(Dis)information for Hire?

Dispelling Exaggerated Concerns about Social Media Influencers' Role in Spreading Misinformation

K. Jonathan Klüser

klueser@ipz.uzh.ch

Department of Political Science, University of Zürich, Switzerland

Emma Hoes

hoes@ipz.uzh.ch

Department of Political Science, University of Zürich, Switzerland

July 2024

Keywords: Misinformation, Digital Media, Social Media Influencers, Information Ecosystem

Abstract

Amidst rampant concerns about disinformation, social media influencers (SMIs) can capitalize on their often enormous outreach to spread false claims among their followers. However, despite their sizable potential, the extent to which SMIs sow discord and endorse false narratives is uncharted territory. In this paper, we explore the scale at which SMIs engage with misinformation. We begin by gathering posts from English-speaking influencers with over 500,000 followers on Instagram using CrowdTangle. We then identify instances of disputed content by (i) cross-referencing posts with verified false claims from Politifact, and (ii) manual fact-checking of a random sample of 1000 political posts. This research is pioneering in providing empirical evidence on SMIs' participation in spreading falsehoods. Yet, we find that the concerns are exaggerated, as the involvement of SMIs in propagating false claims is minimal, with only 0.003% of the more than 1.3 million posts analyzed actually supporting statements flagged as disputed by Politifact.

1 Introduction

It is well known that disinformation often comes from the top (Bovet and Makse 2019), but there is a 'new' group of political elites — social media influencers (SMIs, cf. Suuronen et al. 2021) — that has not been studied, despite indications they may play a role in the dissemination of falsehoods.[1] Occasionally, these accounts have, at least inadvertently, used their platforms to trigger misinformation scandals. Recent examples include the provision of inaccurate health or COVID-19 advice (Abidin et al. 2021), which their audience tends to trust (Mena, Barbe, and Chan-Olmsted 2020; Schmuck and Harff 2023). Furthermore, media reporting claimed that there had been a secretive industry at play which 'hired' SMIs to sow discord, meddle in elections and seed false narratives. Recently, the New York Times published an article which exposed a public relation agency that offered payment to SMIs in return for promoting falsehoods on behalf of a client (Fisher 2021). Yet, this has only been the most prominent case of a 'disinformation-for-hire' industry, which, as some claim, has secretively evolved into a booming business, triggering growing and serious concerns.

SMIs can capitalize on their outreach to influence people's perceptions of given issues, which turns them into a useful pawn for actors who strive to foster misperceptions, i.e., make people believe something that is factually inaccurate. There are several reasons to believe that SMIs can have a significant impact on shaping political opinions, whether factually accurate or not. First, popular SMIs have a massive number of followers, which often reaches tens of millions. Second, the power of SMIs lies in how they engage with their followers differently compared to other informational sources such as traditional media. Following influencers on social media is more like having a personal interaction, rather than passively consuming content one encounters while scrolling through a social media feed (Kaskazi and Kitzie 2023). Following the ideas of a "parasocial relationship" (Horton and Richard Wohl 1956), SMIs create a sense of intimacy by sharing carefully selected private content, making their audience feel like they are being addressed di-

[1] This group of large influential accounts can also comprise more traditional forms of celebrities or models. For stylistic reasons, we use the term SMI throughout the paper.

rectly. This closeness builds an emotional bond, and followers perceive influencers as friends rather than distant figures (Ballantine and Martin 2005). Over time, this emotional connection leads to higher levels of trust and a greater acceptance of persuasive messages (Phua 2016). Because of this personal trust, SMIs can have a strong and more subtle influence on their followers' political views.

This, on the one hand, legitimizes worries if SMIs' influence is leveraged to spread misinformation. On the other hand, however, (political) actors and organizations may leverage SMIs' potential to combat rather than spread falsehoods (Lorenz 2021). It thus seems that SMIs' often enormous outreach may bear both positive and negative consequences, if they are indeed 'hired' to spread (political) messages. The extent to which SMI actually endorse pieces of disinformation, however, is unknown.

Most research mapping the sources spreading misinformation online typically focuses on news media (Allen et al. 2020; Brest and Cordonier 2023), Twitter (Grinberg et al. 2019; Osmundsen et al. 2021), and Facebook (Boberg et al. 2020), finding that overall few people share or encounter misinformation (Aslett et al. 2022; Guess, Lockett, et al. 2020; Guess, Nagler, and Tucker 2019; Hoes et al. 2022). Importantly, however, even if relatively few people spread or consume misinformation, these particular few may be especially (politically) active and thus disproportionately influential. Indeed, extant studies point out that only a concentrated narrow subset of the population engages in the dissemination and spread of misinformation (Nyhan 2019).

The most worrisome misinformation therefore is not necessarily that which is widespread, but rather misinformation echoed by actors who dominate people's information consumption — be it news coverage, such as elected officials, or indeed social media feeds, such as SMIs. For instance, Donald Trump's single Tweet — which was covered by the media almost instantaneously — stating that the 2020 election was rigged allegedly mobilized his supporters and led to what we now know as the Capitol Riots on January 6th, 2021. Similarly, celebrity and SMI Olivia Rodrigo — with back then over 30 million followers on Instagram — posted about her visit to the White

House in order to push the COVID-19 vaccine, which resulted in more than 5 million likes.[2] Both instances showcase how few but visible actors have can have a significant impact, the latter of which is an example of combating and the former one of spreading misinformation.

Remarkably, despite SMIs' capacity to influence political discourse and media concerns about a supposed 'disinformation-for-hire' industry, there has been a lack of research on the prevalence of misinformation on platforms heavily used by SMIs, like Instagram and TikTok. In our study, we quantify the extent to which SMIs either propagate or counteract misinformation, with a particular focus on Instagram. Employing a blend of semantic search techniques and manual analysis, we discovered that only 0.65% of SMIs in our sample engaged with verified false claims from the fact-checking organization, Politifact. These posts represent a scant 0.005% of all 1.3 million posts in our data. Furthermore, within this small subset, just 60% were true endorsements of false claims, whereas the remainder effectively challenged and corrected misinformation. Extrapolating from a manually fact-checked random sample, we infer that 1.3% of the *political posts* that SMI publish on Instagram contain a false claim, which corresponds to 0.03% of the entire content we observe.

Our findings therefore show that the vast majority of SMIs does not engage with (political) misinformation. While this finding does not alleviate every concern about misinformation, it does counter alarmist narratives in the media (Altay, Berriche, and Acerbi 2021; Hoes et al. 2022) suggesting that 'disinformation-for-hire' is at large.

2 Results

SMIs rarely endorse verified false claims

Contextualizing common societal and political apprehensions, our findings strongly indicate that SMIs typically refrain from using their platforms to disseminate false claims. By employing a supervised semantic search strategy, we only found 67 Instagram posts that directly addressed a

[2]For the Instagram post, see `https://www.instagram.com/p/CRU9pq4rEKj/?hl=de`

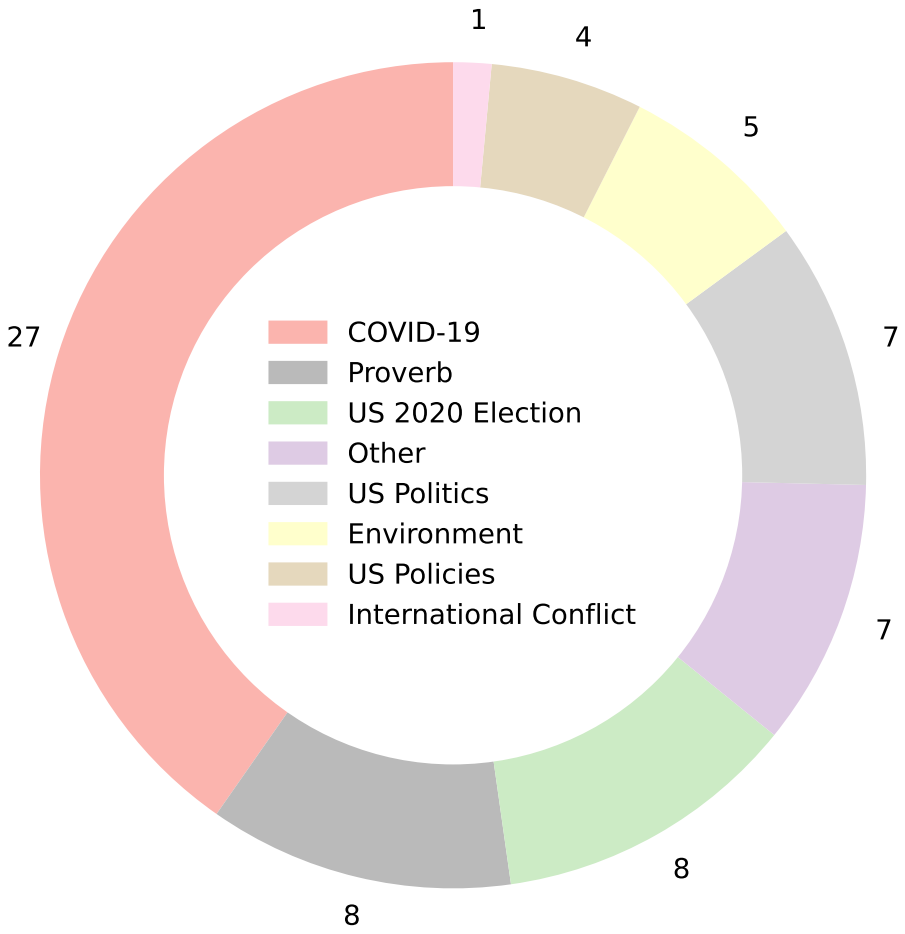

Figure 1: Distribution of engagement with false claims over categories. Categories are based on manual classification. Non-political categories are colored in gray.

verified piece of false information. These posts represent only 0.005% of the 1.3 million posts in our dataset. Moreover, only 60% of these identified posts actually endorse the false statements, while in the remaining instances SMIs have utilized their voice to dispel them.

To better understand the nature of these reiterations of false claims, we manually clustered the original false claims that SMI addressed in their posts. Given the timeframe of our study comprises the year 2020, a large share (40%) of the endorsements pertains to COVID-19, vaccines, or the pandemic in general. Other prominent political topics include the US 2020 election and US politics (11% each), which include unflattering caricatures of both presidential candidates, and alleged foreign and domestic interference in US political affairs. Specific policy topics such as the

environment or domestic US policy decisions are both only endorsed in the low single-digits. Remarkably, the second-most endorsed category of false claims are proverbs that either are wrongly used or lack empirical foundation. For instance, in September 2019, an Instagram account impersonating the US comedian Jim Carrey claimed that the Roman emperor Marcus Aurelius once said "Everything we hear is an opinion, not a fact. Everything we see is a perspective, not the truth." Although there is no historical evidence for this quote, it was later reiterated and endorsed by SMI accounts in our data (cf. Figure 3).

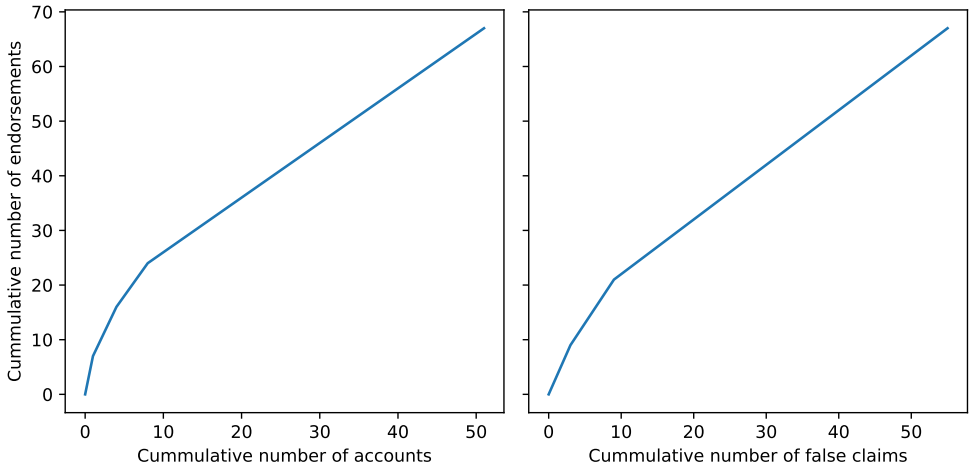

Figure 2: Cumulative distributions of endorsements of false claims. The left plot shows the cumulative number of endorsements versus the cumulative number of Instagram accounts that posted them. The right plot shows the cumulative number of endorsements versus the cumulative number of verified false statements from the fact-checking website Politifact. Numbers are absolute.

Endorsements are rather evenly distributed

In terms of the distribution of false claims across accounts, the 67 total endorsements or debunks of false claims were posted by 51 different SMI accounts (Figure 2). Some accounts were slightly more active in endorsing false claims than the average, with four accounts posting three or more pieces of disinformation over the one-year period of this study. However, more commonly, SMIs shared only one false claim. Moreover, approximately 10 false claims constitute 20% of the endorsements posted by SMIs, whereas the remaining false claims have only been endorsed once.

Figure 3 illustrates the temporal distribution of endorsements and debunks of false claims. The x-axis represents months, while the y-axis indicates days within each month. Each square corresponds to one day, with the color of the square denoting the number of endorsements posted on that day. Dark-colored squares denote one shared false claim per day, whereas bright-colored squares mark days when two false claims were reiterated. The graph indicates a relatively uniform distribution of false claim endorsements throughout 2020, with a notable concentration of activity in the early months of the pandemic. Despite this, most of the posts were isolated incidents occurring on individual days. Examining specific examples, we observe that on May 7, 2020, fashion blogger Zaklina Pisano propagated misinformation about the COVID-19 pandemic. In her post, she endorsed the incorrect assertion that sunlight acts as a deterrent against the coronavirus and encouraged her followers to rally behind her to prevent being "shut down" by social media platforms:

> US acting homeland security under-secretary for science and technology, William Bryan, talked about the effects of sunlight, temperature, humidity and bleach on the coronavirus that is removing virus in seconds‼ Prof Paul Hunter, Professor in Medicine, UEA, said: That UV light inactivates SARS-CoV-2 is not surprising. UV inactivates most viruses very efficiently. Indeed UV disinfection is widely used for disinfection of drinking water. Take vitamin D regularly and you'll take your power back (...) Thank you for your private messages of support. But, you see, we are less and less on the social platforms, that try to give back power to the people... You tube, google, facebook, Instagram are shutting down our accounts... and if you don't give us support OPENLY ‼! soon will be no one left, that will speak in your name

Regarding the refutations of false claims, their temporal distribution mirrors that of the endorsements, with debunks evenly dispersed throughout the year and typically occurring as single events on distinct days. An exemplary case is a post by Niecy Nash, a US actress, who uses her platform to counter a false narrative. Specifically, she addresses the misinformation surrounding 17-year-old Kyle Rittenhouse, accused of shooting three individuals during a protest in Kenosha, Wisconsin, by clarifying the inaccuracy of claims that he was legally permitted to carry a rifle under state law. She articulates this correction as follows:

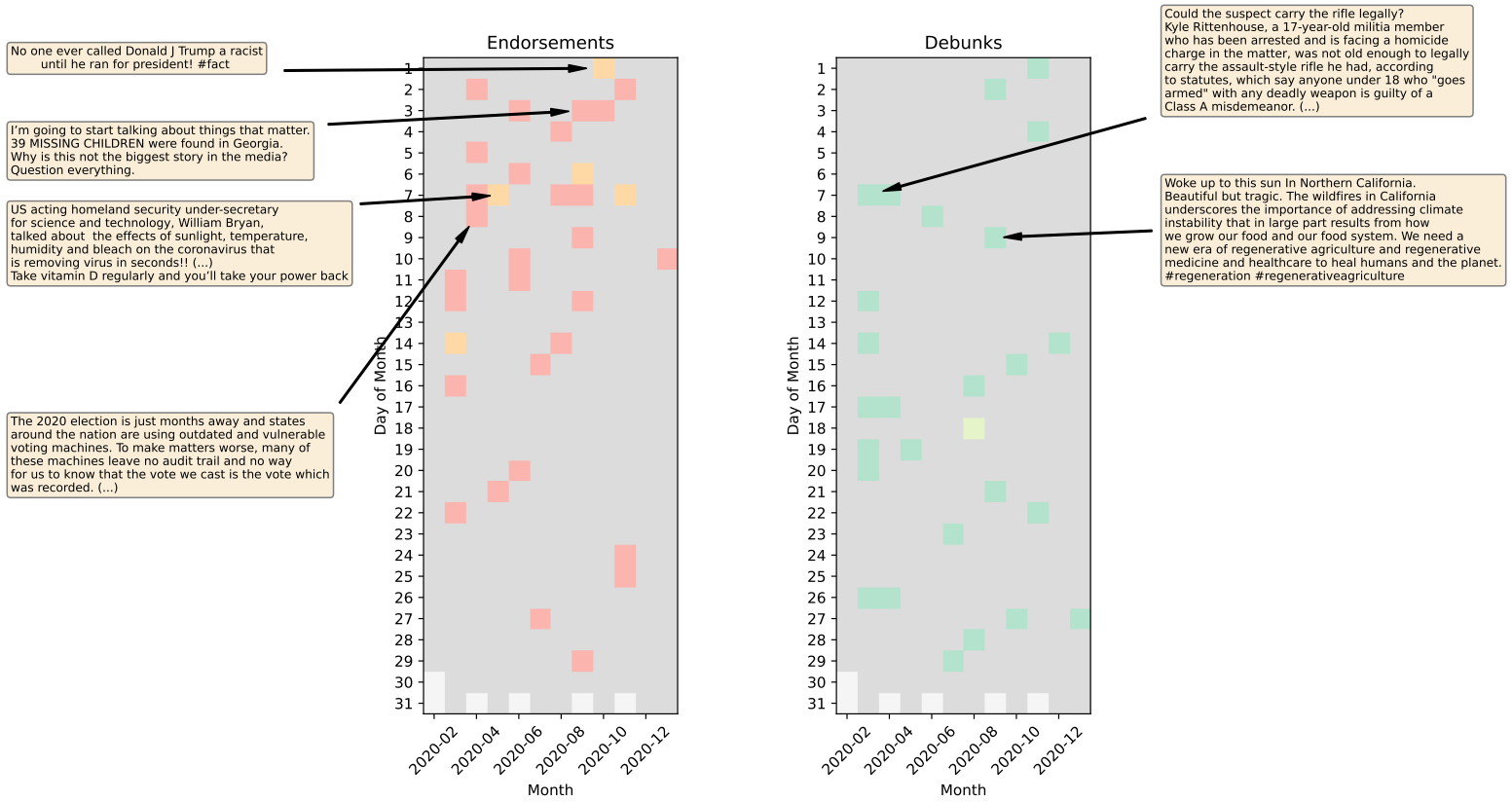

Figure 3: Temporal distribution of endorsements of false claims. The x-axis shows months, whereas the y-axis indicates days within each month. Hence, each square represents one day. The color of the square indicates the number of endorsements posted on that day.

> Could the suspect carry the rifle legally? Kyle Rittenhouse, a 17-year-old militia member who has been arrested and is facing a homicide charge in the matter, was not old enough to legally carry the assault-style rifle he had, according to statutes, which says anyone under 18 who "goes armed" with any deadly weapon is guilty of a Class A misdemeanor. John Monroe, a lawyer who specializes in gun rights cases, believes an exception for rifles and shotguns, intended to allow people age 16 and 17 to hunt, could apply.
>
> He could be in violation of having a gun within a gun-free zone, if there was, for instance, a school nearby. Also, Illinois law requires anyone who owns any kind of firearm in that state to have a Firearm Owners Identification card, but that is only available to someone 21 or older, or someone with a sponsor who is 21 and eligible for a card.

Endorsements perform very similar to regular posts

Lastly, we examine how endorsements of false claims perform in comparison to regular Instagram posts. Previous research indicated that SMI and celebrity accounts are generally wary to engage with political topics as they fear the potential repercussions for their online persona (Suuronen et al. 2021; Ki et al. 2020). Hence, we expect that endorsements of false claims affect both low- and high-cost user engagement by eliciting fewer likes and prompting less-positive comments compared to regular posts by the same account. Regarding low-cost engagement, Figure 4 shows that there is no statistical significant difference between endorsements of false claims and other posts in terms of received likes on Instagram. Comparing the scaled number of likes, i.e., subtracting the account mean and dividing by the account's standard deviation, reveals that both types of posts essentially receive the same number of likes on Instagram.

Manual fact-checking of political posts finds slightly more false claims

The previous results relied on cross-referencing Instagram posts with fact-checked statements from Politifact. This approach has a notable downside: the set of false claims we use for cross-referencing is necessarily incomplete, as no fact-checking organization can feasibly verify all false claims that exist. Hence, we do not know how much 'novel' disinformation SMIs produce and spread on Instagram. To partially address this limitation, we manually fact-checked a random

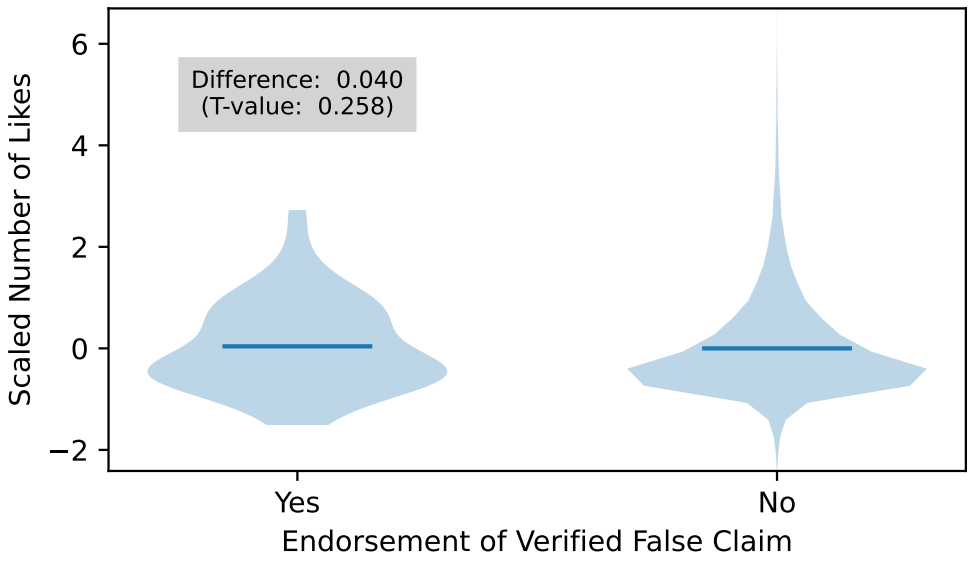

Figure 4: Performance of endorsements of false claims vs. regular Instagram posts. The y-axis shows the scaled number of likes per post, i.e., subtracting the account mean and dividing by the account's standard deviation.

sample of 1000 political claims.

Political posts were defined as any posts that speak to a policy topic as defined by the Comparative Agendas Project (CAP; Baumgartner, Breunig, and Grossman 2019). Our classification approach involved a two-step process. First, we used a classifier to identify posts that appeal to any civic or political topic, broadly conceived, achieving an F1 score of 0.84. We then refined the classification using a pre-trained classifier trained on data from the CAP project (F1 (weighted): 0.82; Sebők et al. 2024). Posts not aligned with any CAP categories were removed, resulting in 37,000 posts classified as political, which is 2.9% of the entire sample. See SI B for more information on the classification process.

Manual fact-checking found 13 political posts in the random sample that contained a false claim. This indicates that 1.3% of political posts contain a false claim, which translates to 0.03% of the entire content observed. Thus, the overwhelming minority of SMI's political communication on Instagram contains verified misinformation.

3 Discussion

Our study provides the first thorough examination of the extent to which social media influencers (SMIs) and celebrities, conventionally apolitical accounts, participate in spreading false claims on Instagram. We discovered that these accounts seldom utilize their reach to propagate falsehoods, with a mere 0.005% of the 1.3 million posts in our dataset constituting clear, manually verified instances of claims labeled false by the fact-checking organization Politifact. Relying on manual fact-checking instead of cross-referencing posts with information from Politifact, we find that 1.3% of all *political content* contains a false claim, which corresponds to 0.03% of the entire content we observe. Additionally, our findings suggest that, often, SMIs actively combat misinformation by addressing false claims and disseminating accurate information. Furthermore, we demonstrate that users generally disapprove of SMIs or celebrities endorsing false statements, as posts containing such endorsements tend to attract a higher volume of negative comments.

Our study, while providing valuable insights into the role of social media influencers (SMIs) in the dissemination of political content and false claims, has a central limitation that should be acknowledged. Our study solely relies on large accounts, specifically those with over 500,000 followers. While these accounts bear the largest potential for harm if they were to engage in the spread of false claims, they may not be the most likely SMIs accounts to endorse false claims. Beyond a certain size, SMI are likely to make their living based on their Instagram presence, which means they are very wary of posting content that does not align with their 'authentic' brand or content that might alienate or offend their followers (Ki et al. 2020). In fact, abusive comments by followers are particularly common in settings that include political topics (Cicchirillo, Hmielowski, and Hutchens 2015) and fora, which are generally geared towards the discussion of leisure topics, are the most likely to spur political disagreement (Wojcieszak and Mutz 2009). Hence, large, brand-wary professional SMIs might effectively be deterred to posts debatable political content. However, given CrowdTangle's API limitations and the observation that the number of accounts decreases logarithmically with the number of followers — meaning that the number of accounts increases exponentially with smaller follower numbers — a full analysis of smaller

accounts has been infeasible. Nevertheless, considering that public fears referenced at the beginning of this study explicitly focus on large SMI accounts, we believe this research is well-suited to put the fears of an "influencer army" into perspective (Fisher 2021).

We urge future work to build upon our initial findings and provide a more comprehensive understanding of the role of SMIs in the dissemination of political content and false claims. Our current analysis relies solely on Politifact as a source of verified false claims. We opted for this approach since Instagram, unlike Facebook or X, does not allow posting any URLs, which in most studies looking at the dissemination of misinformation on social media are used as an indicator of untrustworthy websites (e.g., Boberg et al. 2020; Grinberg et al. 2019; Osmundsen et al. 2021). Still, to obtain a more comprehensive understanding of the extent to which SMIs engage in the spread of false claims, expanding the sourcing to include other reputable fact-checking websites is recommendable to obtain a fuller picture of the extent to which SMIs engage in the spread of misinformation.

In light of the observation that SMIs also use their voice to refute false claims, future work should look into the effectiveness with which SMIs can use their outreach to confront false claims with factual information in an effort to combat misinformation. Exploring this line of inquiry could reveal the extent to which SMIs contribute positively to the information ecosystem. Such a scenario could thus suggest that (political) actors and organizations may leverage SMIs' potential to combat rather than spread falsehoods Lorenz 2021, which may be especially valuable in reaching social-media users in information-poor environments (cf. Schmuck and Harff 2023).

Lastly, more in-depth inquiries should be made in order to understand to what extent — if at all — SMIs are actually contacted by (political) organizations to spread certain messages. This would allow making more meaningful claims about an alleged disinformation- or information-for-hire industry, the former of which at least our computational exploration finds little proof, since there does not seem to be much endorsed disinformation to begin with.

Methods

Collection of Instagram posts

In this study, we used the data provider CrowdTangle to collect all posts made by Social Media Influencers and celebrities on Instagram between February 2020 and February 2021. To make data collection manageable, we focused on predominantly English-speaking accounts with a minimum of 500,000 followers. The data collection process was executed in several stages, as explained below.

First, we compiled a list of all accounts that had more than 500,000 followers at any time during our study's timeframe from the data provider CrowdTangle. More specifically, we repeatedly queried the "leaderboard" of the best- and worst-performing 1000 accounts until our query did not return any additional accounts. This resulted in a total of 40,100 unique Instagram accounts that have been active during the timeframe of this study.

Besides generally apolitical social elites, such as SMIs or traditional celebrities, the set of accounts also contains other account categories, such as journalists, corporations, or politicians. In order to filter out these types of accounts, we leveraged recent advancements in generative AI, particularly ChatGPT. We capitalized on the "memorization" capability of deep generative models, which allows them to recall portions of the input data encountered during training (Burg and Williams 2021). Since our data collection period precedes ChatGPT's knowledge cut-off (September 2021), this approach is technically feasible. More precisely, we prompted ChatGPT to return a brief summary of the topics an Instagram conventionally posts about, based on the full name and the Instagram user handle of an account. Based on this brief textual summary, we used contrastive learning to train a classifier that reliably distinguished between SMI and traditional celebrity accounts on the one hand, and other large Instagram accounts on the other hand. The entire classification process – summary creation and subsequent classification – achieves a F1 -score of 0.81 (accuracy: 0.88). Please refer to SI A for additional information on the account classification process. This filtering process reduced the set of analyzed Instagram accounts to

about 9,400.

Using these accounts as a foundation, we collected all posts from the data provider CrowdTangle, leveraging their API access to gather posts programmatically. Our focus on textual content necessitated filtering the posts to include only those with adequate text. Consequently, we established a minimum textual requirement of at least 20 characters, encompassing both image text and captions. Posts with less than 20 characters were excluded to ensure sufficient content for the subsequent analysis. Moreover, as our database of disputed claims predominantly contains claims in English, we subset our data to posts from predominantly English-speaking accounts. Although CrowdTangle supplies information about the language of the caption accompanying an Instagram post, we discovered that this data was unreliable. To resolve this issue, we enriched our dataset with language information obtained from Google's language-detection API.[3] We then eliminated all posts with captions not written in English.

Following this filtering process, we obtained approximately 1.3 million Instagram posts created by SMIs and traditional celebrities between February 2020 and February 2021. These posts serve as the basis for our study.

Identification of Endorsements of False Claims

In this section, we describe the methodology employed to ascertain whether the caption of an Instagram post directly addresses a claim that has been verified as being false. Our approach to identifying endorsements of false claims involves a comprehensive semantic search for verified disinformation within our dataset of Instagram posts. We obtained these false claims from the fact-checking website Politifact, spanning the period from January 2018 to January 2021. We retained only claims categorized as "false" or "mostly false," resulting in a set of 2600 false claims.[4]

To detect instances in which SMIs or celebrities disseminate misinformation within our filtered dataset, we encoded all textual information, encompassing both social media posts and false claims, in the same 768-dimensional vector space. This was achieved through the use of the

[3] We used the Python library "langdetect" (`https://pypi.org/project/langdetect/` to call the API.
[4] `https://www.politifact.com`

"sentence-transformers" Python library, which is tailored for creating sentence embeddings with language-transformer models. Specifically, we employed the *paraphrase-multilingual-mpnet-base-v2* model, which is optimized for tasks such as clustering or semantic search (Reimers and Gurevych 2019).

Considering the size of our Instagram dataset, we opted for an efficient semantic search approach and used the captions' embeddings to construct a FAISS (Facebook AI Similarity Search) index. A FAISS index is a data structure that facilitates efficient similarity search and clustering of dense vectors (Johnson, Douze, and Jégou 2017). Utilizing this FAISS index, we conducted a search of all false claims against our data of verified false claims. As a result, we performed a total of 3.4 billion comparisons between individual verified false claims and all Instagram posts in our dataset. We retained those comparisons that were semantically close, with a squared L2 distance of less than 5. This process yielded not more than 1300 close comparisons. Given the small number of potential addresses of false claims, we manually verified both all accounts who had authored these posts and the posts themselves. Regarding the former, we identified two non-Western news accounts that had falsely been classified as SMI by our classification approach, which jointly had been responsible for about half of the close comparisons.[5] Accordingly, we purged them from our results. The final set of close comparisons was manually verified by research assistants, who were tasked with deciding, for each close comparison, whether the Instagram caption constitutes (a) an endorsement of a verified false claim, (b) a refutation of a verified false claim, or (c) is topically similar, but does not directly address the false claim.

[5]These were `https://www.instagram.com/dawn.today/` and `https://www.instagram.com/lindaikejiblogofficial/`

References

Abidin, Crystal et al. (2021). "Influencers and COVID-19: Reviewing Key Issues in Press Coverage across Australia, China, Japan, and South Korea." In: *Media International Australia* 178.1, pp. 114–135.

Allen, Jennifer et al. (2020). "Evaluating the fake news problem at the scale of the information ecosystem." In: *Science Advances* 6.14, eaay3539.

Altay, Sacha, Manon Berriche, and Alberto Acerbi (2021). "Misinformation on misinformation: Conceptual and methodological challenges." In.

Aslett, Kevin et al. (2022). "News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions." In: *Science advances* 8.18, eabl3844.

Ballantine, Paul W and Brett AS Martin (2005). "Forming parasocial relationships in online communities." In: *ACR North American Advances*.

Baumgartner, Frank R., Christian Breunig, and Emiliano Grossman, eds. (2019). *Comparative Policy Agendas: Theory, Tools, Data*. 1st ed. Oxford; New York, NY: Oxford University Press. 405 pp.

Boberg, Svenja et al. (2020). "Pandemic populism: Facebook pages of alternative news media and the corona crisis–A computational content analysis." In: *arXiv preprint arXiv:2004.02566*.

Bovet, Alexandre and Hernán A Makse (2019). "Influence of fake news in Twitter during the 2016 US presidential election." In: *Nature communications* 10.1, p. 7.

Brest, Aurélien and Laurent Cordonier (2023). "Does Exposure to Online News Media Depend on Individuals' Political Attitudes and Trust in These Media? A Comparison Between Declarative and Behavioral Data." In: *Mass Communication and Society*, pp. 1–30.

Burg, Gerrit J. J. van den and Christopher K. I. Williams (2021). *On Memorization in Probabilistic Deep Generative Models*. arXiv: 2106.03216 [cs, stat]. Pre-published.

Cicchirillo, Vincent, Jay Hmielowski, and Myiah Hutchens (2015). "The Mainstreaming of Verbally Aggressive Online Political Behaviors." In: *Cyberpsychology, Behavior, and Social Networking* 18.5, pp. 253–259.

Fisher, Max (2021). "Disinformation for Hire, a Shadow Industry, Is Quietly Booming." In: *New York Times*.

Grinberg, Nir et al. (2019). "Fake news on Twitter during the 2016 US presidential election." In: *Science* 363.6425, pp. 374–378.

Guess, Andrew, Dominique Lockett, et al. (2020). ""Fake news" may have limited effects beyond increasing beliefs in false claims." In: *Harvard Kennedy School Misinformation Review* 1.1.

Guess, Andrew, Jonathan Nagler, and Joshua Tucker (2019). "Less than you think: Prevalence and predictors of fake news dissemination on Facebook." In: *Science advances* 5.1, pp. 45–86.

Hoes, Emma et al. (2022). "The Cure Worse Than the Disease?" In.

Horton, Donald and R Richard Wohl (1956). "Mass communication and para-social interaction: Observations on intimacy at a distance." In: *psychiatry* 19.3, pp. 215–229.

Johnson, Jeff, Matthijs Douze, and Hervé Jégou (2017). *Billion-Scale Similarity Search with GPUs*. arXiv: 1702.08734 [cs]. Pre-published.

Kaskazi, Amana and Vanessa Kitzie (2023). "Engagement at the margins: Investigating how marginalized teens use digital media for political participation." In: *New Media & Society* 25.1, pp. 72–94.

Ki, Chung-Wha (Chloe) et al. (2020). "Influencer Marketing: Social Media Influencers as Human Brands Attaching to Followers and Yielding Positive Marketing Results by Fulfilling Needs." In: *Journal of Retailing and Consumer Services* 55, p. 102133.

Lorenz, Taylor (2021). "To Fight Vaccine Lies, Authorities Recruit an 'Influencer Army'." In: *The New York Times*.

Mena, Paul, Danielle Barbe, and Sylvia Chan-Olmsted (2020). "Misinformation on Instagram: The Impact of Trusted Endorsements on Message Credibility." In: *Social Media + Society* 6.2, p. 2056305120935102.

Nyhan, Brendan (2019). "Why Fears of Fake News Are Overhyped." In: *Medium*.

Osmundsen, Mathias et al. (2021). "Partisan polarization is the primary psychological motivation behind political fake news sharing on Twitter." In: *American Political Science Review* 115.3, pp. 999–1015.

Phua, Joe (2016). "The effects of similarity, parasocial identification, and source credibility in obesity public service announcements on diet and exercise self-efficacy." In: *Journal of health psychology* 21.5, pp. 699–708.

Reimers, Nils and Iryna Gurevych (2019). "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. EMNLP-IJCNLP 2019. Hong Kong, China: Association for Computational Linguistics, pp. 3982–3992.

Schmuck, Desiree and Darian Harff (2023). "Popular Among Distrustful Youth? Social Media Influencers' Communication About COVID-19 and Young People's Risk Perceptions and Vaccination Intentions." In: *Health Communication* 0.0, pp. 1–14. pmid: 38099315.

Sebők, Miklós et al. (2024). "Leveraging Open Large Language Models for Multilingual Policy Topic Classification: The Babel Machine Approach." In: *Social Science Computer Review*, p. 08944393241259434.

Suuronen, Aleksi et al. (2021). "When Social Media Influencers Go Political : An Exploratory Analysis on the Emergence of Political Topics Among Finnish Influencers." In: *Javnost-the Public*, pp. 1–17.

Tunstall, Lewis et al. (2022). *Efficient Few-Shot Learning Without Prompts*. arXiv: 2209.11055 [cs]. Pre-published.

Wojcieszak, Magdalena E. and Diana C. Mutz (2009). "Online Groups and Political Discourse: Do Online Discussion Spaces Facilitate Exposure to Political Disagreement?" In: *Journal of Communication* 59.1, pp. 40–56.

A Account Classification

A.1 ChatGPT Prompt

```
{"role": "system",
"content": "You will be given one user name from an Instagram account.
    ↪ Your task is to provide a short (about 50 words) summary of what
    ↪ the account is about. This includes information on what they
    ↪ usually post about and what the account is known for. Also name
    ↪ the country where the account is based in.

Format your answer as follows:
{summary: SUMMARY, location: LOCATION}

If you do not know the account, answer "NA"

Your answer must be formatted as valid JSON"}
```

A.2 Contrastive Learning Classifier

Based on the generated account summaries, we used supervised learning to classify Instagram accounts. We applied contrastive learning techniques, specifically using a few-shot classifier, to differentiate between Social Media Influencers (SMI)/celebrities and other types of accounts. The core of our classifier was the SetFit method, which was implemented using the sentence-transformers/all-mpnet-base-v2 transformer model.

Our training dataset consisted of 135 summaries of Instagram accounts, as described by the ChatGPT prompt. We configured the classifier with a batch size of 16 and trained for 3 epochs. This approach proved effective, as evidenced by the classifier's performance on the test set. The classifier achieved an accuracy of 0.879 and an F1 score of 0.811, indicating a high level of precision

and recall in classifying the Instagram accounts into the correct categories based on the entire classification process.

B Classification of Political Content

The identification of political content among sourced Instagram posts involves a two-stage process. First, a few-shot learning technique using SetFit was utilized to build a binary classifier capable of identifying posts referencing civic, economic, or political topics (Tunstall et al. 2022). *paraphrase-multilingual-mpnet-base-v2* was chosen as the underlying language model. Trained on 500 examples, this classifier achieved acceptable performance (F1 (weighted): 0.84), serving as an initial filter for broadly political content. For a more precise definition of political content, the Comparative Agenda Project's (CAP) coding scheme was applied. This second step used language-specific classifiers, fine-tuned on CAP training data, to identify posts associated with any of the 20 policy topics (Sebők et al. 2024). The classifier achieved an F1 score of 0.82 (cf. Table 1), which is consistent with the performance reported by the developers.

Category	F1 Score
Weighted	0.8189
Macroeconomics	0.4000
Civil Rights	0.8571
Health	0.8605
Agriculture	0.7500
Education	0.7778
Environment	0.9286
Energy	0.7500
Immigration	1.0000
Transportation	0.6667
Law and Crime	0.8364
Social Welfare	0.8510
Housing	0.6667
Domestic Commerce	0.7500
Defense	0.8000
Technology	0.6667
International Affairs	0.6813
Government Operations	0.8750
Public Lands	1.0000
Other	0.8391

Table 1: F1 Scores for Different Policy Topics