Harnessing social media influencers or using X ads to combat misinformation? Pre-analysis plan *

Antonella Bandiera,[†] Joaquin Barrutia, [‡] Jeremy Bowles,[§] Horacio Larreguy,[¶] Shelley Liu,[‖] John Marshall,[**] & Eduardo Zago[††]

Current draft: August 6, 2024

Abstract

In partnership with the fact-checker Africa Check, we assess whether social media influencers (SMIs) can be effectively recruited to increase awareness of, and knowledge about, misinformation among their social media audiences. Our partnering fact-checker recruited "micro" SMIs—individuals, such as journalists and social activists, with significant followings—on Twitter in Kenya and South Africa. We then provided a random half of the SMIs with digital literacy training videos, as well as contemporaneous fact-checks, for them to share with their audiences through social media. We additionally cross-randomized the dissemination of identical content to their audiences via regular Twitter ads to compare whether SMIs are more effective than ads in ensuring that digital literacy training and fact checks reach and are internalized by their followers. This at-scale intervention ultimately seeks to understand how fact-checking institutions can effectively combat misinformation.

*The study received IRB Approval from Columbia University (IRB-AAAU4365), Duke University, and Strathmore University (SU-ISERC1593/23). We acknowledge funding from the Social Science Research Council (SSRC)'s Mercury Project and the New Venture Fund. Larreguy acknowledges IAST funding from the French National Research Agency (ANR) under grant ANR-17-EURE0010 (Investissements d'Avenir program).

[†]Assistant Professor, Political Science, ITAM. antonella.bandiera@itam.mx.

[‡]PhD student, Economics, Emory University, joaquin.barrutia.alvarez@emory.edu.

[§]Assistant Professor, Political Science and School of Public Policy, University College London. jeremy.bowles@ucl.ac.uk.

[¶]Associate Professor, Economics and Political Science, ITAM. horacio.larreguy@itam.mx.

[‖]Assistant Professor, Public Policy and Political Science, Duke University. shelley.liu@duke.edu.

[**]Associate Professor, Political Science, Columbia University. jm4401@columbia.edu.

[††]PhD student, Political Science, New York University. ez2490@nyu.edu.

Introduction

Potentially harmful misinformation about social, political, and public health issues is a growing problem. The problem is particularly acute in the Global South, where digital media literacy remains low and social media is often the only affordable source of online information (Arechar et al., 2022). The COVID-19 infodemic, which has likely led to harmful individual health choices around the world, highlights the need to identify scalable ways to counter the spread of misinformation and its effects. Indeed, conspiracy theories surrounding the pandemic and the vaccines may play an important role in contributing to low vaccination rates across the Global South (Mallapaty et al., 2022; Ackah et al., 2022).

Existing research highlights the utility of both *debunking* and *prebunking* interventions (Cook, Lewandowsky and Ullrich, 2017; Nyhan et al., 2020; Walter et al., 2020; Guess et al., 2020; Vraga, Bode and Tully, 2022), but each mode of combating misinformation faces major obstacles to implementation. Debunking, which principally involves providing fact-checks designed to counteract misinformation (van der Meer and Jin, 2020; Nyhan and Reifler, 2015), faces two connected problems. First, fact-checks are usually released too late: fact-checking occurs when the virality of the misinformation on social media is already naturally decaying, and misinformation has already caused harm. Second, fact checks have very limited virality relative to the misinformation they aim at debunking, so they fail to reach the wide audience affected by misinformation.

Prebunking, which instead relies on literacy training and warnings as a way of inoculating people against misinformation, mitigates these timing and virality issues by attempting to prevent belief in misinformation *before* citizens encounter misinformation (Cook, 2013; McGuire, 1964). However, while prebunking interventions have demonstrated effectiveness in various settings, they still involve recruiting and incentivizing people to consume information. If participants are not incentivized to consume and internalize literacy training and misinformation warnings, it is unclear (i) whether they would consume such information; (ii) the extent to which they would internalize the information; (iii) and whether the results are as long-lasting. Moreover, we have a particularly limited understanding of whether these interventions can be scaled at a low cost.

We tackle these outstanding issues in a study developed in partnership with Africa Check, the first and largest fact-checking organization in sub-Saharan Africa. Specifically, we ask whether "micro" social media influencers (SMIs) can be recruited to increase awareness and knowledge of misinformation among their online audiences by posting digital literacy training materials and fact-checks. The effectiveness of fact-checking interventions relies heavily on users' trust in and attention to a source (Bowles, Larreguy and Liu, 2020). Consequently, micro SMIs with engaged audiences may be an especially effective mode of delivery to target large numbers of individuals at

minimal cost. However, this raises a critical question: can such SMIs be incentivized to disseminate information about misinformation? And will their followers engage with it? While piloting indicated that financial incentives for posting digital literacy training materials and fact-checks are required for compliance (in the form of compensating the fact-checkers for internet data costs and time), their effectiveness is unclear. To benchmark the effectiveness of leveraging SMIs to disseminate digital literacy training materials and fact-checks, we additionally use an equivalent amount of funds to instead target ads to social media followers. Our goal is to inform the programming decisions that fact-checkers face relating to how to optimally mitigate misinformation's potentially deleterious effects.

Due to piloting and difficulties in the recruitment of sufficient SMIs, the intervention was implemented in three phases with 302 SMIs across Kenya and South Africa: *pilot*, *first batch*, and *second batch*. The *pilot* was implemented with 40 SMIs between December 5th, 2022 and January 29th 2023, the *first batch* with 152 SMIs between March 13th, 2023 and May 7th, 2023, and the *second batch* with 110 SMIs between May 1st, 2023 and June 25th, 2023.

We focus on four main sets of outcomes, which we measure using behavioral Twitter data among the followers of SMIs. First, we are interested in whether social media usage patterns change as a result of our intervention. Specifically, we measure the extent to which the treated SMI followers become *less likely* to post and engage with (1) content of worse quality, in the sense that they are not verifiable and are predicted to be, or resemble, fake content and (2) content from less reputable sources, and *more likely* to post and engage with (3) content that reports fact-checks and news from more reputable sources. To that end, we have scraped Twitter posts and used machine/deep-learning techniques (developed in the context of Bandiera et al. (2023)) to classify whether the posts are verifiable and, if so, whether they constitute likely misinformation. We similarly coded the links shared in posts according to their content and source reputation.

Second, we investigate how the intervention changes posts and engagement with social media content about topics particularly affected by misinformation, such as COVID-19 and vaccines against it, and politics. Importantly, we use Natural Language Processing (NLP) to distinguish whether the tone of the posts and engagement was positive, neutral, or negative. Third, to assess the extent to which the treatment content spilled over to the followers of the SMI followers, we measure interactions with the posts discussed in the first two sets of outcomes. Fourth, we assess the extent to which continued exposure to digital literacy training materials and fact-checks induced followers to follow, and engage with, content from Africa Check and SMIs during and after the intervention.

Regarding data collection, we first scraped baseline Twitter posts for the pilot and first and second batches of SMI followers (3 months before the start of each intervention). Additionally, we

scraped the Twitter posts for all eight weeks of the first-batch followers and some of the second-batch followers. Lastly, we scraped one month of post-intervention Twitter posts for the first batch and some followers for the second batch. We are in the process of scraping Twitter posts for the intervention and post-intervention periods for the remainder of the second-batch followers. The change in pricing of the Twitter API significantly slowed data collection efforts.[1] In the early stages of data collection, we conducted a preliminary analysis to understand the implications of data restrictions (e.g., removing very active users who are likely bots or users with very low activity on social media) for statistical power. None of the pre-registered analyses have been conducted to date.

Research design

Recruitment of SMIs

We study whether micro SMIs can be recruited to increase their followers' awareness and knowledge of misinformation, and in turn alter their online behaviors. We investigate this question by partnering with Africa Check to recruit SMIs to disseminate digital literacy training materials and fact checks to their followers.

Small-scale SMIs in Kenya and South Africa were first identified using the academic Twitter API, which allowed us to collect information on all public Twitter accounts. These SMIs were screened according to three inclusion criteria. The first criterion is location. Having tracked all tweets based in Kenya and South Africa for approximately a year, we selected handles that explicitly referenced one of those countries as the user's location. Thus, we excluded profiles that did not have a study country as their location. The second criterion is the popularity of Twitter users. We *initially* selected users with more than the country's average number of followers among our sample of Twitter users in Kenya (4,490) and South Africa (3,266) at the time of SMI selection. The third criterion is whether the individual engages in high-quality news sharing. To this end, we retrieved the information shared by the users during a fixed period. Using a list of reputable media outlets in each country validated by Africa Check, we identified posts that contain links to a high-quality news article or fact-check and *initially* kept only users who had made at least one such post.

Our inclusion criteria yielded an initial sample of 732 SMIs from Kenya and 1,309 from South Africa. Africa Check then vetted these potential SMIs, filtering those they did not think would

[1] In February 2023, Twitter announced that access to the API would no longer be free for academics and introduced paid tiers to access the API. Twitter's cheapest tier with the functionalities we require costs $5,000 monthly, and it only allows access to 1,000,000 posts or interactions with posts monthly. Depending on additional scraping, our overall data collection is likely to comprise approximately 200,000 followers and hundreds of millions of posts and interactions with posts prior to, during, and after the study's completion.

be good ambassadors for their institutions, leaving a sample of around 150 potential SMIs across both countries. Finally, the fact-checkers approached these potential SMIs to gauge interest in participating in the study.

Due to the small sample of potential SMIs and even smaller yield, we revisited our sampling decisions. We relaxed the second and third inclusion criteria. We allowed for potential SMIs that were below, but close to, the threshold of the country's average number of followers and had at least 2,000 followers. We removed the constraint that the users made at least one post that contained links to a high-quality news article or fact check. Moreover, Africa Check recruited SMIs through Twitter ads, which we also ensured had sufficient followers after Africa Check vetted them.

Ultimately, we recruited a total of 302 SMIs across both countries in what we denote three phases: **pilot** (40), **first batch** (152), and **second batch** (110).[2] Table 1 summarizes the distribution of recruited SMIs in each phase and country by recruitment criteria.

Table 1: Distribution of SMIs according to recruitment strategy

		Recruitment Criteria			
Phase	Country	Both	Only Popularity	Only News Sharing	Twitter Ads
Pilot	Kenya	20	0	0	0
	South Africa	20	0	0	0
First Batch	Kenya	47	11	18	0
	South Africa	14	34	28	0
Second Batch	Kenya	3	2	4	49
	South Africa	3	21	12	16
Mean [Followers]		31,497.551	14,607.706	2,873.952	10,985.446
Std. Dev. [Followers]		55,361.358	18,352.048	695.712	30,804.893

Notes: This table presents the distribution of SMIs at each phase and country by recruitment criteria. Column 1 and 2 refers to the phases of the intervention and the country of origin of the SMIs. Columns 3-7 present the number of SMIs obtained by each inclusion criteria. Column 3 shows the SMIs that complied with both the popularity and high-quality news-sharing criteria, meaning those SMIs that shared at least one high-quality news or fact check and had more than 4,490 followers in Kenya and more than 3,265 followers in South Africa. Column 4 shows all SMIs with more than the average number of followers in each country but who did not share any news or fact checks. Column 5 shows all SMIs that shared at least one news or fact check and had more than 2,000 followers but less than the average number of followers in each country. Column 5 shows all SMIs recruited through Twitter Ads.

In practice, our lists of micro SMIs comprise mostly journalists and social activists in each country. A natural concern with this pool of SMIs is the possibility of floor effects, given that these

[2]Out of the 40 SMIs recruited for the pilot, 20 SMIs corresponded to each country. Out of the 152 SMIs recruited for the first batch, 76 SMIs corresponded to each country. Out of the 110 SMIs recruited for the second batch, 58 SMIs corresponded to Kenya and 52 corresponded to South Africa.

SMIs usually share high-quality information and their followers probably have higher social media literacy and are already less susceptible to misinformation. While this is a reasonable concern, Africa Check was understandably unwilling to recruit potential misinformers to disseminate the content of the intervention. Nevertheless, it remains unclear whether even the followers of relatively credible SMIs possess sufficient social media literacy, while they may also be particularly receptive to intervention content.

Treatments and randomization

SMI treatment

Upon enrollment into the study, we randomized half of the SMIs within each country to receive the *treatment* condition;[3] the other half served as the *control* group.

Treated SMIs were encouraged to share two types of content with their followers through their Twitter account each week: (i) an informative video produced by Africa Check that aimed to increase media literacy; and (ii) one fact-check produced by Africa Check pertaining to their country. These prebunking and debunking efforts were designed to be complements and were disseminated by treated SMIs for a period of eight weeks. The SMIs received small payments for their time and internet data costs. However, the SMIs were *not* contracted based on actually sharing the provided materials. In contrast, SMIs assigned to the *control* group were asked not to share content by Africa Check. Consequently, followers of control SMIs were very unlikely to receive the information disseminated by treated SMIs, since such content was not widespread on social media. We discuss below how we handle the potential overlap in the sets of followers between treated and control SMIs.

Ads treatment

We additionally sought to compare whether micro SMIs are more effective than regular social media ads in ensuring that fact-checking messages reach, and are internalized by, users. To assess this, we cross-randomized some of the followers,[4] irrespective of whether they followed treated or control SMIs, into two groups. The first *ads* group received targeted social media ads that delivered Africa Check's videos and fact-checks. These ads were disseminated using Twitter's ad targeting tool around the same days that treated SMIs were asked to promote Africa Check's weekly content. The second *no ads* group was not targeted with ads.

[3] See randomization details in Appendix section 2.
[4] See randomization details in Appendix section 3.

Importantly, we mainly targeted Twitter ads to followers who we identified as having very strong or strong ties with at least one SMI, which are roughly 7% of the followers, and a randomly selected 1% share of followers with weak ties to SMIs. We define a tie as very strong when a follower liked *and* replied to content shared by an SMI approximately a year before the intervention began, a strong tie when a follower liked *or* replied to content shared by the SMI before the intervention began, and a weak tie when a follower did previously not interact with the SMI. The logic of restricting ads to followers with very strong and strong ties reflected the difficulty of targeting a meaningful share of followers without a considerable budget and that SMIs were unlikely to reach all their followers. We primarily targeted strong and very strong ties with ads to approximate the likelihood that the Twitter algorithm is much more likely to show the SMI posts to followers with stronger ties.[5] Thus, our data collection efforts also focused on collecting user data of strong and very strong ties. This was also due to the limited tokens we had for the Twitter API while it was free for academics, and how slow it has been to scrape the remainder of the posts and interactions once Twitter changed its pricing strategy.

We created and uploaded lists of followers assigned to be treated with Twitter ads. Not all users are always uploaded, so successfully uploaded lists may be smaller than the original uploaded ists. Moreover, Twitter might not target all users who are included in the successfully uploaded lists. Out of 73,363 targeted followers, 36,459 (50%) were successfully targeted with ads in the first batch, and out of 34,403 targeted followers, 25,680 (75%) were successfully targeted with ads in the second batch.

Overview of intervention

In sum, followers of SMIs participating in the study were, at random, either (1) treated with information from one or more SMIs they followed, (2) treated with information from ads, (3) treated with both, or (4) assigned to receive no information as part of this intervention.

Following the recruitment and randomization of SMIs and followers, the intervention then proceeded as follows. Over the course of eight weeks, treated SMIs in each phase of the intervention received materials from Africa Check twice a week and were encouraged to tweet about these materials to their followers. The research team tracked selected SMI followers' social media activity throughout the eight weeks of treatment and four weeks after.

Tables 2 to 7 show the content of the intervention in Kenya and South Africa.

All SMI posts included the hashtag #FactsMatter. Following the findings from Bowles et al. (2024), they use empathetic language that highlights that it is understandable that fears led people

[5]We will verify the assumption that strong and very strong ties are more likely to receive SMI posts.

to fall for and share misinformation. Moreover, they prime shared values and social images, such as highlighting the importance of fact-checking before sharing to protect friends and family.

Estimation

Our primary goal is to estimate the effect of the intervention on the behaviors of the followers of treated micro SMIs or users who received the intervention content via targeted Twitter ads. To do so, we will employ standard regression analyses to identify average and heterogeneous treatment effects. Specifically, we estimate non-parametric and parametric regressions of the following form:

$$\begin{aligned} Y_{ft} &= \alpha_k + \sum_{j=1}^{J} \tau_j 1[T_f = j] + \sum_{j=1}^{J} \sum_{k=1}^{K} \tau_{j,k} 1[T_f = j](1[N_f = k] - \overline{N}_k) \\ &\quad + \sum_{b=1}^{B} \beta_b 1[b_f = 1] + \gamma Y_f^{pre} + \boldsymbol{\delta} \mathbf{X}_f^{pre} + \varepsilon_{ft}, \end{aligned} \tag{1}$$

$$\begin{aligned} Y_{ft} &= \alpha_k + \tau_{linear} T_f + \sum_{k=1}^{K} \tau_{linear,k} 1[T_f = j](1[N_f = k] - \overline{N}_k) \\ &\quad + \sum_{b=1}^{B} \beta_b 1[b_f = 1] + \gamma Y_f^{pre} + \boldsymbol{\delta} \mathbf{X}_f^{pre} + \varepsilon_{ft}, \end{aligned} \tag{2}$$

where Y_{ft} is an outcome for follower f at stage of the intervention t, $T_f \in \{1, ..., J\}$ is the total number of *treated* SMIs followed at the point of randomization (or treatment dosage), α_k are fixed effects for K groups defined by the total number of experimental (treated or control) SMIs followed ($N_f \in \{1, ..., K\}$) at the point of randomization, $\overline{N}_k$ is the share of the sample following k SMIs, and ε_{ft} is an error term. Because followers who follow more than one SMI in the experimental sample are more likely to follow treated SMIs, the α_k fixed effects are critical for identification; following Lin (2013), the interaction between treatment and each demeaned level of N_f enables us to estimate the average treatment effect of each dosage across these strata. The first specification then non-parametrically estimates the average intent-to-treat effect of following j treated SMIs, τ_j. Because it is rare to follow more than a couple of experimental SMIs, τ_j will likely be relatively imprecise for more than a few treated SMIs. The second specification instead estimates a sample-averaged linearized effect, τ_{linear}, across treatment dosages. To estimate just the effect of exposure to the first treated SMI, we restrict the sample to followers of only a single experimental SMI; the two specifications are identical for this subsample.

To increase precision, we further include the following covariates: indicators $b_f \in \{0, 1\}$

for following an experimental SMI from a given randomization block (i.e. fixed effects for each experimental SMI a follower follows), baseline outcome Y_f^{pre} (wherever possible), and baseline covariates $\mathbf{X}_f^{pre}$. We will select X_f^{pre} in two different ways. First, we will follow the Belloni, Chernozhukov and Hansen (2014) approach using LASSO to select the union of covariates correlated with outcome and treatment. Second, we will consider those covariates interacted with the likelihood of assignment to treatment, as well as the individual variables. Roth and Sant'Anna (2023) show that the latter approach is the most powered in designs such as ours. Moreover, this set of controls takes into account that some followers are much more likely to receive certain treatment assignments.

Further, we will consider two approaches to weighting the estimates to deal with heterogeneity in the number of followers per SMI in the experimental sample: providing unweighted estimates at the follower-level, and inversely weighting by the number of followers in the sample of each SMI. While the former is potentially the policy-relevant parameter, given large potential variation in the reach of different SMIs, the latter will estimate the effect for the average SMI (treating them equally, and thus potentially increasing the power of the estimates).

Our analysis is divided into four relevant stages, denoted by t. The first is the baseline period ($t = 0$), which covers three months before the start of the intervention. Then, to understand whether the treatment had immediate or cumulative effects throughout the intervention period, we will divide the intervention period into two stages of four weeks each ($t = 1$ and $t = 2$, respectively). Finally, to study medium-term effects, we will also analyze four weeks after the intervention ($t = 3$).

We conduct inference using randomization inference, using 1000 pseudo-treatment assignments, to account for the fact that different individuals follow different SMIs. Since there are many follower configurations, there is no obvious way to produce valid multi-way clustered standard errors.

We additionally analyze the effect of Twitter ads providing identical intervention content to that disseminated by SMIs. To do so, we estimate regressions of the following form:

$$Y_{ft} = \tau_{ads} Ads_f + \beta_b + \gamma Y_f^{pre} + \boldsymbol{\delta} \mathbf{X}_f^{pre} + \varepsilon_{ft}, \tag{3}$$

where Ads_f indicates being targeted with Africa Check's ads, and β_b are randomization-block fixed effects for the ads assignment.[6] Due to the individual-level randomization and lack of overlapping treatment assignments, we estimate robust standard errors.

In a secondary analysis, we explore whether content from SMIs and Twitter ads are complements or substitutes.

While Kenya and South Africa are clearly different contexts, we will consider estimates with

[6] Appendix 3 provides a detailed description of the stratification/blocking used for this randomization.

the pooled sample for ease of exposition. Moreover, our analysis will focus on the first and second batches, and thus exclude the pilot batch, for two primary reasons. First, the pilot phase was designed to assess the experiment's feasibility, identify best practices to communicate with treated and control SMI, and understand how to collaborate effectively with Africa Check. Second, the change in pricing of the Twitter API significantly complicated data collection, making it particularly challenging to recover pilot phase information through traditional web scraping techniques.

Data

Sample of SMI followers

Our analysis relies on data scraped from Twitter (X). We measure behavioral outcomes by first identifying a sample of followers with public profiles associated with each experimental SMI. This sample includes followers with very strong or strong ties to the SMI (as defined above), as well as a random sample of followers with weak ties to the SMI. Due to restrictions imposed by the Twitter API, we are limited in our ability to track all weak ties. We monitor these followers' posts and the social engagement (likes, comments, shares, and quotes) with their posts during and after the study.

Among the first batch of SMIs, we also measured whether followers followed Africa Check and SMIs by the end of the intervention. We were unable to do this for the second batch since Twitter API pricing changes prevented us from accessing the endpoint required for this. Instead, we measured the follower's social engagement with Africa Check and SMI's posts.

To enhance the precision of our estimates, we filtered out followers at both extremes of the SMI engagement distribution. Baseline information was employed to exclude potential bots, which we defined as users posting an anomalously high volume of tweets within a short period.[7] Moreover, we also filtered out followers who did not post any tweets during their baseline period, thereby ensuring that our analysis concentrated on active users. After aggregating followers from both South Africa and Kenya, the final sample for analysis consists of around 170 thousand followers.[8]

Measurement of outcomes

Our outcomes are organized into six subcategories: activity, verifiability, fake/true, topic sentiment, URLs, and interactions. For each subcategory, we define and, in some cases, create various outcome

[7] Users above the 95th percentile of tweets posted at baseline were excluded.

[8] Note that our final sample size will likely be lower since the followers might close their accounts or make them private. We do not expect this to be different according to treatment status.

variables that enable us to assess the impact of the intervention on users' overall behavior.

Social media activity

At the sender level, we measure the extent to which SMIs made posts with the treatment information, and the impressions and engagement their posts received. We similarly measure the impressions and engagement of Twitter ad posts with the treatment information received.

At the follower level, we measure the extent to which SMI followers posted, shared, replied to, and quoted others' posts. We refer to these forms of engagement collectively as "posts", since we are unable to distinguish between them for the subset of data scraped through traditional techniques following Twitter API's price changes.

Verifiable content

We trained a machine learning model to label followers' posts as verifiable or non-verifiable. Thanks to Africa Check's help labeling a sample of posts, we trained a binary classification model based on the Bidirectional Encoder Representations from Transformers (BERT) architecture. This model labeled posts as either verifiable or non-verifiable, and had a validation accuracy of 85%.[9] Beyond being of interest itself, by identifying verifiable posts, we could subsequently label verifiable posts as fake or true, as we explain next.

Fake/true content

We developed an additional model to label followers' verifiable posts as either approximating content that is fake or true.[10] This binary classification model was trained using posts labeled as "fake" by Africa Check in their fact checks or as "true" if they originate from reputable news sources, such as established newspapers. Similar to our verifiable model, we utilized the BERT architecture. Our model had a validation accuracy of 84%.[11]

URL source

Since information and misinformation are often not explicitly written in posts, but shared through links, we analyze the URLs shared by followers. To do this, we first extracted all URLs from the

[9] Appendix 4 provides a detailed description of the model, the training dataset, and the validation accuracy.

[10] Piloting and experience in another project with Africa Check indicate very low accuracy if we run this model on non-verifiable posts.

[11] Appendix 5 provides a detailed description of the model, the training dataset, and the validation accuracy. See Appendix 6.1 for a comprehensive description of how we built the training data.

posts and grouped them by domain name,[12] and manually classified them as follows. First, we distinguish between links to information sources (e.g. newspapers, fact checks, blogs, etc.) versus other websites that do not provide information (e.g. gambling websites). Second, for information sources, we distinguish between reliable and non-reliable news websites, fact checks, and other information sources.

To categorize the news websites from Africa, we first leveraged a dataset from Africa Check that classified reliable and non-reliable new websites. Additionally, Africa Check reviewed and classified an extra batch of African news websites. For global news websites, we used several sources, including NewsGuard and MediaBias.[13]

Topic sentiments

We analyze the effect of our intervention on the sentiment expressed by followers towards topics usually the subject of misinformation, such as COVID-19 and COVID-19 vaccines, and politics. We first identified and labeled tweets containing any information related to a given topic. For example, in the case of COVID-19 and COVID-19 vaccines, we manually defined a set of keywords, such as "coronavirus", "COVID-19", "mRNA", "Pfizer", etc., and filtered all posts containing these terms. From the subset of posts containing these keywords, we extracted the most frequent words (excluding the initial set of keywords) to ensure that no relevant terms related to the topic were excluded. We then identify the set of posts containing any of these terms.

We employ two different sentiment analysis models to predict the sentiment of each post as positive, negative, or neutral. Our primary analysis utilizes a BERT model, trained on a comprehensive dataset sourced from Kaggle, to perform sentiment prediction on the tweets.[14] Additionally, for robustness, we consider VADER, which is a lexicon and rule-based sentiment analysis tool specifically designed to detect sentiments expressed on social media.

[12]The net domain name for famous sites such as Daily Mail UK and Fox News are "www.dailymail.co.uk" "www.foxnews.com", respectively.

[13]We share four examples of sources we used to categorize the new websites, most coming from NewsGuard. For reliable news websites, we encountered several lists from reliable news sources such as Forbes and also directly from NewsGuard. For non-reliable news websites, we primarily relied on NewsGuard's veracity grading articles, for example for RT and for The Daily Beast Hunter.

[14]Appendix 7 provides a detailed description of the dataset, the model used, and the evaluation metrics.

Social interactions with follower content

We use data on the social interactions with the posts produced by followers to describe the popularity of each post that their own followers interacted with.[15] To conduct this analysis, we computed total interactions (likes, shares, comments, and retweets) with each follower's posts, as well as those specific to different types of posts (such as verifiable and non-verifiable posts) and posts with different sentiments about various topics (such as about COVID-19 and the COVID-19 vaccines).

Missing data

We will drop the data from the few users whose content is no longer available on Twitter because they closed their accounts, were banned, or made their content private. However, we will ensure that there is no differential attrition and that the randomization is not compromised.

Hypotheses

Finally, we describe our hypotheses for our various outcome variables. We will conduct one-sided hypothesis tests when the direction of the test has been pre-registered and the sign of the estimates is consistent with the pre-registered hypothesis, and carry out two-sided tests otherwise.

Hypothesis 1: Treatment delivery and uptake

We hypothesize that treated SMI posts and Twitter ads will successfully deliver Africa Check's media literacy materials and fact-checks to followers. Specifically, we expect followers treated via either the SMIs they follow or Twitter ads to be more likely to engage with the treatment information through likes, replies, shares and quotes. Although there may be negligible spillover to control SMIs and followers, these effects should be greater for those targeted directly.

Whether these effects are greater for treatment delivered by SMIs or Twitter ads is an empirical question. On the one hand, Twitter ads should naturally get more impressions, given that they were paid for. On the other hand, engagement with the posts likely depends on the followers' trust in, and attention, to a source (Bowles, Larreguy and Liu, 2020), and we may expect these to be greater for the SMIs than the Twitter ads posted by Africa Check.

[15] We focus exclusively on original tweets, not retweets, as the API provides interaction metrics for the initial tweet in the case of retweets.

Hypothesis 2: Social media activity

We expect treated followers to increase their knowledge of misinformation traits and, consequently, their ability to discern between fake and true information and knowledge of verification methods, both of which were covered in the treatment information.

We expect any increases in knowledge and discernment induced by either form of the intervention to affect followers' posting behavior on social media. In particular, we expect that exposure to the treatment information will reduce the extent to which treated followers create and share posts that we label as non-verifiable or verifiable but fake, which are respectively less informative or misleading. We are unsure of whether these reductions might be accompanied by increases in the creation and sharing of verifiable true posts, given the greater costs of producing and sharing that type of information. We thus expect an overall reduction in total posting and sharing.

Similarly, we expect exposure to either form of the intervention to reduce posts and shares that include links to unreliable—and thus non-informative—websites. Moreover, we expect a more modest increase in the number of posts and shares with links to fact checks and reliable news sources by treated followers, given the greater cost of sourcing such information. Overall, we expect exposure to the treatment information to decrease the total number of posts and shares *with links*.

Lastly, we will empirically explore whether the effect of content disseminated by SMIs is larger than the effect of information disseminated by Twitter ads on each of these outcomes.

Hypothesis 3: Sentiments toward topics affected by misinformation

We expect exposure to either form of the intervention to improve sentiment towards topics negatively affected by misinformation. This is particularly the case for COVID-19 and the COVID-19 vaccines, given that some of the treatment information was about these subjects. While we cannot measure beliefs and attitudes through a survey, we can observe the overall number of posts about COVID-19 and the vaccines against it, as well as the number of positive, neutral, and negative posts. We expect more posts with a positive or neutral sentiment and fewer with a negative sentiment.

Hypothesis 4: Social interactions

While the type of posts indicates the treatment information's effect on followers, we are also interested in treatment spillover to the followers of the SMI followers. To that end, we focus on the social interactions with the followers' posts, and those specific to different types of their posts, such as non-verifiable and verifiable fake posts, posts with links to fact checks, reliable news sources, and non-reliable news sources, and post with different sentiments regarding COVID-19 and the

COVID-19 vaccines. The expected treatment effects on the different types of followers' posts, which we discussed earlier, should also carry on to the interactions on them.

Hypothesis 5: Sustained exposure to treatment content

Lastly, we are interested in the extent to which fact-checkers can continuously expose social media users to digital literacy training materials and fact-checks. We then assess the effect of either form of the intervention on the extent to which followers follow and engage with content from Africa Check and SMIs during and after the intervention. While we expect not to observe negative effects, it is nevertheless important to assess if there are negative effects that undermine the potential for long-term exposure.

Heterogeneity

We will empirically explore whether stronger ties between SMIs and followers are associated with stronger treatment effects, and whether content from SMIs and Twitter ads are complements or substitutes.

References

Ackah, Betty BB, Michael Woo, Lisa Stallwood, Zahra A Fazal, Arnold Okpani, Ugochinyere Vivian Ukah and Prince A Adu. 2022. "COVID-19 vaccine hesitancy in Africa: a scoping review." *Global health research and policy* 7(1):1–20.

Arechar, Antonio A., Jennifer Allen, Adam J. Berinsky, Rocky Cole, Ziv Epstein, Andrew Gully, Jackson G. Lu, Robert M. Ross, Michael N. Stagnaro, Yunhao Zhang, Gordon Pennycock and David G. Rand. 2022. "Understanding and Reducing Online Misinformation Across 16 Countries on Six Continents." PsyArXiv.

Bandiera, Antonella, Camila Blanes, Horacio Larreguy, John Marshall and Daniela Pinto Veizaga. 2023. "Can journalists be empowered through training and resources to counter misinformation." *Working paper* .

Belloni, Alexandre, Victor Chernozhukov and Christian Hansen. 2014. "Inference on treatment effects after selection among high-dimensional controls." *Review of Economic Studies* 81(2):608–650.

Blei, David M., Andrew Y. Ng. and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3:993–1022.

Bowles, Jeremy, Horacio Larreguy and Shelley Liu. 2020. "Countering misinformation via WhatsApp: Preliminary evidence from the COVID-19 pandemic in Zimbabwe." *PloS One* 15(10):e0240005.

Bowles, Jeremy, Kevin Croke, Horacio Larreguy, Shelley Liu and John Marshall. 2024. "Sustaining Exposure to Fact-checks: Misinformation Discernment, Media Consumption, and its Political Implications." *Americal Political Science Review, https://www.dropbox.com/scl/fi/patpg1fb3d8w1fzhuacfc/WCW*$_A$*PSR*$_R$*esubmission.pdf?rlkey* $=$ *ekp6vxqj93kuzcslwmvrs2s1cdl* $= 0$.

Cook, John. 2013. Inoculation Theory. In *The Sage Handbook of Persuasion: Developments in Theory and Practice*, ed. James Price Dillard and Lijiang Shen. Thousand Oaks, CA: SAGE Publications pp. 220–236.
URL: *https://dx.doi.org/10.4135/9781452218410*

Cook, John, Stephan Lewandowsky and K. H. Ecker. Ullrich. 2017. "Neutralizing Misinformation through Inoculation: Exposing Misleading Argumentation Techniques Reduces Their Influence."

PloS One 12(5):e0175799.
URL: *https://doi.org/10.1371/journal.pone.0175799*

Goodfellow, Ian, Yoshua Bengio and Aaron Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Guess, Andrew M, Michael Lerner, Benjamin Lyons, Jacob M Montgomery, Brendan Nyhan, Jason Reifler and Neelanjan Sircar. 2020. "A digital media literacy intervention increases discernment between mainstream and false news in the United States and India." *Proceedings of the National Academy of Sciences* 117(27):15536–15545.

Kingma, Diederik P. and Jimmy Ba. 2017. "Adam: A Method for Stochastic Optimization." *arXiv:1412.6980v9* .

Lin, Winston. 2013. "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique.".

Mallapaty, Smriti et al. 2022. "Researchers fear growing COVID vaccine hesitancy in developing nations." *Nature* 601(7892):174–175.

McGuire, William J. 1964. Some contemporary approaches. In *Advances in Experimental Social Psychology*, ed. Leonard Berkowitz. Vol. 1 Elsevier pp. 191–229.

Nyhan, Brendan, Ethan Porter, Jason Reifler and Thomas J. Wood. 2020. "Taking Fact-checks Literally But Not Seriously? The Effects of Journalistic Fact-checking on Factual Beliefs and Candidate Favorability." *Political Behavior* 42:939–960.
URL: *https://link.springer.com/article/10.1007/s11109-019-09528-x*

Nyhan, Brendan and Jason Reifler. 2015. "Displacing Misinformation about Events: An Experimental Test of Causal Corrections." *Journal of Experimental Political Science* 2(1):81–93.

Reed, Russell and Robert J. Marks. 2016. *Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks*. MIT Press.

Rosner, Frank, Alexander Hinneburg, Michael Röder, Martin Nettling and Andreas Both. 2014. "Evaluating topic coherence measures." *arXiv:1403.6397v1* .

Roth, Jonathan and Pedro H. C. Sant'Anna. 2023. "Efficient Estimation for Staggered Rollout Designs." *arXiv:2102.01291v7* .

Ruder, Sebastian. 2016. "An overview of gradient descent optimization algorithms." *arXiv:1609.04747v2* .

van der Meer, Toni G. L. A. and Yan Jin. 2020. "Seeking Formula for Misinformation Treatment in Public Health Crises: The Effects of Corrective Information Type and Source." *Health Communication* 35(5):560–575.

Vraga, Emily K., Leticia Bode and Melissa Tully. 2022. "Creating News Literacy Messages to Enhance Expert Corrections of Misinformation on Twitter." *Communication Research* 49(2).
URL: *https://doi.org/10.1177/0093650219898094*

Walter, Nathan, Jonathan Cohen, R. Lance Holbert and Yasmin Morag. 2020. "Fact-Checking: A Meta-Analysis of What Works and for Whom." *Political Communication* 37(3):350–375.
URL: *https://doi.org/10.1080/10584609.2019.1668894*

1 First and Second Batch Intervention Content

Table 2: **Africa Check Pilot Content: Kenya**

Week	Suggested Text	Link to content
I	As technology advances so too has the speed and ease at which misinformation travels. To protect your friends and family, here are 5 questions to ask yourself before forwarding a message. #FactsMatter	Link: 5 Questions to Ask Before Forwarding a Message
I	A video shared online shows a brazen robbery and Kenyan social media users have claimed it illustrates the increase in crime in the capital Nairobi. But it's from the South American country Guyana, not Kenya. #FactsMatter	Link: https://africacheck.info/3OqcoqI
II	The speed at which the Covid-19 vaccine was developed fueled understandable fears. But safety was not compromised - here's why. #FactsMatter	Link: The speed with which Covid-19 vaccine was developed did not compromise its safety.
II	Beware of Covid vaccine misinformation - a recent claim that Microsoft co-founder Bill Gates wants to "'force-jab' the unvaccinated" through vaccinated livestock is pure fabrication. #FactsMatter	Link: https://africacheck.info/pfizer_employees
III	Seeing is believing? Not so fast. Here's how to easily verify images and videos on social media and avoid being fooled by misinformation. #FactsMatter	Link: Quick Steps to Verifying Videos or Images
III	A disturbing video of a girl being assaulted has been shared online with the claim the incident took place in Kenya. But the video was shot in the DRC. #FactsMatter	Link: https://africacheck.info/3RsXyBr

Table 3: **Africa Check Content Plan: Kenya (Continued)**

IV	Heated arguments over dinner? Here's what to do when your loved ones share misinformation, without upsetting them. #FactsMatter	Link: Dealing with family who share misinformation
IV	Have road crashes claimed more lives than Covid-19 in Kenya? Get the facts here: #FactsMatter	Link: https://africacheck.info/3jFojWA
V	Why do people create health disinformation? And what's really the harm in sharing it? Keep your friends and family protected by looking out for these hidden motivations #FactsMatter	Link: Why do people share health misinformation?
V	While widespread antiretroviral treatment has drastically changed the lives of people with HIV, there is still no cure for the virus. Expensive treatments advertised on social media that claim otherwise should be ignored. #FactsMatter	Link: https://africacheck.info/fake_HIV_cure
VI	Here are @AfricaCheck's top tips for spotting false information and stopping its spread #FactsMatter	Link: Top 4 tips to spot fake news
VI	A video claims that a Kenyan doctor has invented a cure for high blood pressure and is therefore in danger. The video includes the logo of KTN News and a clip of anchor Ali Manzu. But the channel – and Manzu – never reported this "news". #FactsMatter	Link: https://africacheck.info/3WY6twP

Table 4: **Africa Check Content Plan: Kenya (Continued)**

VII	Medical advice and health tips are everywhere on social media. But some aren't backed by science, and can even cause harm. These verification steps can help protect family and friends from health misinformation #FactsMatter	Link: Verifying health claims quacks or false cures
VII	Some substances extracted from pineapples may help treat less serious conditions. But "hot pineapple water" definitely doesn't cure cancer. This dangerous claim rehashes old (untrue) advice that "hot lemon water" cures cancer. #FactsMatter	Link: https://africacheck.info/3kGLAIe
VIII	IT'S A SCAM! Make sure you're not fooled by using these simple steps. #FactsMatter	Link: Facebook scams and how to stop them
VIII	Beware of posts offering soft loans on WhatsApp from Kenya's Family Bank. The loan terms are attractive. But the offer isn't real, confirming that if something sounds too good to be true, it usually is. #FactsMatter	Link: https://africacheck.info/3lx7cHt

Table 5: **Africa Check First and Second Batch Content: South Africa**

Week	Suggested Text	Link to content
I	As technology advances so too has the speed and ease at which misinformation travels. To protect your friends and family, here are 5 questions to ask yourself before forwarding a message. #FactsMatter	Link: 5 Questions to Ask Before Forwarding a Message
I	What do we know about race and private sector ownership in South Africa? Three viral claims are investigated here: #FactsMatter	Link: https://africacheck.info/3B86Trx
II	The speed at which the Covid-19 vaccine was developed fueled understandable fears. But safety was not compromised - here's why. #FactsMatter	Link: The speed with which Covid-19 vaccine was developed did not compromise its safety.
II	A claim that South Africans were given the "wrong Covid vaccine" is incorrect. A batch of nasal spray to help treat Covid symptoms was ordered by the defence department, but returned to Cuba after the public protector stepped in. #FactsMatter	Link: https://africacheck.info/3W19A6K
III	Seeing is believing? Not so fast. Here's how to easily verify images and videos on social media and avoid being fooled by misinformation. #FactsMatter	Quick Steps to Verifying Videos or Images
III	No, this photo doesn't show the British ambassador to Somalia with a black child in a cage. #FactsMatter	Link: https://africacheck.info/british_ambassador_somalia

Table 6: **Africa Check Content Plan: South Africa (Continued)**

IV	Heated arguments over dinner? Here's what to do when your loved ones share misinformation, without upsetting them. #FactsMatter	Link: Dealing with family who share misinformation
IV	Is #SouthAfrica's much-criticised power utility @Eskom_SA overstaffed by 66%? And do salaries for 27,543 surplus staff amount to R1.38 billion per year? These claims have been much shared online. But are they accurate? #FactsMatter	Link: https://africacheck.info/eskom_salaries
V	Why do people create health disinformation? And what's really the harm in sharing it? Keep your friends and family protected by looking out for these hidden motivations #FactsMatter	Link: Why do people share health misinformation?
V	While widespread antiretroviral treatment has drastically changed the lives of people with HIV, there is still no cure for the virus. Expensive treatments advertised on social media that claim otherwise should be ignored. #FactsMatter	Link: https://africacheck.info/fake_HIV_cure
VI	Here are @AfricaCheck's top tips for spotting false information and stopping its spread #FactsMatter	Link: Top 4 tips to spot fake news
VI	Government ministers do sometimes put their feet in their mouths. But there's no evidence South African minister of water & sanitation suggested people shower 'in groups', however dire the country's water and electricity crises might be. #FactsMatter	Link: https://africacheck.info/3U7Flt3

Table 7: **Africa Check Content Plan: South Africa (Continued)**

VII	Medical advice and health tips are everywhere on social media. But some aren't backed by science, and can even cause harm. These verification steps can help protect family and friends from health misinformation #FactsMatter	Link: Verifying health claims quacks or false cures
VII	Some substances extracted from pineapples may help treat less serious conditions. But "hot pineapple water" definitely doesn't cure cancer. This dangerous claim rehashes old (untrue) advice that "hot lemon water" cures cancer. #FactsMatter	Link: https://africacheck.info/3kGLAIe
VIII	IT'S A SCAM! Make sure you're not fooled by using these simple steps. #FactsMatter	Link: Facebook scams and how to stop them
VIII	Beware of rings sold with the promise of weight loss or health benefits. It's a scam, confirming that if something sounds too good to be true, it usually is. #FactsMatter	Link: https://africacheck.info/3wphJGV

2 SMI Randomization

We formed blocks of size 4 and 2 considering the mahalanobis distance of all the following covariates.

- *n lists*: Number of lists that include the SMI.
- *n followers*: Number of users who follow the SMI.
- *n very strong*: number of very strong ties to followers.
- *n strong*: number of strong ties to followers.
- *n weak*: number of weak ties to followers.
- *days account creation* Days since creation of account.
- *n posts*: Total posts in the previous 6 months.
- *n posts news*: Total posts in the previous 6 months with a link to a news article.
- *interactions sum*: Total likes, quoted tweets, and replies they made in the previous 6 months.

3 Follower Randomization

For the followers, we first stratified the followers using the combination of the medians they were in for the following covariates:

- *n very strong*: number of very strong ties to an SMI.
- *n strong*: number of strong ties to an SMI.
- *n weak*: number of weak ties to an SMI.
- *n very strong treated*: number of very strong ties to a treated SMI.
- *n strong treated*: number of strong ties to a treated SMI.
- *n weak treated*: number of weak ties to a treated SMI.
- *days account creation* Days since creation of account.
- *interactions sum*: Total likes, quoted tweets, and replies they made in the previous 6 months.

We then formed blocks of size 4 and 2 considering the mahalanobis distance of all the above covariates.

4 Inferring the Verifiability of Twitter Posts Using BERT

To construct a training data set, we identified and scraped viral posts in English from three sub-Saharan African countries, Nigeria, Kenya, and South Africa, and sent them to Africa Check. Their fact checkers labeled 1,000 of these viral posts as *Not Verifiable*, *Verifiable*, and *Somewhat Verifiable*. The initial distribution of these labels was as follows. Out of the thousand posts, 477 (47.7%) were labeled as not verifiable, 158 (15.8%) as somewhat verifiable, and 365 (36.5%) as verifiable. To augment this data set for greater balance in verifiability, we incorporated 121 Twitter posts and 96 Facebook posts fact checked by Africa Check, and thus inherently verifiable. The final distribution of labeled posts is portrayed in Table 8.

Table 8: Distribution of 1,217 labeled posts:3 labels

Type	Label	Number	Percent
Not Verifiable	1	477	0.39
Somewhat Verifiable	2	158	0.13
Verifiable	3	582	0.48

We opted for Bidirectional Encoder Representations from Transformers (BERT), a deep learning model built on the robust transformers architecture to label posts as verifiable or not. Having been trained on millions of data points, this model had already learned the inner structure of the English language, and its features can be used to train a standard classifier. We followed the relevant literature to choose the optimization and loss functions. For the optimization function, we used AdamW, since the results are generally better than every other optimization algorithm, have faster computation time, and require fewer parameters for tuning (Kingma and Ba, 2017; Ruder, 2016). For Binary Classification problems, we use the Binary Cross-Entropy as a loss function, which is the go-to loss function for these types of problems, since the cross-entropy error function is often used for classification problems when outputs are interpreted as probabilities of membership in an indicated class (Reed and Marks, 2016; Goodfellow, Bengio and Courville, 2016).

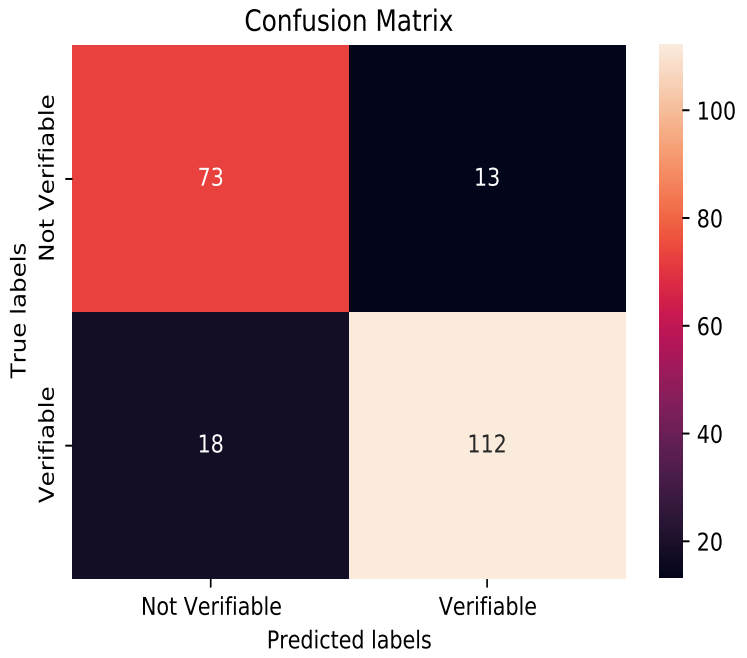
After training the first few models and failing to obtain one that could distinguish between the *verifiable* and *somewhat verifiable* classes, we decided to pool them together into one single class and train a binary classifier instead. The resulting distribution of labeled posts is portrayed in Table 9. Finally, upon trying several hyper-parameters we decided on the model with validation (out of sample) accuracy of 85% and cross-validation accuracy of 84%. The confusion matrix in Figure 1 shows this more clearly. This model kept the stop-words since BERT could utilize them to

better understand the context, and included a DropOut regularizer and a linear activation function.

Table 9: Distribution of 1,217 labeled posts: 2 labels

Type	Label	Number	Percent
Not Verifiable	0	477	0.44
Verifiable	1	600	0.56

Figure 1: Verifiability Model: Confusion matrix



5 Labeling Posts as Fake-True Using BERT

Over the years, Africa Check has manually verified thousands of posts deemed fake news and stored some of them alongside the fake-true predictions made. Out of these posts, we could recover the text from 456 posts, which were all (except for two) classified as fake news. To add true posts and gain balance between fake and true posts, we built a data set containing around 100 thousand newspaper posts from the three most reliable sources in Nigeria (Mobile Punch), Kenya (Standard Kenya), and South Africa (Daily Maverick).[16]. From these 100 thousand posts, we filtered out non-relevant posts related to opinions, sponsored content, and business reviews and randomly sampled 30 thousand Tweets from each newspaper.

We further matched 456 of those newspaper posts to the fake posts fact checked by Africa Check using matching and natural language processing techniques that allowed us to identify which true post was closest to a fake one in terms of the topics they covered (identified in an automated way) and common words used in each topic. This balanced data set, not only label-wise but also topic-wise, reduces the association between fake news and certain topics, which is usually time-dependent and affects fake news labeling outside the period of the data set. To do this, we specifically employed an LDA Topic Model with five topics.[17]

We augmented this data set with 600 political posts scraped from the FakeNet DataSet, which is a data set of US news tweets labeled as true or fake. The label distribution of the training is in Table 10.

Table 10: Distribution of 1,518 labeled posts

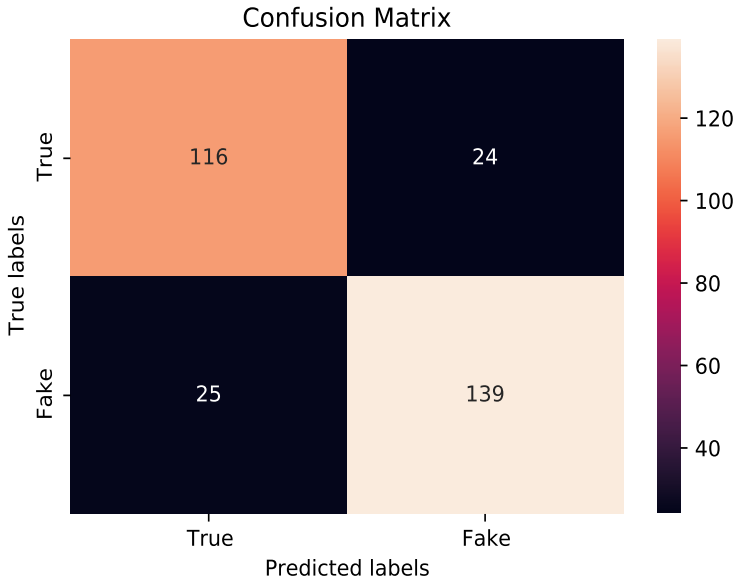
Type	Label	Number	Percent
Fake	0	692	0.456
True	1	826	0.544

As with the verifiability model, after training several models using BERT, we decided on the one model that had the highest test (out of sample) accuracy. We checked the cross-validation accuracy to discard that our results were influenced by the size of our training data set and/or the sampling to split training and validation. We obtained a validation accuracy close to 85%, as reflected in the following confusion matrix in Figure 2.

[16]These were recommended as reliable newspapers by the Chief Editors at AfricaCheck.

[17]A description of the model and the results can be found in Appendix 6.1

Figure 2: Fake/True model: Confusion matrix



6 Topic modeling

6.1 LDA model: fake posts dataset

Latent Dirichlet Allocation (LDA) is a probabilistic generative model widely employed in natural language processing for topic modeling. LDA posits that documents are generated from a distribution over latent topics, with each topic characterized by a distribution of words. The model's generative process entails the assignment of topics to words in documents, guided by prior distributions (Blei, Ng. and Jordan, 2003).[18]

In our experiment, we utilized the LDA model to analyze the topics present in our fake dataset. We then applied this model to predict the topics in a large collection of authentic posts gathered from reputable newspapers across Africa. After making these predictions, we compared the topic distributions of each post using cosine similarity. By doing so, we were able to find the most similar true posts for each fake post based on their topic distributions.

To decide the number of topics, we focus on objective measures of good fit, such as coherence and perplexity. Coherence measures the semantic similarity between high-frequency words within a topic. It helps assess how interpretable and meaningful the topics are. It is calculated by considering pairs of words that frequently co-occur within documents. A higher coherence score indicates that

[18]Given the timing of the deployment of the experiment, compared to when we analyzed the results of it, a few other models came into seen, like BERTopic.

the words in a topic tend to be more closely related in meaning. Therefore, a good coherence score suggests that the model has successfully identified coherent topics that capture distinct themes present in the corpus (Rosner et al., 2014). In turn, the perplexity score measures how well the model predicts the observed documents in the dataset. A lower perplexity score indicates better predictive performance.

The evaluation of LDA models based solely on coherence and perplexity scores may not provide a complete understanding of which model is a better fit. Although the scores across both metrics in Table 11 are better for the model with nine topics, we observed that the model with five topics produced the most human-interpretable results upon closer examination of each model's relevant words. In models with a relatively high number of topics, there was a tendency for repeated keywords across multiple topics. Furthermore, the greater the number of topics, the harder it is to interpret them, making it more challenging to discern meaningful distinctions between them.

Table 11: Topic Coherence and Perplexity across Different Models

Topics	Perplexity	Coherence
4	-8.463	0.414
5	-8.558	0.420
6	-8.673	0.430
7	-8.765	0.430
8	-8.775	0.440
9	-8.799	0.440

Given these observations, we decided to choose the five-topic model. These topics, as we can see in Figure 3, are related to economic conditions in Africa (topic 1), the political context (topic 2), health misinformation related to traditional medicine (topic 3) and health misinformation related with modern medicine (topic 5).

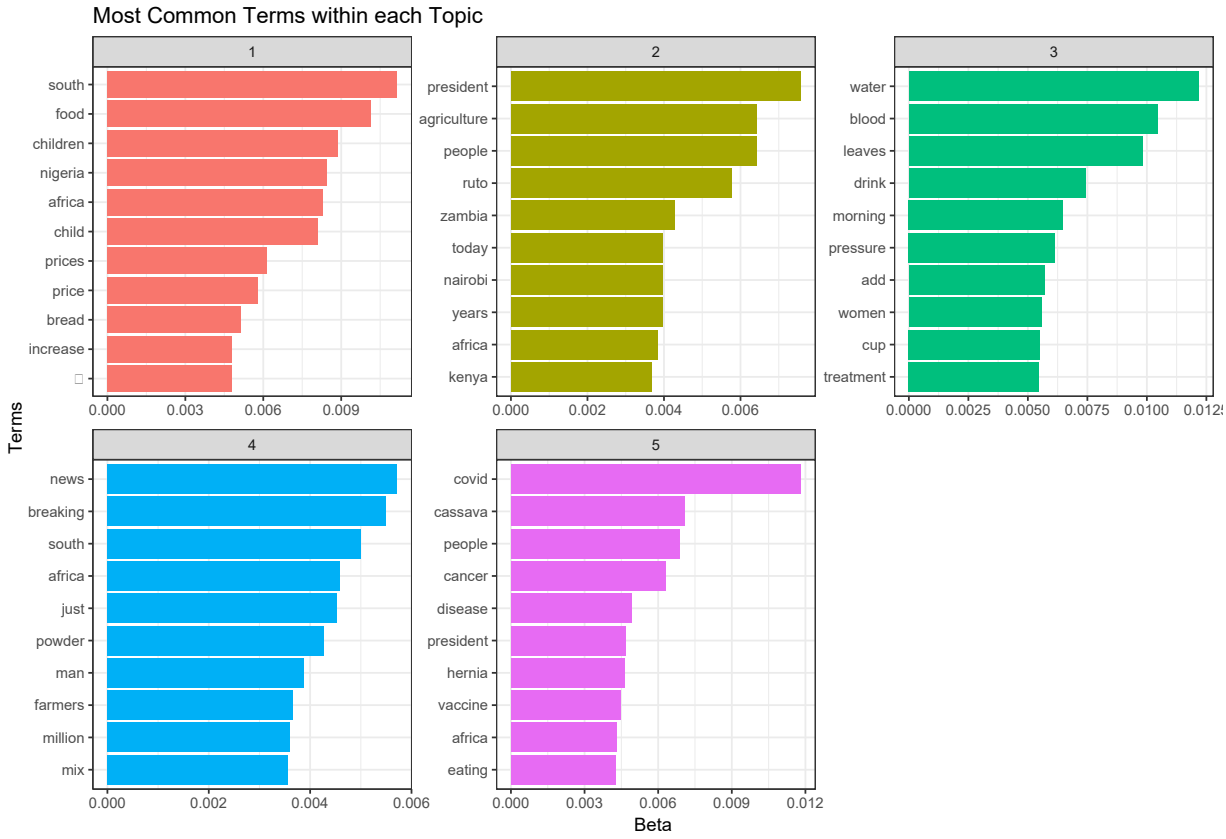


Figure 3: Five-Topic LDA Model: Word scores

7 Sentiment analysis

We analyze the sentiment towards topics subjected to misinformation using machine learning algorithms for sentiment prediction of posts.

For example, due to our particular interest in COVID-19 and vaccination campaigns, we obtained a training data set from Kaggle. This dataset is human-labeled and contains tweets from 2020 regarding COVID-19. It has five labels: extremely positive, positive, neutral, negative, and extremely negative. For practical reasons, we pooled together these five labels into three: positive, neutral, and negative. The data set is divided into training (41,157 observations) and test (3,798) data. Given the large size of the training data set, we fully balanced the data set across labels to improve accuracy, which yielded 18,046 observations for each label.

For our main analysis[19], we use a BERT (Bidirectional Encoder Representations from Transformers) model. Specifically, we fine-tuned the bert-base-uncased and added and trained a classifier layer on top of the BERT architecture. This BERT model is significantly bigger than the Small BERT (over 108 million parameters).

After five training epochs, we obtained the following metrics. We obtained 92% of validation accuracy and a test accuracy of 89%, calculated by the F-Score metric. Moreover, we see no major bias in any of the labels from the confusion matrix in Figure 5, meaning that the model does not confuse any label or has particular issues predicting a particular one. Also, by looking at the learning curve in Figure 4 we see no concern regarding over-fitting and under-fitting, which resulted from trying different sets of regularizers and even the Small BERT.

[19]We also use VADER for robustness of results.

Figure 4: Learning Curve: BERT Base

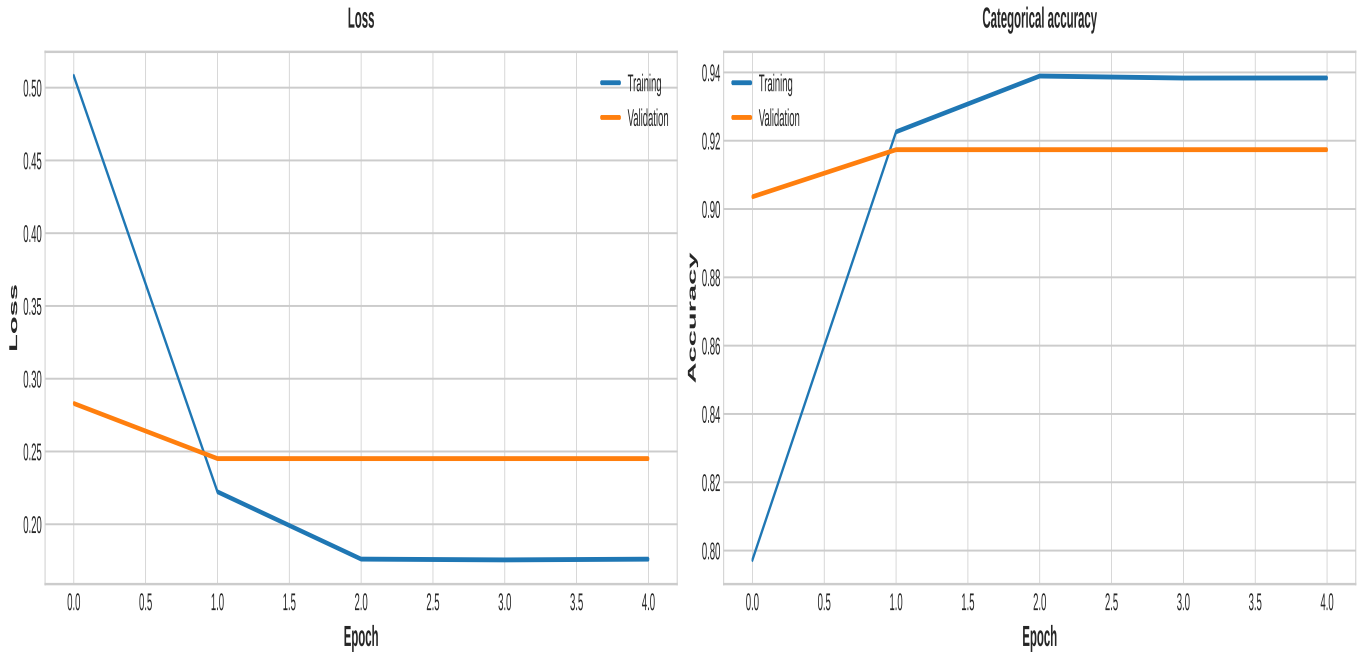


Figure 5: Confusion Matrix: BERT Base