributed into four reporter profiles that can explain most of the inaccuracy (55%) in misinformation reporting. Breaking down the inaccuracy into four types of noise associated with these different profiles, we suggest specific means of actions that can increase the relevance of the signal. Second, we demonstrate the performance of a gradient-boosting classification model trained on a combination of user reports when identifying these types of noise.

2 A reporter-oriented approach to studying misinformation at the content level

In the absence of an existing classification for content reported by social media users, we developed an original one, resulting from the human review of 8,975 content items. This section explains how our approach aims to fill important gaps in the misinformation literature and describes how the dataset was collected, as well as how the annotation methodology was built.

2.1 General approach

Misinformation research benefits from a diversity of methodologies, selected countries, and social platforms. Such diversity, however, makes it difficult to compare results, generalize findings, and conduct meta-analyses. Furthermore, with a few exceptions (e.g., Humprecht, Esser, and Van Aelst 2020; Cinelli et al. 2021), the selection of several countries and platforms in web-data-based research is justified for data augmentation purposes, rather than for comparative analysis. The resulting taxonomies of misinformation content (Wardle 2016; Molina et al. 2021; Innes 2020) thus do not provide information on specific platforms and countries, whereas the annual Reuters Institute Digital News Report (Newman et al. 2020) suggests that misinformation is significantly sensitive to these variables. Our comparative approach aims to contribute to the literature by differentiating misinformation manifestations and practices across platforms and regions for which existing research is abundant (Facebook, especially in the US and the UK) and those for which it is minimal or non-existent (Instagram, especially in France). While the proportion of people using social networks to access the news has been relatively stable across Facebook, Twitter, and YouTube (+0%, +4.9% and +4.6% CAGR, respectively), it increased five-and-a-half-fold for Instagram between 2014 and 2020 (Newman et al. 2020, 29). We thus expect Instagram to provide a better observation point for obtaining new insights on misinformation types and practices.

The high volume of misinformation content and the limited access to relevant datasets have led most researchers to study misinformation indirectly, such as via surveys or at the aggregated data level. Surveys are valuable for capturing people's general sentiments about misinformation (e.g., Newman et al. 2020) and how perceived misinformation impacts their trust in information sources (Altay, Hacquin, and Mercier 2022). However, they are limited by the respondents' memory, sincerity, and ability to detect hoaxes (Barthel, Mitchell, and Holcomb 2016). Aggregated data analyses are also valuable for verifying hypotheses using large datasets, but they do not allow content-level observations and are limited by a series of biases. These biases include a strong dependency on ratings by fact-checkers, who have their own guidelines (as not all false news requires moderation) and only review items reaching specific virality thresholds (in terms of engagement and impressions), and whose ratings serve as feedback used to train detection algorithms (responsible for enqueuing relevant content for fact-checking). We aim to overcome these limitations by observing misinformation at the closest level, by conducting a human review of reported content without any preselection based on relevance criteria, and by establishing our own classification to analyze it. We thereby wish to build a dataset reflecting the users' perception of misinformation, rather than that of the fact-checkers.

2.2 Collected data

Users can report content items for various reasons, such as "hate speech," "impersonation," "illegal product sales," or "false news." We refer to users who share content items as "creators" and to those who report specific content items as "reporters." Out of all the reports submitted by Facebook and Instagram users as "false news" between June 3 and July 3, 2020, we extracted a subset S0 from Facebook's internal servers. S0 contains all the reports submitted by users from France, the UK, and the US that had available items' type (link, video, image, carousel), content (picture, video), and report count, as well as basic de-identified information about their creators and reporters (gender, age, country). It should thus be understood that the total number of reports logged in Facebook and Instagram in June 2020 is greater than that of the items in S0. The selection of countries was justified by the authors' language skills (English and French) and

by the aim of including countries for which little research on misinformation has been conducted (France and UK), while allowing comparison with previous research (exhaustive for the US). In addition, because Instagram has only recently allowed people to register their gender, S0 only contains items reported by Instagram users who have linked their account to a Facebook account with a registered gender.

From S0 (N = 97.4M), we extracted a random sample S1 of 11,463 items. The sampling was stratified to ensure a relative balance between reporters' countries, genders, and age categories while maximizing the diversity of content and reporters (i.e., minimizing the number of identical posts shared or reported by different users and the number of different items reported by the same user). After excluding 2,004 items that were removed by users before being labeled and 484 items that included languages other than English and French, we were left with 8,975 items, which received a class label according to the classification described in the following section. Table 1 presents basic information about the resulting dataset.

	FB US	FB UK	FB FR	IG US	IG UK	IG FR
N items in S0	3,946,934	322,829	116,606	83,092,802	8,021,929	1,857,718
% Distinct items	83.5%	77.7%	83.3%	0.2%	0.4%	1.3%
N items in S1	1,527	1,519	1,422	1,500	1,516	1,491
% Distinct items	98.9%	94.2%	96.3%	97.5%	81.5%	96.9%
Reporters' mean age	37.3 (14.7)	37.4 (15.0)	37.4 (15.4)	38.6 (10.7)	39.4 (11.8)	40.1 (10.9)
Reporters' gender (% males)	50.1%	51.5%	52.9%	49%	46%	39%

Table 1: Key metrics of S0 and S1 datasets. FB is Facebook; IG is Instagram.

2.3 Classification methodology

In the absence of an existing classification for content reported by social media users, we used a four-step review to develop an original one to analyze S1's data. An exploratory review of c. 300 items per country allowed us to lay the foundations for the classification. We then reviewed each of S1's items and associated them with a class. The process was ordered to review content from the same media type, then platform, then country to assess differences and similarities across these three layers. Whenever all items from a given platform in a given country were labeled, a confirmation review was conducted for each content class to ensure label coherence. Once all items were classified, a final review was conducted for each class label across all content. In addition to detecting misclassified items and ensuring coherence of the labeling across platforms and countries, this final round also permitted us to include fact-checkers' most recent ratings (as of December 7, 2020). Figure 1 presents the resulting classification composed of six classes and 14 subclasses.

The annotation task was performed by the authors in a sociological manner, which allowed them to take extensive notes about the main topics represented in the reported content and the manipulative techniques observed for content likely to contain a hoax at each step of the review process. Note that this classification does not aim to assess the reporters' performance—that is, in distinguishing reported content containing accurate vs. false claims—but their credibility: in other words, their ability to differentiate content that could contain false news from that which could not. In fact, reporters may be reliable in the sense that they report content items that they genuinely believe to contain false news, but they may fail to report the type of content relevant to moderators—for instance, because it contains a controversial claim that is actually accurate or because fact-checking it would require extensive resources exceeding those available to moderators (e.g., someone accusing their neighbour of poisoning their cat in a small

village). For these reasons, the annotation was performed with a reporter-oriented approach, which relies on two main assumptions: reporter's best intention (RBI)—if any of the possible claims expressed in any layer of the item could reasonably be considered controversial, then it is assumed that the user reported the item because of this claim—and best fact-checking resources (BFR)—considering all identified controversial claims as relevant regardless of the resources required to verify them. These assumptions, together with the four main difficulties associated with the labeling of user reports, are discussed in Appendix A. As a result of the reporter-oriented approach it builds on, our classification is relevant from a research perspective to understand user reports and classify them for the purpose of improving misinformation detection with user-level signals; it should not, however, be interpreted as a turnkey tool to operationalize content moderation.

Class	Sub-class	Description	Examples
Irrelevant (I)		Purely irrelevant content from a moderation perspective. This is content devoid of any claim, announcement, or inappropriate image.	[1-2]
Joker (J)		Content which does not fit in any other class.	
Misreported (M)		Content which, by nature, cannot support misinformation (absence of claim) but which can be considered inappropriate for other reasons.	
	M1	Content that can arguably be considered inappropriate or shocking by some people but is not likely to violate policies (e.g., vulgarity, suggestive but not sexual photos).	[3-6]
	M2	Content that can arguably be considered inappropriate or shocking by many people, with good chances of violating policies and being removed by moderators (e.g., nudity, pornography, violence, hate, bullying).	[7-11]
	M3	Content that can arguably be considered as referring to a problematic interaction but whose problematic dimension is, however, not captured by the current reporting	[12-17]
	MS	Content that may be associated to a scam, spam, fake offer, or false announcement.	[18-20]
Humour or irony (H)		Humoristic or ironic content, which may or may not contain a claim.	
	H1	Content whose ironic aspect is relatively obvious: the claim can hardly be seriously considered as misleading.	[21-25]
	H2	Content whose ironic aspect is quite ambiguous: it is not clear whether the claim is ironic or not.	[26-27]
	H3	Content whose ironic aspect is associated to the mockery of specific individuals or groups of individuals based on a shared characteristic (e.g., dark humour).	[28-30]
Opinion (O)		Content including a claim which typically cannot be fact-checked and is rather presented as a personal opinion.	
	O1	Content referring to an opinion, which may not be arguably considered potentially dangerous.	[31-34]
	O2	Content referring to an opinion, which may arguably be considered potentially dangerous or particularly offensive.	[35-36]
Controversial claim (C)		Content including a claim which typically can be fact-checked, regardless of whether it seems a priori accurate or not or whether it is easy or hard to fact-check.	
	C0	Content including a questionable piece of information that is of low impact, making it too complicated and not significant enough for moderators to fact-check.	[37-39]
	C1	Content including a piece of information which seems to be obviously false and relatively easy to detect for an educated person.	[40-42]
	C2	Content including a questionable piece of information which may or may not be false and would benefit from background checking. This is the content for which the fact-checkers' added value is the greatest.	[43-46]
	C3	Content including a questionable piece of information which may or may not be false and requires some expert-level knowledge to assess. This is content for which fact-checkers' capacities may be too low for a relevant assessment (e.g., expert debates around the efficiency of hydroxychloroquine against Covid-19).	[47]
	C2*	Content including a questionable piece of information which may or may not be false but can arguably be considered harmful, whatever the expertise required to debunk it (e.g., drinking bleach as a cure for Covid-19).	[48]

Figure 1: General classification of reported content. The examples referenced in the fourth column are included in Appendix C. By "claim," we refer to an assertion that can be true or false, regardless of whether it is practically feasible to fact-check it. By "announcement," we refer to a type of claim with relatively low impact that is not associated with a judgement (e.g., fake competitions for the purpose of generating activity on a post or low-scale information such as "our pizzeria will be closed next Monday").

3 Country and platform specificities of content reported as misinformation

Previous research has found that misinformation topics may vary among countries (Humprecht 2019) and that people express concern about different types of content across regions (Newman et al. 2020). In this section, we examine whether these observations are consistent with the content reported as "false news" in S1 and find that such content greatly differs in volume, type, and manipulation technique between countries and platforms. This prompts us to formulate the hypothesis of convergence in the content circulating on both platforms and to identify emerging trends in manipulation techniques used to propagate false news.

3.1 Discrepancies in countries' exposure to misinformation

We start by examining the distributions of S1's items for each class and country, illustrated in Figure 2. The dark blue bars represent the normalized distribution of reported content per class, which appears to be mostly concentrated around the I (irrelevant content), MS (scams, spams, fake offers or announcements), O1 (opinions) and C2 (controversial content) classes on both platforms across the three countries. Examples of such content are represented in Figure 3. The light blue bars represent the deviation of the number of reports between the two platforms as a measure of the homogeneity of content reported on the two platforms in the same country.

A first observation is that the relative homogeneity in M1, M2, and M3 content across countries breaks down at the platform level in France and the UK. M1 can arguably be considered inappropriate by some people but is unlikely to violate moderation policies (e.g., vulgarity, suggestive but not sexual images), M2 is likely to violate these policies (e.g., nudity, pornography, violence, hate, bullying), and M3 has a problematic dimension that is not obviously captured by the reporting labels. This latter type of content seems to be more specific to Facebook, with 159 items on FB FR and FB UK vs. 36 for IG FR and IG UK, while M1 and M2 are more specific to Instagram (resp., 214 on IG FR and IG UK vs. 60 on FB FR and FB UK). This difference may be explained by the fact that M3 items mostly refer to denunciations, warnings, and calls for support, which better fit semantic posts than pictural ones. In contrast, most M1 and M2 items have violent, offending, or sexually suggestive content, for which pictural posts are better suited.

A second point relates to the large difference in the volume of controversial content reported, in particular C1, which is content including a piece of information appearing to be obviously false and relatively easy to detect for an educated person, and C2, which is content including a questionable piece of information that may be false and would benefit from background checking. C1 and C2 are 1.8x less numerous on FB FR than on FB UK and 2.0x less numerous on FB FR than on FB US. This trend is even more apparent on Instagram (resp., 3.6x and 7.3x lower), where only four items were classified as C1 in France (vs. 40 on IG UK and 86 on IG US). These results could be explained by:

(a) a lower volume of false news circulating in France,

(b) lower reporting activity from French users, or

(c) a lower capacity or willingness of French users to report C content.

Hypothesis (b) should be rejected, at least for Instagram, where the number of distinct content items reported per 1,000 monthly active users is very similar across the countries (0.78 FR, 0.81 UK and 0.81 US). Hypothesis (c) is also difficult to defend, because 950 of the 1,492 reported content items on IG FR were associated with a policy

offense (vs. 186 for IG UK, and 133 for IG US), suggesting a serious attitude toward reporting in France. Humprecht, Esser, and Van Aels (2020) also recently observed that French people were four times more resilient to online misinformation than Americans. Finally, a qualitative review of S1 provides another argument supporting hypothesis (a) over hypothesis (c), as we observed that while the great majority of C2 content in the US subsets was very likely to contain hoaxes, in the FR subsets that content was much less likely to express false news. While the latter content was mostly classified as C2 because of the RBI assumption, the content expressed general skepticism toward the government rather than toward false news, especially concerning the management of masks, which was at the center of a political scandal at that time (Moullot and Halissat 2020).

Despite similar reporting rates and polarized contexts (COVID-19 restrictions and anti-police riots occurring in all three countries), a body of corroborating evidence suggests that the circulation of misinformation was significantly smaller in both volume and severity in France than in the US, with the UK occupying an intermediate position. Relatively unconcerned by misinformation issues, French Instagram users, however, experienced a different issue manifested by the significant number of reports associated with spams, with MS items accounting for 58% of IG FR.

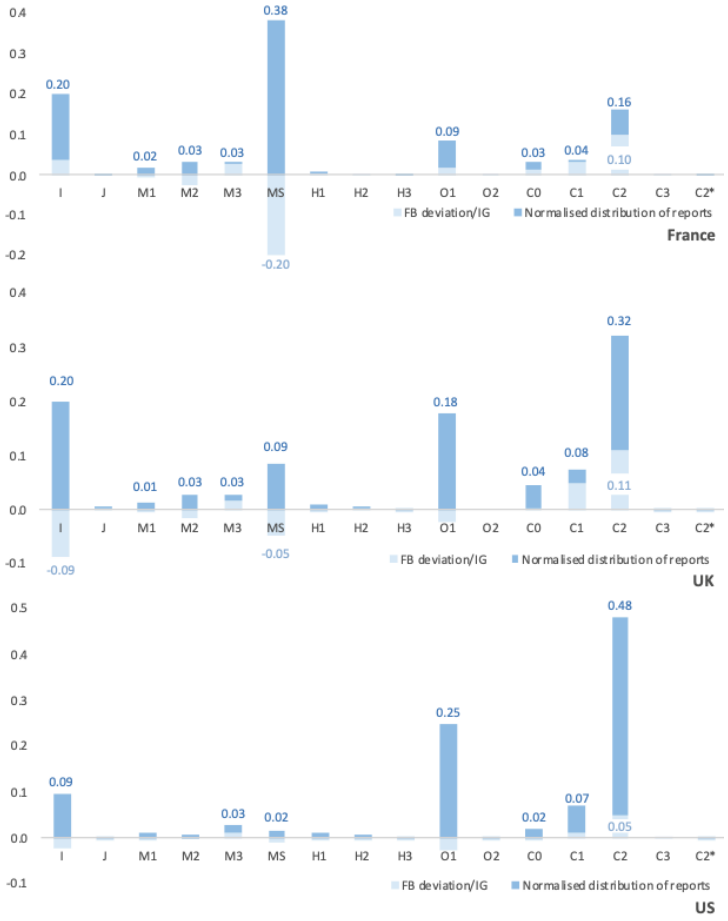


Figure 2: Distribution of reports in S1 for each class, platform, and country. Misinformation appears to be a greater concern on both Facebook and Instagram in the US than in France, where the type of reported content is more heterogeneous.

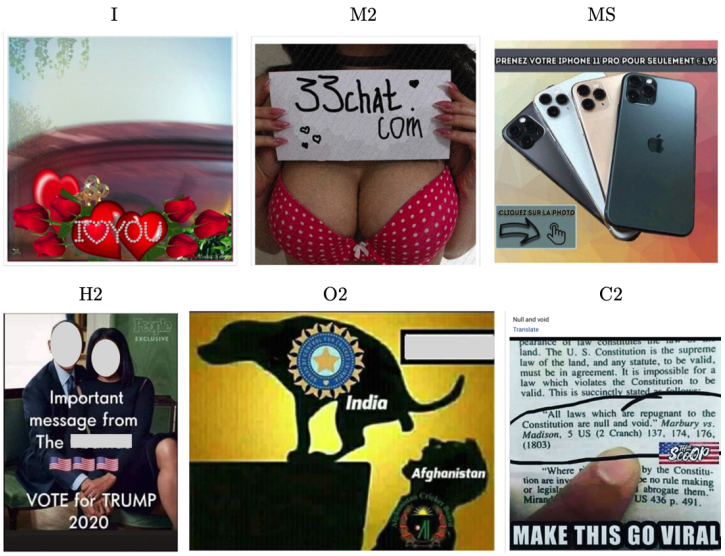


Figure 3: Examples of content rated I, M2, MS, H2, O2, and C2 in S1

3.2 Probable convergence in misinformation content between platforms

The deviation of the number of the reports presented in Figure 2 strongly indicates distinct degrees of similarity in the content circulating on Facebook and Instagram. In France, and to a lesser extent the UK, the distribution of reported content in each class significantly differs between the two platforms. It is clear that C content is predominant on Facebook, rather than Instagram, which is consistent with the belief that misinformation is primarily an issue on Facebook and not on Instagram. Such variations, however, tend to disappear in the US, where the high similarity in the reported content types suggests an increasing uniformization of content on both platforms, revealing that such a belief is not justified anymore. To complement our analysis, we took extensive notes on the main topics expressed in reported content across the subsets of S1 at each step of the annotation process. These topics are summarized in Figure 4 with a qualitative scale representing their prevalence across all subsets.

Figure 4 indicates that misinformation reports differ not only in volume and type across regions and platforms, but also in the topics considered polemical (as reported) by users. Less intuitively, we observe that the relative homogeneity in content between the two platforms in the US is less prevalent in the UK and absent in France. This supports the hypothesis of convergence in misinformation reports between the platforms. In addition to the two universal topics present in all subsets (the pandemic and riots), only one other topic is present on both FB FR and IG FR, namely animal welfare, and it is not even expressed in the same way—IG posts express political claims related to animal rights, whereas FB posts denounce particular cases of animal cruelty. In contrast, reports raised similar concerns about weight loss products and the Armenian conflict on IG FR and IG UK. Furthermore, even when user activity is monopolized by common topics associated with global events, it does not necessarily focus on the same aspects. For example, while most controversial posts on FB US and IG US relate to serious hoaxes about masks, vaccines, and the reality of the pandemic, those on FB FR and IG FR generally criticize the government's restrictions.

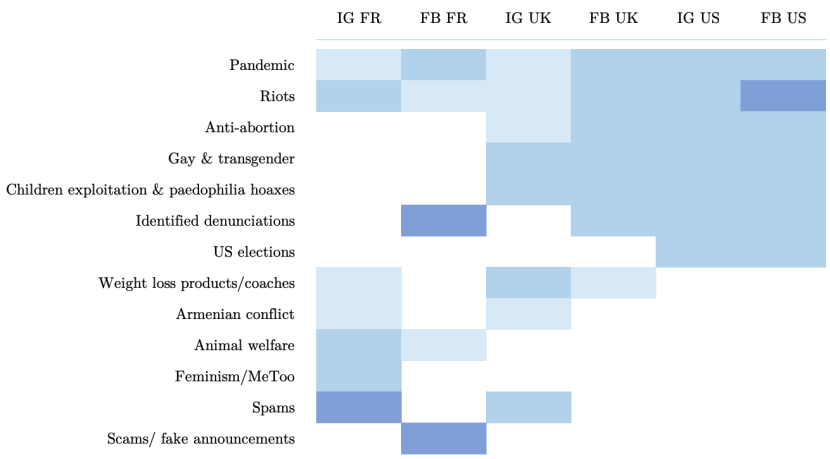


Figure 4: Qualitative representation of the main topics present in subsets of S1. The qualitative scale ranges from light blue (a few items), to medium blue (a significant number of items) and dark blue (a large number of items). While the topics of reported content on Facebook and Instagram are similar in the US, they diverge significantly in France.

We also observed that most of the C2 items from FR and UK with a reasonable likelihood of being misinformation were concentrated in a small cluster of identical posts (n < 10) replicated dozens of times. This contrasts with the US subsets, where C2 posts with a high likelihood of being hoaxes are highly diversified. Furthermore, we found some identical controversial posts in both the FB US and FB UK samples and to a lesser extent in the FB FR samples. These posts are likely to have originated from the US and are often related to COVID-19 but also to US-centered events (Black Lives Matter [BLM] riots, including local events such as hoaxes about vandalized statues and cemeteries). In contrast, no posts related to French or English topics were found in the US samples. This suggests that several hoaxes are transmitted from the US to Europe and better circulate between countries on the same platform than between platforms in the same country. This was not obvious: one could have expected language to be a strong barrier to content circulation, all reporters of S1 content also have a Facebook account linked to their Instagram account, and many IG US C2 items appear to be screenshots of Twitter posts. Another possibility, although less likely, is that such content relating to US topics was generated outside the US.

Overall, while the value of the above qualitative observations should not be overestimated, the regional discrepancies in the deviation of the number of reports and in the similarity of reported topics between platforms, together with the nonreciprocal presence of US content in the UK and FR subsets, support the hypothesis of convergence in misinformation reports—and likely, by extension, of content in circulation—between the platforms, probably driven by the US.

3.3 Manipulation techniques vary by region and platform

The notes taken alongside the annotation process allowed us to identify six main manipulation techniques expressed by the reported content in S1 and likely to derive from disinformation actors (see Appendix B for additional information):

1. *The revelation* challenges people's egos and encourages them to "wake up" instead of being "sheep," making direct references to the cabal, Masons, a world elite, and a new order.

2. *The critical tipping point* leverages a real fact (e.g., a public scandal) as an entry point to encourage people to reconsider their beliefs ("If they lied about this, what else did they lie about?").
3. *False facts supported by false evidence* are typically false statements about events that have allegedly occurred and false quotes from public figures, often backed by inauthentic documents, scientific data, or personal testimonies.
4. *The misleading presentation of facts* presents authentic documents or accurate facts in a misleading way to encourage a targeted erroneous interpretation.
5. *The confusion of feelings* presents either false news or authentic facts in a twisted way to provoke an emotional reaction against a given target, often mobilizing symbols (e.g., protesters' violence against a veteran, profanation of a military cemetery).
6. *The excuse of casualness* characterizes items with humor or an artistic dimension, making an implicit reference to a widely known hoax using a frivolous tone.

A well-established finding in the social psychology literature on persuasion is that people do not respond identically to different types of rhetorical strategies (Barbier and Fointiat 2015). Hovland and Lumsdain (1949), for instance, found persuasion to be more efficient on targets with low knowledge about a given topic, while conviction works better on informed subjects. The former are also more receptive to simple messages (McGuire 1968) and the latter to the use of complex language (Eagly and Warren 1976). Our six techniques can be associated with different psychological leverages: (1) and (2) play on people's skepticism, (3) and (4) seem better designed for people sensitive to pragmatic arguments, (5) builds upon empathy, and (6) plays on the register of *coolness*. While the first five techniques are consistent with existing false news typologies (Wardle 2016; Yurkova 2018; Molina et al. 2021; Innes 2020) and mostly apply to traditional disinformation strategies on Facebook, the *excuse of casualness* seems both recent and specific to Instagram, especially IG US, emerging from the type of content shared on the platform. It characterizes items with humor (jokes, ironic statements, memes, caricatures) and/or an artistic dimension (cartoons, short format videos, songs), making implicit reference to a hoax using a frivolous tone. This content differs from both "parodies" (Wardle 2016) and "satires," as it is not "meant to be perceived as unrealistic" (Molina et al. 2021, 198) but rather hides an assumed claim behind a veil of frivolity. Based on suggestion and sarcastic humor, this content does not introduce a new hoax but rather may constitute an efficient second-layer relay to support a hoax's viral propagation. The apparent lack of seriousness makes the content more difficult to both detect and moderate.

We observed significant discrepancies in the range of manipulation techniques deployed across countries and platforms: the small amount of controversial content on FB FR was mostly associated with (1) and (2), while (1), (2), (3), and (5) were leveraged in the UK, and all strategies were used in the US—(1) to (5) on Facebook and principally (6) on Instagram. Additionally, fake reshares and screenshots of fabricated posts of public figures were found to be particular to IG US (and also found on FB US but not elsewhere), while the testimonies of false individuals (e.g., "my friend working at the CDC said…") are particular to FB US. From a general perspective, C2 content was found to be significantly more subtle on IG US than elsewhere: claims were more suggestive and often based on multimodal combinations (picture and caption), and satires and parodies were less obvious, such as memes presenting rioters with streamers, whose text was edited in a nonobvious way. Another typical example is the trend to create parodies of Donald Trump's style of Twitter posts, which may result in the unintentional diffusion of misinformation. In addition to the excuse of casualness, the increasing use

of video formats makes it more difficult to detect false claims made within long videos that also include personal opinions and testimonies.

The strategies used to spread false news by coordinated groups of inauthentic actors are still technically evolving to avoid bulk detection (Goldstein and Grossman 2021). At the content level, however, misinformation innovation does not seem to proceed from technological advances such as deepfakes—similarly to Brennen et al. (2020), we observed that altered posts originate from low-tech photo and video edits—but from language refinement and social cues. Despite the apparent convergence between Facebook and Instagram in the types and topics of reported content in the US, the techniques used to propagate false news still differ. By design, social media platforms encourage the sharing of different types of content, and misinformation propagators adapt their message accordingly. This follows the evolution of marketing techniques on social networks, recently illustrated by tobacco companies' efforts to leverage the *coolness* of Instagram influencers to advertise e-cigarettes, vapes, and nicotine pouches (Chapman 2021), or Mike Bloomberg's strategy to hire Instagram influencers "to make him seem cool" with memes (Noor 2020). While people do not necessarily believe the false news they share (Pennycook and Rand 2021), Fazio, Rand, and Pennycook (2019) observed that repeating a claim increased its perceived truthfulness, which could result in a new kind of "grey" content that is just as harmful as fabricated news. Presenting these posts as frivolous not only makes them more difficult to detect and moderate but may also increase their virality potential and thus their impact due to their repetitiveness.

4 A plurality of reporters' profiles allows various leverages for noise reduction

While misinformation appears to vary with the country and platform, we postulate the main reasons why users report content to be more universal. In this section, we define four profiles of reporters and aim to detect the most "credible" ones, i.e., those whose motivations are aligned with moderators' expectations. We find that our classification can explain 55% of the inaccuracy in user reports, demonstrating that most of the noise in the reporting signal does not derive from a lack of seriousness of reporters but from other reasons, and we suggest means of actions to reduce the inaccuracy. We then train a model capable of identifying the reports most relevant to fact-checkers, those least relevant, and those relevant for content moderators but misclassified.

4.1 Four main profiles of reporters

It is certainly questionable to attempt to infer users' intentions based on their reporting activity. The diversity of reported content, however, supports the hypothesis of a plurality of reporting profiles associated with distinct goals, which several signals allow us to identify.

1. *Reporting false news to trigger moderators' actions*. This is the expected use of the reporting feature, and several signals strongly support this hypothesis. For example, 41.8% of all labeled content pieces in S1 were classified as controversial, of which 17.3% were confirmed to be false news by fact-checkers, suggesting the accurate use of the reporting tool by a significant proportion of reporters. Comment sections are also used by a number of users to label misinformation by posting comments expressing disbelief (e.g., "fake," "fake news," "this is a hoax"), communicating that they have reported a post and are encouraging others to do so (e.g., "reported!," "this is fake, report it"), often

providing links to material from fact-checking websites that debunks claims. Some users even act as "super reporters," flagging a large number of relevant content items. For instance, 43 Instagram users from S0 were responsible for more than 1,000 reports each over 90 days (May–July 2020). One user logged 2,962 reports for 60 distinct content pieces, of which 11 were labeled in S1, containing 8 C2, 2 C0, and 1 MS. Another one logged 1,175 reports for 37 distinct items; 10 were in S1, among which 5 were rated C2, 4 C1, and 1 C0, and 8 out of the 10 were confirmed to be misinformation by fact-checkers.

2. *Reporting to annoy the content creator*. The significant amount of O content, especially on FB US (22.3%), suggests that many users report opinions they disagree with and news they believe but dislike. This is consistent with previous research on Instagram (Grossman et al. 2020; Smyrnaios and Papaevangelou 2020). In addition, a significant proportion of reported items do not even contain a claim, nor do they qualify for other policy violations. This suggests that reporting is used not only as a "dislike button," namely to send negative feedback to users expressing divergent opinions, but also to annoy them, perhaps to express jealousy resulting from negative social comparison. This hypothesis is suggested by the very large amount of I content, of which many items are related to romantic relationships (pictures of couples, often with captions expressing love and happiness, or public notifications such as "X is in a relationship with Y") and body image (selfies at the gym), two topics known for triggering negative social comparisons (Burke, Cheng, and Gant 2020). As a limitation to this a priori irrelevant reporting, it should be noted that a number of I items, although not problematic at the content level from a misinformation perspective, were associated with user-level offenses (e.g., fake accounts, spam, impersonation, property rights). This association applies to 7.5% of I content in S1, but it may apply to a larger number of reported items for which the violation has not yet been detected.

3. *Reporting inappropriate content that is not misinformation*. All of the M content (23%, of which 55% was confirmed to be policy-breaking) supports the hypothesis that users "misreport" some items; this content may nonetheless be problematic, but the user has selected an incorrect option. M1 content, especially on Instagram, often contains images that are sexually suggestive or show quasi-nudity but do not qualify as pornography or sexual solicitation. M2 content often includes expressions of brutality, and MS content relates to scams, spams, and fake accounts. It is understandable that users may associate scams and inauthentic accounts with false information, although dedicated categories exist to specifically report them. This supposed "mistake" is, however, more surprising for M1 and M2 content, because to report a post as false news on Instagram a user must select "inappropriate" instead of "spam," then scroll down to the "false news" option, which is the last one in a list of eight categories that better fit all types of items classified as M1/M2.

4. *Reporting to draw the moderators' attention to a problematic situation*. M3 items are posts soliciting the user community to support or be aware of an issue that is not primarily political and usually has a personal connection to the content sharer. They vary from warnings (reporting scams and bad experiences with businesses or artisans) to unidentified accusations (e.g., "the waiter of company X was racist to me") and identified exposure ("this man is a racist, expose him," accompanied by screenshots of private messages), often calling for public shaming and sometimes including serious indictments ("X sexually assaulted me last week"). We found these accusations in every country (51 on FB US, 74 on FB UK, 60 on FB FR),

but it is difficult to tell whether the aim of the reporters was for moderators to take action against the content creator (a few of these posts turned out to be associated with harassment or created by false accounts), to flag a danger associated with the public shaming of a potentially innocent person, or to support the post sharer in the hope that moderators might alert the police. This latter hypothesis could explain why, when users explicitly ask the community for support ("please block X," "please report X," "report this false news"), users often report the whistleblower instead of the post or user they are asked to flag [50,51,52].

The absence of other typologies of reporting profiles prevents us from comparing this typology with other ones. Therefore, we suggest that these categories are considered a first draft of a reporting profile typology, which would benefit from further research using different methods, notably psychometric analysis and sociological interviews.

4.2 Splitting the noise to increase reporting accuracy

By postulating various profiles of reporters, we assume that not all reports have the same value for content moderators. While reports from profile (1) reporters are the most relevant for fact-checkers, the others are nevertheless not equally "inaccurate." For instance, reports from profile (2) reporters are truly irrelevant, while those from profile (3) reporters are simply misclassified. This leads us to reevaluate the accuracy of user reporting and assess its performance using more suitable metrics.

The traditional approach captures the accuracy of reporting using the ratio between the total number of reports and that of verified hoaxes confirmed by fact-checkers. To calculate this ratio, we compare our labels with those from the databases of Facebook and Instagram moderators; we denote the items in S1 that were reviewed by fact-checkers and confirmed to contain false news as VM (verified misinformation), and we denote those detected by moderators as containing a policy violation other than "false news" as VO (verified other). We obtain a value of VM/N(S1) = 0.07, indicating that 93% of reported items in S1 are inaccurate.

This metric is nevertheless too simple; it is limited by the subjectivity of fact-checkers, with whom reporters may argue and who may also be incorrect, as illustrated by the significant changes in ratings observed as late as nine months after our data collection. It also ignores how many of the S1 items were reviewed by fact-checkers, as not all reported content is necessarily reviewed by moderators. Finally, it does not distinguish between purely irrelevant reports and those that are just misclassified. A better approach involves splitting this "relative" inaccuracy into different degrees of noise, which can be linked to our reporter profiles and based on which distinct actions can be taken to clean the overall signal.

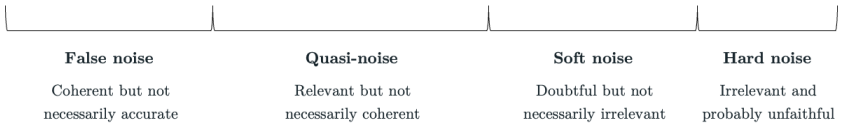
- *False noise* = $\frac{\sum(C1,C2,C3,C0)}{N(S1)}$ indicates the proportion of coherent, although not necessarily accurate, reporting. It accounts for 0.35 of the overall inaccuracy and could be associated with reporter profile (1). We call it false noise as it probably results mainly from credible reporting by people with a low capacity to detect false news or understand the moderation policies and types of misinformation prioritized by fact-checkers (e.g., C0). The use of educative campaigns, communication to explain platforms' moderation policies, and self-fact-checking materials could help reduce this noise.
- *Quasi-noise* = $\frac{\sum(VO,M2,MS)}{N(S1)}$ indicates the proportion of relevant, although not necessarily coherent, reporting. It accounts for 0.2 of the overall inaccuracy and could be associated with reporter profile (3). We call it quasi-noise because it refers to content that can be moderated but not in the context of misinformation.

The reporting is probably credible, and additional investigations should be conducted to understand the source of the confusion surrounding the reporting feature.

- *Soft noise* = $\frac{\sum(M1,M3,H2,H3,O2)}{N(S1)}$ indicates the proportion of doubtful, although not necessarily irrelevant, reporting. It accounts for 0.03 of the overall inaccuracy and could be associated with any reporter profile. We call it soft noise because it is difficult to make any strong assumptions about it.
- *Hard noise* = $\frac{\sum(H1,O1,I)}{N(S1)}$ indicates the proportion of probably irrelevant reporting. Accounting for 0.34 of the overall inaccuracy, it could be associated with reporter profile (2). We call it hard noise because it probably results from unfaithful reporting that should be filtered.

Figure 5 presents the cumulative proportion of S1 items distributed in each content class and associated with reporters' profiles. To reflect the different types of noise, the classes are ordered by relevance, from the most relevant content on the left, associated with profile (1) reporters, to the least relevant on the right, associated with profile (2) reporters. The figure reveals that at least 55% of the reporting inaccuracy (false noise and quasi-noise) should probably not be attributed to the reporters' lack of seriousness but rather to confusion surrounding the reporting features and moderating rules. As a limitation, our RBI assumption probably inflates the amount of C2 content, categorizing content as "false noise" when it actually is unfaithfully reported. Nonetheless, some C items were not reviewed by fact-checkers and could have been rated VM. Likewise, a proportion of the "hard noise" may also contain undetected policy-breaking content, which may be rated VO later.

Profile	1	1	1/2	1/2	3	3	3	3/4	3/2	1/2	2	2	2	
Class	VM	C1	C2/C3	C0	VO	M2	MS	M3	M1	H2/H3	O2	H1/O1	I	J
All	0.07	0.12	0.39	0.42	0.55	0.56	0.62	0.62	0.62	0.66	0.66	0.84	1.00	1.00
FB	0.13	0.20	0.51	0.54	0.56	0.56	0.63	0.63	0.63	0.69	0.69	0.86	1.00	1.00
US	0.19	0.24	0.62	0.63	0.64	0.64	0.65	0.65	0.65	0.70	0.70	0.93	1.00	1.00
UK	0.13	0.23	0.56	0.61	0.62	0.63	0.66	0.66	0.66	0.72	0.72	0.88	0.99	1.00
FR	0.07	0.12	0.34	0.38	0.41	0.41	0.57	0.57	0.57	0.64	0.65	0.76	1.00	1.00
IG	0.02	0.04	0.27	0.30	0.55	0.57	0.61	0.61	0.61	0.64	0.64	0.82	1.00	1.00
US	0.05	0.10	0.49	0.52	0.55	0.55	0.57	0.57	0.57	0.61	0.61	0.88	1.00	1.00
UK	0.01	0.03	0.24	0.28	0.39	0.42	0.50	0.50	0.50	0.53	0.53	0.74	1.00	1.00
FR	0.00	0.00	0.07	0.09	0.72	0.73	0.75	0.75	0.75	0.78	0.78	0.85	1.00	1.00



False noise
Coherent but not necessarily accurate

Quasi-noise
Relevant but not necessarily coherent

Soft noise
Doubtful but not necessarily irrelevant

Hard noise
Irrelevant and probably unfaithful

Figure 5: Cumulative proportion of content items in S1 for each class associated with reporters' profiles. Reading the table from left to right, we observe that while only 7% of S1 items were confirmed by fact-checkers to be misinformation, 62% are very likely to be relevant for content moderators.

4.3 Leveraging multisignal reporting to improve misinformation detection

Considering that the relevance of reports varies among reporters, that their relevance depends on the reporters' profiles, and that the inaccuracy in reporting can be distributed among different types of noise, we should now find a means of detecting these noise types among reported content. Our objective is to identify false noise, most relevant for misinformation moderation purposes, and quasi-noise, relevant for other moderation channels. We choose to focus on the Instagram subsets, as our previous observations suggest that it is the platform on which innovative

manipulation techniques emerge, challenging current computational tools used to detect misinformation, and therefore the platform with the greatest need for a novel detection approach. To better match the types of profiles and the types of noise previously identified, we aggregate classes as C (C0, C1, C2, C3, C2*), M (M2, M3, MS), HM (H2, H3, M1), OH (H1, O1, O2), and I.

The first three graphs in Figure 6 represent the distribution for each aggregated class of the number of reports received for each Instagram item of S1 over 90 days. We consider a 0.999 quantile, excluding 4 outliers (1 C, 1 OH, 1 M, 1 I) with a number of reports exceeding 329,279. The total number of misinformation reports received for an item clearly appears to be a meaningful signal that can be used to distinguish C from other classes, especially from I and M. However, it seems to be less accurate in differentiating C and OH below a threshold of 20,000 reports and incapable of separating M from I. We confirm these observations by comparing the means of the dependent variable, i.e., content reports per class, for each pair of aggregated classes. The results are presented in the fourth graph, where the p-values displayed correspond to each two-by-two comparison for all statistically significant relationships. We use Welch's t-test to accept a one-sided alternative hypothesis with a significance level of 5% (or 10% for grey arrows). We do not assume the same variance among classes since the reporting profile may significantly vary, and it is empirically verified. The mean normality assumption is verified by the D'Agostino–Pearson test. Kolmogorov–Smirnov tests were also conducted to reveal the significant shift in the distribution between the aforementioned categories.

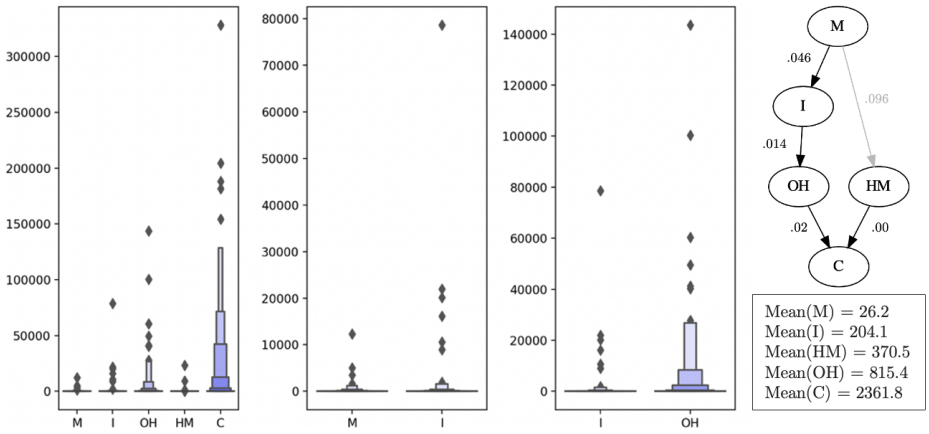


Figure 6: Distribution of Instagram content report count for each aggregated class and partial order of aggregated classes based on number of reports. The total number of reports logged against each content item appears to be a meaningful signal for distinguishing C from other classes, especially from I and M.

Finally, having seen that misinformation reporting is a complex signal whose overall accuracy is undermined by relevant but misclassified content, we expect that combining this signal with other reporting signals will increase its quality. We train a gradient-boosting classification model to identify four classes (C, M, I, others) from 10 features corresponding to the main reporting signals of the platforms (false news, nudity/sexual solicitation, violence, harassment, suicide/injury, spam, hate speech, unauthorized sales, inappropriate content, "I don't like it") on IG(S1) with a test sample of 10%. The model's general performance, presented in Figure 7, reaches $F1 = 0.56$ for $C/\neg C$ and $F1 = 0.63$ for $M/\neg M$. More interesting is the performance for each country: the model gives performances of $F1 = 0.84$ for $M/\neg M$ on IG FR, where spam was identified as the

main issue, and $F1 = 0.72$ on IG US, where misinformation was identified to be the main issue. While they significantly vary in order between countries, the most important features are "false news," "spam," "hate speech," and "inappropriate content."

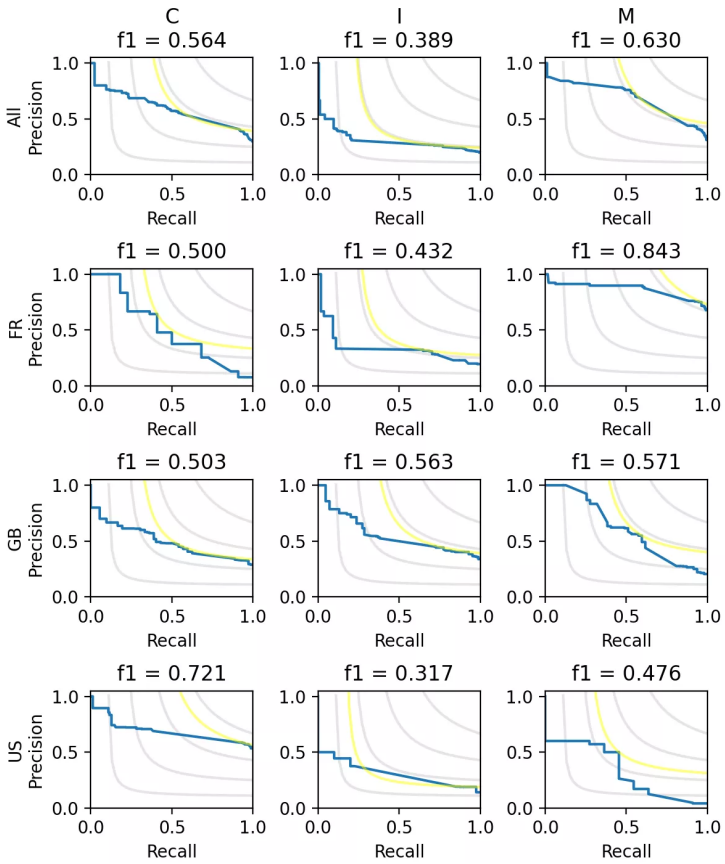


Figure 7: Precision–recall curves for each country for aggregated class detection on S1(IG). The performance increases when the model is trained at the country level; it is particularly interesting when detecting misclassified content in France and controversial content in the US.

5 Discussion

Our original approach to studying misinformation from the reporters' perspective, comparatively between regions and platforms, and at the content level allowed us to make important contributions to the literature on misinformation. Several limitations of this study should nevertheless be considered.

In Section 2, we introduce the first classification of content reported as misinformation and describe the classification methodology as well as the constitution of our novel dataset in the hope that they will serve future research. A first limitation relates to the fact that the dataset was annotated by the authors, who may have introduced excess subjectivity into the labels. As discussed in Appendix A, the review of misinformation reports is a complex task requiring great rigor, expertise related to misinformation, and familiarity with the political context of the considered countries. For these reasons, we decided that professional annotators would not be suitable for conducting this task, informed by a previous similar study in which the recourse to professional annotators

did not provide a sufficiently good annotation to analyze the data. In addition, as further explained in the Appendices, while the GCRC's subclasses are admittedly more subjective than classes, the latter are based on the content's nature (e.g., does it contain a claim or not?), making them sufficiently objective to support our analyses. As supporting evidence of this, 97% of the S1 items that had been reviewed and classified as misinformation by fact-checkers were rated C, with the remaining 3% relating to O1 content, for which we disagree with the fact-checkers.

A second limitation derives from the assumptions underlying the classification, namely the RBI and BFR. These assumptions result from our reporter-oriented approach and are justified by the paper's goal: unlike most research in the misinformation literature, we did not aim to detect false news and "bad actors" but rather to assess the credibility of user reports in order to improve the relevance of this specific signal for supporting false news detection methods. As a result, many C2 items that we interpret as credible reports (because they contain a falsifiable claim) may in fact be unfaithful, resulting from profile (2) reporters rather than profile (1) reporters. It is impossible to know whether reporters genuinely consider an item to contain a false claim or believe it is accurate but report it because they dislike it.

In Section 3, we combine qualitative and descriptive statistical methods to analyze the dataset. We conclude that countries are not equally exposed to misinformation, which appears to be a significantly smaller issue in France than in the US; we observe probable convergence in misinformation content between platforms, which seems to be driven by the US; and we identify a novel manipulation technique emerging on Instagram in the US. The two latter findings result from the extensive notes that were taken when reviewing the content and the body of converging signals. Therefore, they should not be interpreted as solid evidence of emerging trends in online misinformation, but rather as exploratory hypotheses, calling for further investigation in the absence of other misinformation research on Instagram.

With regard to the first finding, we tend to draw conclusions vis-à-vis the amount of misinformation circulating on the platforms from the study of misinformation reports. Such an association is questionable; for instance, reports could fail to account for specific types of false news that users reported particularly poorly. Although limited, we hold user reports to be a better proxy for misinformation than fact-checkers' labels, which are commonly used by researchers. Because the perspective of reporters avoids a number of observation biases discussed in Section 1, we believe that adopting this perspective rather than that of fact-checkers provides a more comprehensive view of misinformation. An illustration of this is the novel manipulation technique we were able to identify, whereas such content would not likely be captured by the detection algorithms for review by fact-checkers. For these reasons, we believe that our findings about the distribution of user reports are generalizable to that of false news in circulation.

In Section 4, we identify four profiles of reporters, split the inaccuracy of the reporting signal into four types of noise, and regroup content classes to associate them with each noise type. The proposed metrics cast a new light on the credibility of user reports, showing that 62% of content in S1 is likely to be credible and relevant for content moderation. They also allow us to offer explanations for 55% of the inaccuracy in misinformation reporting, together with specific means of actions to reduce each type of noise. We then show that the total number of reports logged against a content item is a relevant signal for distinguishing between three key categories of content (irrelevant items that should be filtered, controversial claims that are relevant for fact-checkers, and misclassified items that should be redirected toward other moderators) and that combining misinformation reports with other reporting signals enables a classifier to

better detect these categories. We also show that training a classifier on regional data rather than on a language-based dataset, as is common practice in the industry, results in better performance with regard to the specific moderation issues of each country, namely misinformation in the US and spam in France for Instagram.

The assumptions underlying our classification certainly inflate the number of credible reports, as many C2 items may result from unfaithful reporting. Nonetheless, this effect is somewhat counterbalanced by the number of faithful reports categorized as "soft noise" or "hard noise," which contain policy violations other than false news and are undetected by moderation tools. Our model's performance should also be interpreted in light of the research goal. False news detection classifiers are considerably more sophisticated than our gradient-boosting classifier. Sepúlveda-Torres et al. 2021, for instance, achieved 94% accuracy on the Fake News Challenge FNC-1 dataset. The aim of our model is not to detect the type of content that fact-checkers would rate false news, but rather to identify the most relevant reports to be sent to fact-checkers and to redirect misreported items to more suitable moderators. In contrast with highly sophisticated detection models trained on massive datasets and leveraging many input signals, the value of our model is its capacity to achieve a promising performance despite its extremely simple architecture, its training on a very small dataset (n = 4,056), and its processing of only 10 basic input human signals associated with the categories users employ to report a content piece. In summary, while state-of-the-art misinformation classifiers leverage multimodal architectures to grasp the meaning of a post in relation to its semantic and pictural components, our model is blind to the content itself and ignores personal information about both the content creator and the reporter, as well as contextual information about the content's sharing and reporting.

Finally, the identification of "super reporters" complements that of reporters acting in bad faith, such as users weaponizing the reporting feature to censor divergent opinions (Grossman et al. 2020; Smyrnaios and Papaevangelou 2020). While the analysis in Section 3 is based on the total number of reports logged against a content piece, all the observations presented in Sections 1 and 2 converge on the idea that the credibility of reporters would be a better predictor of the likelihood of an item containing false news than the number of reports. Some actors report content for malicious reasons and others report the same item hundreds of times. Consequently, we strongly believe that weighting user reports by a score reflecting the credibility of the reporting user based on the relevance of their reporting history offers an effective way to improve the accuracy of reporting and, therefore, of misinformation detection tools.

6 Conclusion

By studying misinformation from the reporters' perspective and at the content level, our goal was to refute the idea that user reports are a low-accuracy signal poorly suitable for online misinformation detection. We demonstrate that user reporting is a complex signal composed of different types of feedback that should be understood and assessed separately. When approached as such, reporting offers a valuable signal not only for enhancing the performance of false news detection algorithms, but also for identifying emerging trends in misinformation practices. This approach paves the way to more participative moderation frameworks that reward the best reporters by prioritizing their feedback over those of adversarial actors. The meaningful variations in the volume, type, topic, and manipulation technique of misinformation observed between countries and platforms also support the claim that misinformation is anything but a globally uniform phenomenon. Instead of generalizing the findings of US-centered studies,

researchers, industry players, and policymakers should examine misinformation with respect to the specificities of each country and platform. As the first of its kind, this study, however, remains exploratory, inviting further research to investigate the observations presented and to challenge its provisional conclusions.

References

Altay, Sacha, Anne-Sophie Hacquin, and Hugo Mercier. 2022. "Why do so few people share fake news? It hurts their reputation." *New Media & Society* 24 (6): 1303–24.

Barbier, Laura, and Valérie Fointiat. 2015. "Persuasion et Influence: changer les attitudes, changer les comportements. Regards de la psychologie sociale." *Journal d'Interaction Personne-Système* 4 (1): 1–18.

Barthel, Michael, Amy Mitchell, and Jesse Holcomb. 2016. *Many Americans believe fake news is sowing confusion.* Pew Research Center.

Brennen, J Scott, Felix M Simon, Philip N Howard, and Rasmus Kleis Nielsen. 2020. *Types, sources, and claims of COVID-19 misinformation.* Reuters Institute for the Study of Journalism.

Burke, Moira, Justin Cheng, and Bethany de Gant. 2020. "Social comparison and Facebook: Feedback, positivity, and opportunities for comparison." In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems,* 1–13.

Chapman, Matthew. 2021. "New products, old tricks? Concerns big tobacco is targeting youngsters." *The Bureau of Investigative Journalism.*

Cinelli, Matteo, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. "The echo chamber effect on social media." *Proceedings of the National Academy of Sciences* 118 (9): e2023301118.

Eagly, Alice H, and Rebecca Warren. 1976. "Intelligence, comprehension, and opinion change." *Journal of Personality* 44 (2): 226–42.

Fazio, Lisa K, David G Rand, and Gordon Pennycook. 2019. "Repetition increases perceived truth equally for plausible and implausible statements." *Psychonomic Bulletin & Review* 26:1705–10.

Goldstein, Josh A., and Shelby Grossman. 2021. *How Disinformation Evolved in 2020.* Brookings.

Grossman, Shelby, Ross Ewald, Jennifer John, Asfandyar Mir, Kim Ngo, Natasha Patel, and A. R. 2020. *Reporting for Duty: How A Network of Pakistan-Based Accounts Leveraged Mass Reporting to Silence Critics.* Stanford Internet Observatory.

Hovland, Carl Iver, and Arthur A Lumsdaine. 1949. *Experiments on Mass Communication.* Princeton University Press.

Humprecht, Edda. 2019. "Where 'fake news' flourishes: a comparison across four Western democracies." *Information, Communication & Society* 22 (13): 1973–88.

Humprecht, Edda, Frank Esser, and Peter Van Aelst. 2020. "Resilience to online disinformation: A framework for cross-national comparative research." *The International Journal of Press/Politics* 25 (3): 493–516.

Innes, Martin. 2020. "Techniques of disinformation: Constructing and communicating 'soft fact' after terrorism." *The British Journal of Sociology* 71 (2): 284–99.

McGuire, William J. 1968. "Personality and attitude change: An information-processing theory." *Psychological Foundations of Attitudes* 171:196.

Molina, Maria D, S Shyam Sundar, Thai Le, and Dongwon Lee. 2021. "'Fake news' is not simply false information: A concept explication and taxonomy of online content." *American Behavioral Scientist* 65 (2): 180–212.

Moullot, Pauline, and Ismaël Halissat. 2020. "Masques: comment le gouvernement a menti pour dissimuler le fiasco." *Libération.*

Newman, Nic, Richard Fletcher, Anne Schulz, Simge Andi, and Rasmus Kleis Nielsen. 2020. *Reuters Institute Digital News Report 2020.* Reuters Institute for the Study of Journalism.

Noor, Poppy. 2020. "Mike Bloomberg will pay you $150 to say nice things about him." *The Guardian.*

Pennycook, Gordon, and David G Rand. 2021. "The psychology of fake news." *Trends in Cognitive Sciences* 25 (5): 388–402.

Rashkin, Hannah, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. "Truth of varying shades: Analyzing language in fake news and political fact-checking." In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing,* 2931–37.

Sepúlveda-Torres, Robiert, Marta Vicente, Estela Saquete, Elena Lloret, and Manuel Palomar. 2021. "Exploring summarization to enhance headline stance detection." In *Natural Language Processing and Information Systems,* 243–54.

Smyrnaios, N, and C Papaevangelou. 2020. "Le signalement sur les réseaux sociaux, un moyen de modération mais aussi de censure." *La Revue Des Médias,* https://larevuedesmedias.%20ina.%20fr/signalement-reseaux-sociaux-moderation-censure.

Wardle, Claire. 2016. "6 types of misinformation circulated this election season." *Columbia Journalism Review* 18.

WHO. 2020. "Managing the COVID-19 infodemic: promoting healthy behaviours and mitigating the harm from misinformation and disinformation." *Joint statement by WHO, UN, UNICEF, UNDP, UNESCO, UNAIDS, ITU, UN Global Pulse, and IFRC, 23 September 2020.*

Yurkova, Olga. 2018. "Six Fake News Techniques and Simple Tools to Vet Them." *Global Investigative Journalism Network,* https://gijn.%20org/sixfake-news-techniques-and-simple-tools-to-vet-them.

Authors

Hubert Etienne is an AI ethicist and UX researcher at Meta, New York.

(hae@meta.com)

Onur Çelebi is a research engineer at Meta, Paris.

Acknowledgements

We would like to thank Joelle Pineau, Antoine Bordes, and Jerome Pesenti for their support in having this paper published.

Data availability statement

All data generated or analyzed during this study are included in this published article within the associated data.csv file. The pictures associated to users' posts are, however, not released for privacy reasons and to comply with current regulations.

Funding statement

Both authors are Meta's employees and acknowledge that this affiliation presents a conflict of interests considering the topic of this research. However, the research was conducted with the highest integrity standards and the authors engage their probity to attest that their work did not suffer from any influence.

Ethical standards

The research was approved by Meta's internal review board and complies with existing regulations.

Keywords

Misinformation; false news; reporting; Facebook; Instagram.

Appendices

Appendix A: Four difficulties in labeling reported misinformation content

Labeling content from a misinformation viewpoint is a complex task and not an exact science. It requires someone to (1) identify a claim within a structured piece of content. If this claim is explicit, they must (2) determine whether it could theoretically be falsified and how difficult the verification process would be in practice. When it is implicit, they must (3) assess the degree of obviousness of the underlying claim. Finally, (4) the difference between the subjective perception of the content sharer and the reporter should also be considered. We identified four main difficulties associated with these steps.

The first difficulty is design-specific, resulting from the content's structure. Some items are multilayer posts, composed of content pieces (e.g., a link, picture, or video), metacontent (e.g., a caption or edits of a video), and a second level of metacontent (e.g., a caption reacting to a reshared post that already included a caption), making identifying the reported claim difficult. To address this difficulty, we adopted a holistic reporter-oriented approach, employing the RBI assumption: if any of the possible claims expressed in any layer of the item could be reasonably considered controversial, we assumed that the user reported the item because of this claim. We herein aimed to minimize the number of false negatives—items labeled as irrelevant that the reporters tried to report accurately—at the cost of false positives. Consistent with our position of assessing the content's falsifiability and not its veracity, there is a high ratio of C2 classes among shared links on Facebook (c. 40%), as article headlines often contain factual claims. The RBI assumption, however, allowed us to realize that many reports that seemed irrelevant at first did not originate from frivolous reporting but were associated to posts containing policy violations that were not misinformation.

The second difficulty relates to a claim's degree of falsification, as two claims may be equally true or false but differ in the resources required to verify them. Consistent with our reporter-oriented approach, we adopted the BFR assumption, considering all identified controversial claims as relevant regardless of the resources required to verify them. Reporters have little information about the fact-checkers' capacities, which should thus be orthogonal to their reporting credibility. While we did not aim to rank controversial claims by importance, it was nevertheless possible to draw a relevant distinction between news with small-scale impact and news with large-scale impact. The impact scale differs from the geographic scale, as events with a small geographic scale may have a large-scale impact. Much of the M2/C2 content reported in the US herein does not necessarily relate to large protests but to local events from which larger associations could be made—for example, an individual's act of violence presented as epitomizing the whole BLM movement. While an event's geographic scale was often found to have little relevance in assessing its potential impact, a small-scale impact was usually associated with a small geographic scale. As the reporting of such content also seems to proceed from a different intention, we distinguished content with small-scale impact (C0) from that with large-scale impact (C1, C2, C3).

The third difficulty is semantic, grounded in the subtle distinction between assertion and suggestion. Many items do not explicitly endorse a controversial claim but instead suggest it in various ways. It is even more complicated to assess this content when it includes emojis or multimodal associations (e.g., on its own, the caption and image are not controversial but their combination is [49]), or refers to a commonly known

idea without a direct reference. Furthermore, although semantically accurate, it may be unreasonable to change a comment's label from "false news" to "opinion" when a user simply adds "I think that" at the beginning of their posts. The several rounds of reviews allowed us to develop a general understanding of the top viral topics and reclassify the items whose suggestive references had previously been missed. The RBI assumption also allowed us to classify items with a suggested claim as belonging to C1 or C2 according to the obviousness of the reference and the level of controversy.

The fourth difficulty is metacognitive, resulting from a double asymmetry. The first asymmetry is that between the actual intention IA(A) of user A when posting post PA and the given intention IB(A) inferred by user B of user A when seeing PA. This is particularly the case when A makes a metaphorical use of statistics (e.g., "99% of people recover from COVID-19," suggesting that a large majority of people recover, which is accurate even though the exact statistic may not be) or hyperbole (e.g., "everybody recovers from COVID-19," suggesting that most people do). The second asymmetry affects A's expected reception of PA by B and B's actual reception. This especially applies to humorous posts, understood as such by some people but taken seriously by others. This pitfall is central, as it places an agent's subjectivity in tension with that of others; while we aimed to retain the RBI assumption, some reported items of content clearly expressed irony. However, many humorous posts also contained a subclaim that was often controversial, making humorous posts a difficult-to-moderate vector for hoax dissemination. To satisfy the diversity of cases, we classified content using humor into three subclasses according to its obviousness and potential to offend. Many items also contained a mixture of opinions and fact-checkable news or combined inappropriate elements and controversial claims. We classified the former as C instead of O because mixed posts remain relevant from the perspective of misinformation reporting, and the latter as C instead of M, aligned with the RBI assumption.

Finally, the RBI assumption was found to be a valuable asset that preserved the labeling process against reviewers' personal opinions. In such a context of great uncertainty, this assumption removed the temptation for us to classify posts as I when they contained claims that seemed obviously accurate, were debunked later, or were obviously false but whose veracity was ultimately confirmed. Based on the content's nature, GCRC classes are sufficiently objective to be robust to the plurality of opinions that reviewers may have, while classification into subclasses is more strongly influenced by reviewers' opinions. This two-level classification thus combines the advantages of a highly consensual labeling process at the class level and the integration of meaningful additional signals that are, however, less univocal at the subclass level. By comparison with fact-checkers' ratings, we found that 97% of confirmed false news items in S1 were rated C. The other 3% related to O1 content, for which we disagreed with these ratings.

Appendix B: Six manipulation techniques to convey misinformation

The first two strategies target people with a certain degree of skepticism.

1. **The revelation** technique mostly characterizes typical conspiracy theories (C1). It challenges people's egos and encourages them to "wake up" instead of being "sheep," making direct references to the cabal, Masons, a world elite, and a new order. These posts are usually marked by semantic patterns related either to the general concepts of truth, trust, the elite, and the establishment (e.g., "news," "media," "wake up," "masons," "governments," "truth," "sheep,"

"facts," "distracting," or to more contextual theories such as QAnon or anti-vax conspiracies (e.g., "pizza," "pedophile," "Hollywood," "children," "Clinton," "Trump," "Bill Gates," "chip," "Facebook policy") and intend to gain virality by soliciting viewers to reshare ("spread the word," "share before it gets deleted," "share to expose them"). Congruent with other research (Rashkin et al. 2017), such findings also confirm that semantic cues can constitute a useful signal for detecting misinformation, notably by monitoring the frequency of words contained in posts verified as hoaxes by fact-checkers. This signal, however, might mostly be useful for detecting the most caricatured conspiracies on Facebook (C1), such as hoaxes that can be debunked with a quick web search. Moreover, the people most susceptible to these messages are likely those already in contact with conspiracy theories, and they may not be sensitive to debunking material—being a conspiracist is less a case of being given inaccurate information than a psychological posture. Semantic methods may also be challenged by counterdetection strategies (e.g., "c0r0navirus").

2. **The critical tipping point** technique consists of leveraging a real fact (e.g., a public scandal or polemical claims from a controversial figure) as an entry point to encourage people to reconsider their beliefs. The fact is usually presented in a twisted way, often accompanied by exaggerated empathy ("this is despicable") or an invitation to generalize ("if they lied about this, what else did they lie about?"). A variant involves creating a false mystery around an accurate fact ("this is happening, why is nobody talking about it?"). This technique may be most effective in encouraging people who are already skeptical or indignant about a recent scandal to start accepting conspiracies.

The following two techniques target people who are sensitive to pragmatic arguments.

3. **False facts supported by false evidence** are typically false statements about events that have allegedly occurred and false quotes from public figures, often backed by inauthentic documents (e.g., inauthentic "leaked" documents from the FBI, CDC, or BLM management), unreferenced scientific data, or personal testimonies from a mysterious authority ("a friend working at the NHS," "the head of the resuscitation department of this hospital") whose source is impossible to verify. This technique also includes modified pictures and videos, but we did not find any deepfakes.

4. **The misleading presentation of facts** involves presenting authentic documents or accurate facts in a misleading way to encourage a targeted erroneous interpretation (e.g., quotes and pictures taken out of their original context, truncated videos, partial references to history). While similar to the previous technique, this technique is more difficult to debunk because of the accuracy of the facts it is based on.

The fifth technique plays on people's empathy.

5. **The confusion of feelings** presents either false news or authentic facts in a twisted way to provoke an emotional reaction against a given target. This was particularly observed when someone leveraged an isolated action to discredit a whole movement through the mobilization of symbols (e.g., protester's violence against a veteran, acts of police brutality against a peaceful protester, profanation of a military cemetery, destruction of the statue of a public figure).

The sixth technique plays on the register of *coolness*.

6. **The excuse of casualness** characterizes items with humor (jokes, ironic state-

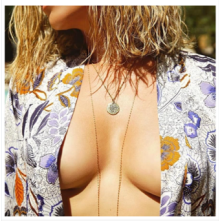
ments, memes, caricatures) and/or an artistic dimension (cartoons, short format videos, songs), making an implicit reference and reacting to a subclaim using a frivolous tone. This content differs from both "parodies" (Wardle 2016) and "satires," as it is not "meant to be perceived as unrealistic" (Moullot and Halissat 2020, 198) but rather hides an assumed claim behind a veil of frivolity.

Appendix C: Selected examples of reported content from S1

[1]

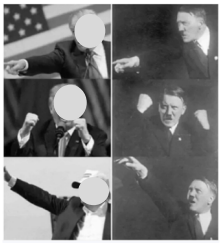
[2]

[3]



[4]

[5]



[6]

[7]

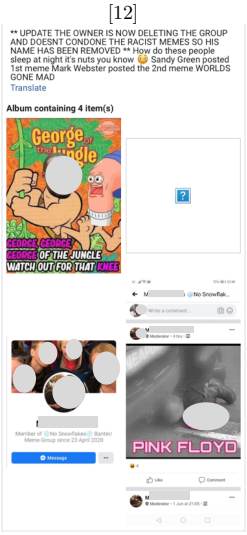
[8]

[9]

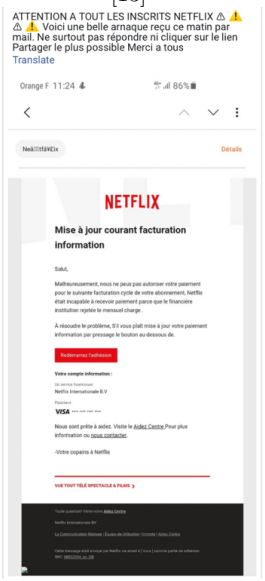
[10]

[11]

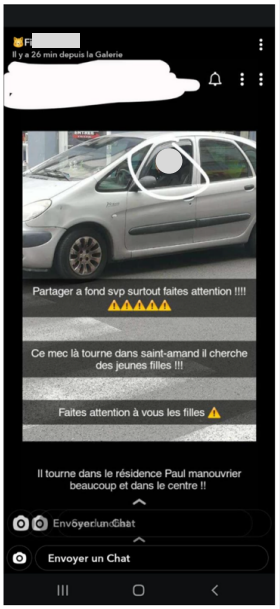
[12]



[13]

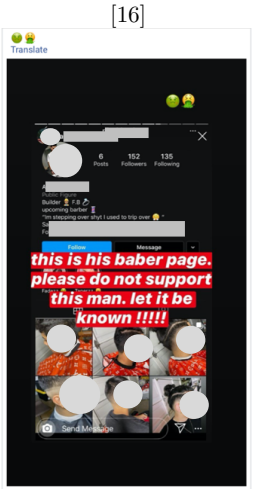


[14]

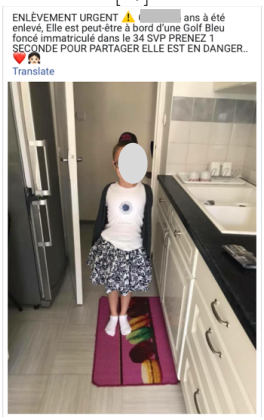


[15]

[16]



[17]



[18]

[19]

[20]

[21]

[22]

[23]

[24]

[25]

[26]

[27]

[28]

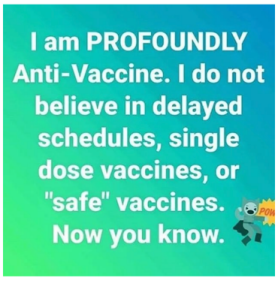
[29]

[30]

[31]

[32]

[33]

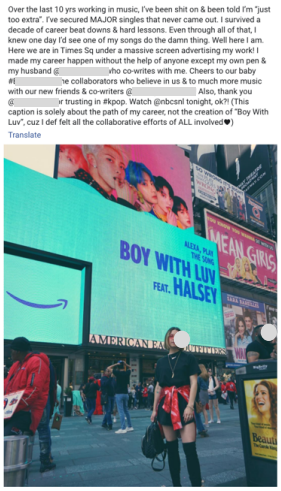


[34]

[35]

[36]

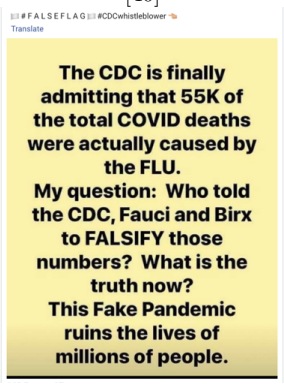
[37]



[38]

[39]

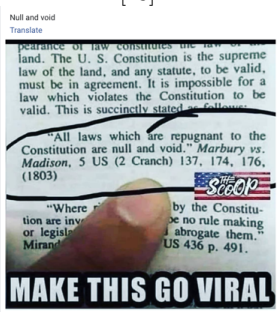
[40]



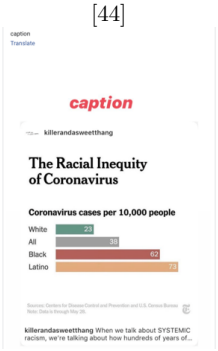
[41]

[42]

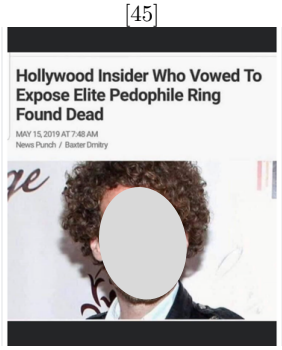
[43]



[44]



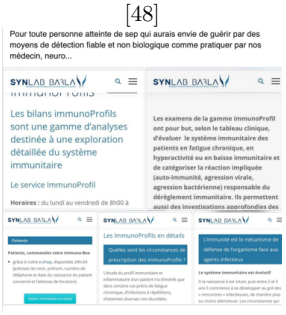
[45]



[46]

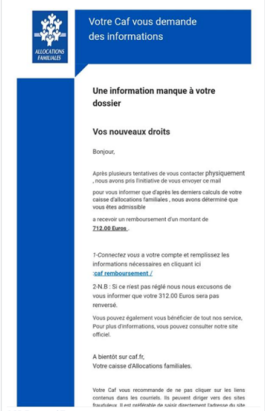
[47]

[48]



[49]

[50]



[51]

[52]