

Ershov, Daniel; Yanting, He; Seiler, Stephan

Working Paper

How Much Influencer Marketing Is Undisclosed? Evidence from Twitter

CESifo Working Paper, No. 10743

Provided in Cooperation with:

Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Ershov, Daniel; Yanting, He; Seiler, Stephan (2023) : How Much Influencer Marketing Is Undisclosed? Evidence from Twitter, CESifo Working Paper, No. 10743, Center for Economic Studies and Ifo Institute (CESifo), Munich

This Version is available at:

<https://hdl.handle.net/10419/282431>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

How Much Influencer Marketing Is Undisclosed? Evidence from Twitter

Daniel Ershov, Yanting He, Stephan Seiler

Impressum:

CESifo Working Papers

ISSN 2364-1428 (electronic version)

Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH

The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute

Poschingerstr. 5, 81679 Munich, Germany

Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de

Editor: Clemens Fuest

<https://www.cesifo.org/en/wp>

An electronic version of the paper may be downloaded

- from the SSRN website: www.SSRN.com
- from the RePEc website: www.RePEc.org
- from the CESifo website: <https://www.cesifo.org/en/wp>

How Much Influencer Marketing Is Undisclosed? Evidence from Twitter

Abstract

We quantify the prevalence of undisclosed influencer posts on Twitter across a large set of brands based on a unique data set of over 100 million posts. We develop a novel method to detect undisclosed influencer posts and find that 96% of influencer posts are not disclosed as such. Despite stronger enforcement of disclosure regulations, the share of undisclosed posts decreases only slightly over time. Compared to disclosed posts, undisclosed posts tend to be associated with younger brands with a large Twitter following and are posted from smaller accounts that generate higher engagement per follower.

JEL-Codes: C550, M310, M370, M380.

Keywords: social media, influencer marketing, advertising disclosure, consumer protection.

Daniel Ershov
University College London / United Kingdom
daniel.ershov@tse-fr.eu

Yanting He
Imperial College London / United Kingdom
y.he21@imperial.ac.uk

Stephan Seiler
Imperial College London / United Kingdom
stephan.a.seiler@gmail.com

This draft: October 26, 2023

We thank seminar participants at Imperial College London and Helsinki GSE for their feedback. None of the authors received external funding for this paper. All errors are our own.

1 Introduction

Influencer marketing has emerged as a popular new channel to reach customers which, according to a 2019 survey, 93% of advertising and marketing professionals rely on (Michaelsen et al. (2022)). In contrast to traditional forms of advertising, influencer marketing involves private individuals posting about brands and being remunerated for their posts. Marketers often believe that influencer marketing can be more impactful than traditional advertising because messages from private individuals are perceived as more authentic. In response to this trend, regulators in many countries now require any paid content to be disclosed so that consumers are able to distinguish paid from genuinely organic content. However, it is not clear whether influencers will necessarily adhere to this regulation.¹ In this paper we use a novel data set of over 100 million brand-related posts from Twitter and a new classification approach to identify undisclosed sponsored content. The aim of this paper, is to quantify the overall importance of undisclosed influencer posts, track their evolution over time, and to characterize the brands and accounts from which they originate.

A key contribution of our analysis is a novel method for detecting undisclosed sponsored posts. Our approach uses a text-based classification algorithm that is trained on a set of organic and (disclosed) sponsored posts. We deal with two fundamental issues that arise in the context of detecting undisclosed sponsorship: First, the presence of undisclosed sponsored posts implies that we do not have access to a clean sample of organic posts to train our classifier. Second, any classification algorithm will classify organic and sponsored posts with error which needs to be accounted for to correctly quantify the share of undisclosed posts. We address the issue of a contaminated training sample with a pre-processing step that isolates a small sample of "true organic" posts. Then we train a machine learning algorithm on the set of true organic posts and disclosed-sponsored posts and use it to classify the remaining posts. We keep track of the classification error we observe in the training sample and show how to adjust the classification results appropriately.

We apply our approach to a large data set of over 100 million brand-related posts on Twitter for a set of 268 brands with a strong presence on the platform. Twitter constitutes an excellent test-bed to quantify the role of influencer marketing because it is prominently used for influencer advertising (Geysler (2019)). Moreover, the platform has been an earlier adopter of paid influencer content, thus allowing us to study its evolution over a relatively long time period from 2014-2021. Twitter allows us to collect particularly rich data through their data-API² and the text-based nature of the platform enables us to classify posts based on their content.³

Our main finding is that 96% of sponsored content is undisclosed. We find that despite stronger

¹Experimental evidence suggests that disclosure of commercial relationships between influencers and brands can reduce influencer trustworthiness (Karagür et al. (2022)).

²The Twitter data-API allowed researchers, until recently, to obtain the universe of all tweets discussing a particular brand. By comparison, Instagram, through the platform CrowdTangle, allows researchers to access a limited and non-random sample of posts.

³Our classification approach does not easily extend to image- or video-based content which is prevalent on Instagram and TikTok.

enforcement of disclosure regulation over time, the share of undisclosed content only decreases slightly from 98.9% in 2014 to 95.4% in 2021. Undisclosed sponsored posts tend to originate from younger brands with a large Twitter following and from smaller accounts that generate more engagement per follower relative to accounts associated with disclosed content. The high share of undisclosed content and the modest change over time suggest that current regulation does not effectively incentivize disclosure. The preponderance of non-disclosure among younger brands with strong social media presence indicates that future compliance rates may remain low.

This paper contributes to two strands of literature. The first is the growing literature in economics and marketing on influencer marketing. Most papers in this literature identify the effect of sponsored influencer posts on sales outcomes (Huang and Morozov (2022), Li et al. (2021), Gong et al. (2017), Hughes et al. (2019), Rajaram and Manchanda (2020), and Yang et al. (2021)), or examine the network structure of user-generated social media influencer markets (i.e., Lanz et al. (2019) and Valsesia et al. (2020)). A small number of theoretical papers discuss disclosure regulations (Fainmesser and Galeotti (2021), Pei and Mayzlin (2022), and Mitchell (2021)). Ershov and Mitchell (2020) empirically study the impact of disclosure regulation on influencers’ disclosure decisions and content production, and Karagür et al. (2022) experimentally evaluate the effects of disclosure on engagement with sponsored content.

The second strand of literature is on deceptive commercial content online, which so far has primarily focused on studying hidden fake reviews. Mayzlin et al. (2014), Luca and Zervas (2016), and He et al. (2022) document the presence of such content on hotel recommendation, food recommendation and online shopping sites. We contribute to this literature by showing the existence and extent of hidden commercial content on social media platforms using Twitter data. Similar to the fake review literature, our results show that there is a substantial amount of hidden commercial content online. There is also a parallel computer science literature on the identification of undisclosed sponsored content, which includes Kim et al. (2021) and Silva et al. (2020). These papers either use a human-labeled training sample or adopt a classification approach trained directly on posts with and without disclosure, without first identifying a set of “true organic” posts.

2 Background: Influencers & Regulation

We define influencers as social media users who monetize their content by posting about brands in sponsored posts. Brands typically maintain significant control over the content influencers produce related to their products. Sample contracts provided by marketing agencies include lists of “deliverables” and “mandatories,” such as hashtags, terms and links that the brand contractually requires the influencer to incorporate in posts (theinfluencers.com.au (2021)). Brands are also encouraged to advise influencers on terms they should avoid, such as mentions of competitor products or negative language (Geysler (2018)). Due to the level of control by brands, the European Advertising Standards Alliance specifically defines influencer marketing as having editorial content from sponsoring brands, including “a pre-suggested message script, scenario or speech for the influencer

[...] before its publication” (EASA (2018)).

In principle, influencers are governed by the same FTC advertising rules as agents in the traditional media. These rules stipulate that advertising content must be clearly labeled as such, and must be visually distinguishable from non-advertising content. The first regulatory action by the FTC specifically aimed at social media platforms and influencers was an “enforcement statement” and business guidance issued in December 2015 (FTC (2015a) and FTC (2015b)). This guidance transposed traditional media regulations to social media; anyone receiving compensation in return for posting about brands on social media had to provide clear and unambiguous disclosure of it being a paid advertisement. There was no mention of particular hashtags or language that the FTC required influencers to use, but both brands and influencers were potentially liable for improper disclosure and potentially subject to fines. Around the same time, the FTC initiated several cases against brands such as Warner Bros Home Entertainment Inc, the fashion brand Lord & Taylor, and the diet tea and skincare brand Teami for improper disclosure of influencer advertising. As well as dealing with brands, in 2017 the FTC sent “educational” letters to 90 prominent influencers and marketers reminding them to clearly disclose “material connections” between the influencers and advertisers (FTC (2017a)). Updated guidelines and information packages were released in 2017 and 2019 (FTC (2017b) and FTC (2019)). These included clear instructions to influencers about which disclosure hashtags or language they should use, such as #ad or #sponsored. The 2019 guidelines also stressed that the location of the disclosure should be prominent - i.e., at the beginning of a post.

In summary, there has been a tightening of disclosure regulations, characterized by more explicit rules and increased enforcement through cases and warning letters, over the course of our sample period. Despite these regulatory efforts, the probability of detecting violations may remain low. Social media companies have neither been held liable for non-disclosed advertising content on their platforms nor have they engaged in systematic attempts at monitoring or uncovering non-disclosed advertising. The FTC primarily learns about new cases of undisclosed advertising through consumer complaints. This means that popular influencers’ non-disclosure may be detected, but for less popular influencers, the incidence of detection is much lower. Warning or informational letters from the FTC about violating disclosure rules were only sent to the biggest/celebrity influencers. Moreover, punishments conditional on detection have been relatively lax. The FTC has not fined any influencers and most brands caught violating disclosure, such as Lord & Taylor and Warner Bros., settled their cases without paying any fines.⁴

3 Data

Our analysis is based on a sample of over 100 million brand-related posts which we downloaded directly from Twitter using their official research API.⁵ We first find a set of brands that have

⁴The only brand that was ever fined by the FTC was Teami. The magnitude of the fine, 1 million dollar, was small in comparison to the payment rates for celebrity influencer advertising (on the order of 250,000 dollars per post).

⁵This research API has been discontinued in early 2023.

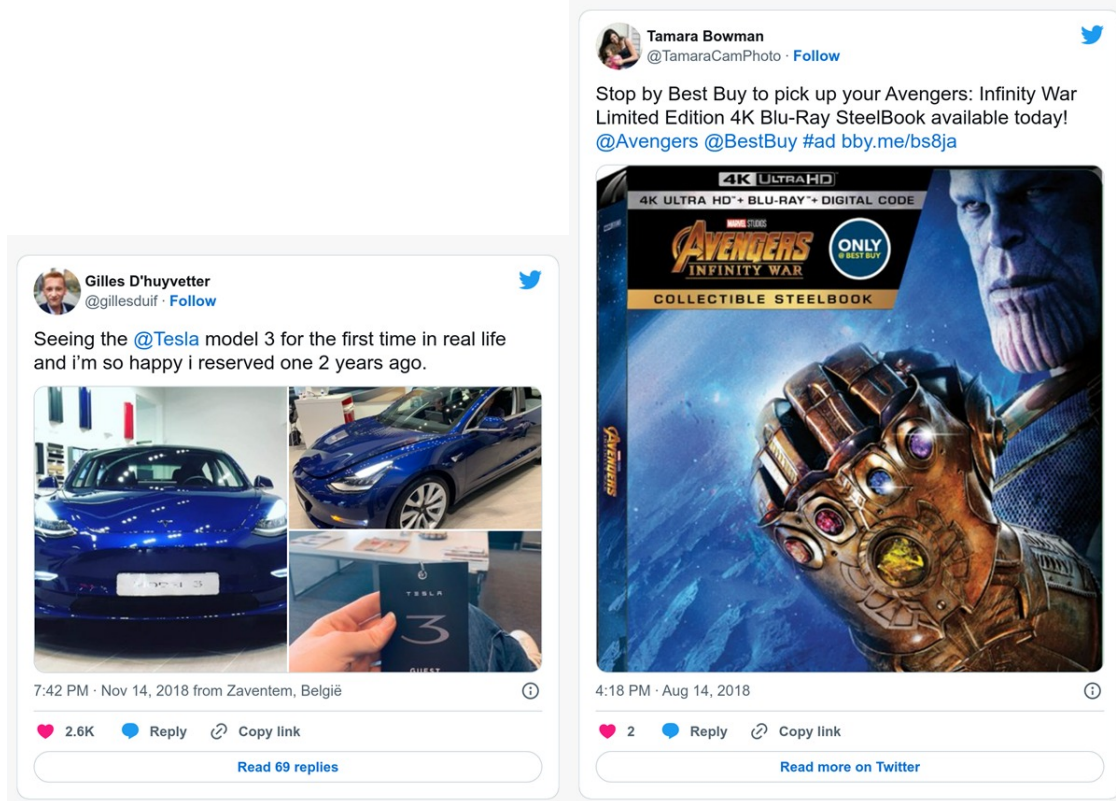


Figure 1: Example: Post without Disclosure (left) & Sponsored Post (right).

a large presence on Twitter as measured by their follower count and for which we observe some disclosed sponsored posts. Specifically, we select brands with verified accounts with more than 500,000 followers and at least 10 sponsored posts in 2021.⁶ We provide more details on how we select brands in Appendix A. Next, we obtain all posts that mention one of the 268 selected brands (a “mention” is defined as an “@” followed by the account name) between 2014 and 2021. Our final sample contains 101,514,289 posts: 667,774 posts with a sponsorship disclosure (#ad or #sponsored) and 100,846,515 posts without disclosure. Although secondary for our main analysis, we also collect all 2,314,160 posts from the brands’ official accounts during our sample period.

Posts with and without sponsorship disclosure have different content / language on average. We provide an illustrative example of a typical set of posts in Figure 1, as well as the top 20 bigrams (sets of two consecutive words) in both types of posts in Table 1.⁷ We find that the set of top bigrams for sponsored posts predominantly contain commercial language such as “available via” or “use code”, whereas organic posts are more likely to contain casual and conversational language.

⁶Because we use a low thresholds of at least 10 sponsored posts we are likely to capture most brands that engage in sponsored activity.

⁷Whenever two bigrams that are part of the same expression appear in the list (such as “laughing out” and “out loud”), we collapse them into one three word sequence. For posts without disclosure, we remove bigrams that appear in posts for fewer than 20 brands. In Appendix B we provide additional detail on how we compute the list of top bigrams.

No Disclosure	Sponsored	Official Account
laugh out loud	available via	chance win
customer service	use code	join us
what the fuck	best buy	pre order
first time	chance win	get ready
looks like	back school	in case you
last night	enter win	you missed it
brand new	limited edition	dont miss
oh my god	holiday season	coming soon
bring back	get free	brand new
social media	twitter party	weve got
years ago	dont miss	tell us
looking forward	holiday gift	happy birthday
every time	dropped via	behind scenes
feel like	giving away	stay tuned
anyone else	blog today	make sure
would like	supplies last	new york
dont know	get ready	take look
well done	try new	limited edition
direct message	perfect gift	modified tweet
new video	pre order	new year

Table 1: **Most Frequent Bigrams in Sponsored Posts and Posts without Disclosure.**

This is consistent with brands requiring influencers to generate content that is commercially viable, as described in Section 2. Moreover, sponsored posts and posts from brands’ official accounts (also in Table 1) contain similar language, suggesting that both types of posts fulfill a similar commercial purpose.

While posts without disclosure likely contain undisclosed sponsored posts, their non-commercial top bigrams suggest that organic posts constitute the majority of posts without disclosure. We will later use the top bigrams to isolate a set of undisclosed posts that belong to the majority class of organic posts.

4 Identifying Undisclosed Influencer Posts

We assume there are two dictionaries for sponsored and organic content with overlapping words but different frequencies of specific words (co-)occurring. As illustrated above, sponsored posts are more likely to contain commercial language, whereas organic posts contain more conversational language. These differences in the content suggest that we can use a classifier to identify which words are associated with sponsored and organic posts. Because we do not directly observe labeled undisclosed sponsored content, we assume that disclosed and undisclosed sponsored posts are drawn from the same dictionary. This allows us to learn about the content of undisclosed sponsored posts from

disclosed-sponsored posts. We believe that this is a reasonable assumption because brands exercise control over the content of sponsored posts and they want influencers to use commercial language that drives purchase behavior. Moreover, since the probability of detection and punishment for non-compliance with disclosure regulations is low, influencers and brands should have little incentive to alter the content of undisclosed-sponsored posts relative to disclosed-sponsored posts.

Next, we address two fundamental issues when trying to detect undisclosed sponsorship. First, while we have access to a sample of sponsored posts, we lack a clean sample of organic posts that we can use to train the classification model. Second, because dictionaries overlap, it is possible that a sponsored-looking post (i.e., a post with some commercial language) is actually organic and might be mis-classified by a content-based classifier. In the next two sub-section we show how to isolate a sample of “true organic” posts to train our classifier and how to correct for mis-classification error.

4.1 Compiling a Clean Training Sample

To identify a subset of organic posts, we focus on the most common terms in posts with and without disclosure from Table 1. We assume the top bigrams for posts without disclosure represent “organic” language. As discussed above, these bigrams do not contain commercial terms, suggesting that they come from organic rather than undisclosed-sponsored posts. The absence of commercial terms likely occurs because undisclosed-sponsored posts are the minority class among posts without disclosure and therefore the *top* bigrams are associated with organic posts.⁸

We compile a list of “true organic” posts by selecting posts that do not contain any of the top sponsored bigrams and that contain at least one of the top bigrams that appear in posts without disclosure. This selection produces a total of 3 million true organic posts which constitute 3% of all posts without disclosure. We then use these posts together with the disclosed-sponsored posts to train a classifier which we describe in more detail in Section 4.3. Before running the classifier, we remove the top organic and top sponsored bigrams from the posts in our sample. By removing the top bigrams, we only train our model on information that was not used to label the training sample of true organic posts.

Later we show that classification results are similar when using ChatGPT to generate a training sample by classifying a set of posts without disclosure as sponsored or organic. We describe results from this approach when presenting classification results in Section 4.3 and provide additional details in Appendix F.

4.2 Adjustment for Classification Error

After training a classifier on the sample of posts described above, we apply the model to the remaining set of posts without disclosure in order to detect undisclosed sponsored posts. To correctly quantify the prevalence of undisclosed sponsored posts, we need to adjust for the classification error of our algorithm. To fix ideas, consider the case where true organic posts have a 5% chance of being

⁸Our classifier results label approximately 17% of posts without disclosure as sponsored, confirming this assumption ex-post.

mis-classified as sponsored. Based on this mis-classification probability we would expect a similar rate of posts being classified as sponsored even if our test sample does not contain any sponsored posts. Because we are able to observe mis-classified organic and sponsored posts in the training sample, we can correct for classification error using an adjustment procedure.

We can write the probability of observing a post that is classified as organic as follows:

$$Pr(O) = Pr(O|O^*) \times Pr(O^*) + Pr(O|S^*) \times Pr(S^*)$$

where O and S stand for organic and sponsored posts and the variables with stars denote the true state of a post and the variables without stars denote how a post is classified. In the equation above $Pr(O)$ is observed as are the shares of organic and sponsored posts that are classified as organic $Pr(O|O^*)$ and $Pr(O|S^*)$. We also note that $Pr(S^*) = 1 - Pr(O^*)$ and therefore the only unknown variable is the share of true organic posts. Re-arranging the equation above gives us the formula for the share of organic posts that accounts for mis-classification:

$$Pr(O^*) = \frac{Pr(O) - Pr(O|S^*)}{Pr(O|O^*) - Pr(O|S^*)}. \quad (1)$$

It is useful to build some intuition based on special cases. If all posts are perfectly classified, i.e. $Pr(O|O^*) = 1$ and $Pr(O|S^*) = 0$, then the share of posts classified as organic is equal to the true share of organic posts. If sponsored posts are correctly identified, but organic posts are not it holds that $Pr(O|S^*) = 0$ and $Pr(O|O^*) < 1$ and therefore the share of organic posts is larger than the share of classified organic posts because the observed share needs to be adjusted upwards due to the mis-classification error.

4.3 Classification Results

Our classification results are based on a random forest classification algorithm. To train the algorithm, we take all 667,774 disclosed-sponsored posts and randomly select a sample of equal size from the 3,767,996 true organic posts, which we identified based on the procedure laid out in Section 4.1.⁹ We allocate 20% of these organic and sponsored posts to a hold-out sample which we use to assess the performance of the classifier. The remaining undisclosed posts constitute the set of posts we aim to classify. We refer to them as “maybe organic” posts because they could be either truly organic posts or undisclosed sponsored posts. We provide details regarding the size of our training and hold-out samples in the top panel of Table 2.

We first verify that our classifier performs well using standard metrics of model performance. We report accuracy, precision, recall and “area under the ROC curve” evaluated based on the hold-out sample which was not used in estimation. Overall the classifier performs very well with an AUC-ROC value of over 97%. We re-iterate that we are able to achieve very high performance despite removing the top bigrams that we used to compose the training sample. It appears that

⁹We apply several pre-processing steps to all posts which we describe in Appendix C.

Sample Construction:				
	# Posts	Training	Hold-out	
Sponsored	667,774	534,432	133,342	
Non-disclosed	100,846,515			
True organic (based on top bigrams)	3,767,996	534,432	133,342	
Maybe organic	97,078,519			
Total Brand-related Posts	101,514,289			

Classifier Performance:				
	Accuracy	Precision	Recall	AUC-ROC
Random Forest Classifier	92.71%	93.5%	91.8%	97.87%
	Prob. Classified as ...			
	Organic	Sponsored		
True Organic	0.936	0.064		
True Sponsored	0.082	0.918		

Classification Results:		
	Organic	Sponsored
Maybe Organic Posts	0.791	0.209
Adjusted Prediction	0.830	0.170
Including True Organic Posts	0.836	0.164

Table 2: **Classification of Undisclosed Sponsored Posts.**

loosing some of the posts’ content due to the sample selection step does not materially affect our ability to classify posts based on the remainder of the posts’ text. We also report class-specific error rates in the middle panel of Table 2. We find that out of all true organic posts in the hold-out sample, 93.6% are correctly classified as organic. We achieve a similar success rate with regards to sponsored posts, of which 91.8% are correctly classified.

Finally, we apply the classifier to the remaining posts without a sponsorship disclosure. We find that 79.1% of posts are classified as organic whereas 20.9% of posts are classified as (undisclosed) sponsored posts. Next, we apply the adjustment procedure laid out in Section 4.2 that corrects for the known mis-classification error of our model. Based on Equation (1) we find that 17% of maybe-organic posts constitute undisclosed sponsored posts. More concretely, we input the observed mis-classification rates into Equation (1) which yields $Pr(O^*) = (0.791 - 0.082) / (0.936 - 0.082) = 0.83$. When we add the sample of true organic posts, we find that out of all posts without disclosure, 16.4% are undisclosed sponsored posts. Given the large number of posts without disclosure, 16.4% of undisclosed posts translates into a total of 16.5 million undisclosed sponsored posts compared to only 667,774 disclosed sponsored posts. Therefore, the large majority of 96.1% of influencer commercial activity is undisclosed .

Year	Mean Across	Percentiles					Number of Brands
	Brands	P5	P25	Median	P75	P95	
2014	0.989	0.944	0.996	0.999	1.000	1.000	259
2015	0.986	0.922	0.993	0.998	1.000	1.000	260
2016	0.981	0.911	0.986	0.996	0.999	1.000	263
2017	0.972	0.869	0.980	0.991	0.997	1.000	264
2018	0.972	0.916	0.973	0.990	0.996	1.000	267
2019	0.969	0.887	0.967	0.988	0.995	0.999	267
2020	0.958	0.865	0.969	0.988	0.995	0.999	267
2021	0.956	0.851	0.958	0.980	0.993	0.997	268

Table 3: **Distribution of Annual Brand-level Non-Disclosure Rates.** Each row reports the mean and percentiles of the distribution of the share of non-disclosed posts (out of all sponsored posts) across brands.

As discussed in Section 4.1, we use ChatGPT to construct a training sample as a robustness check. The GPT-based classification predicts a 98.0% share of undisclosed posts among all sponsored posts relative to a 96.1% share when using our preferred approach. Therefore, the GPT-based classification approach validates our key finding that a large number of sponsored posts is undisclosed.¹⁰ We provide a detailed description of this alternative classification approach in Appendix F.

5 Determinants of Undisclosed Content

Our data contains a broad set of brands from a wide variety of industries, ranging from “traditional” brick-and-mortar retail to fashion and fintech. We observe the brands’ posts over an 8-year period from 2014 to 2021, during which disclosure regulation tightened. In this section, we describe the distribution of brand-level disclosure, study how disclosure evolves over time, and characterize the brands and influencers that are more likely to disclose sponsored content. We conduct most of the analysis in this section at the brand/year level and compute non-disclosure rates by dividing the number of undisclosed sponsored posts by the number of all sponsored posts for a given brand and year.¹¹

¹⁰We also find that ChatGPT-based predictions have a higher classification error when labeling disclosed sponsored posts (the only set of posts for which their true sponsorship status is observed in the data). We therefore regard our classification results as preferable to the classification based on ChatGPT.

¹¹We calculate the probability of a post being truly organic conditional on being classified as organic $Pr(O^*|O)$ using the standard Bayesian updating formula. We calculate the total number of undisclosed sponsored posts by summing up the posterior probabilities.

Dependent Variable	(1)	(2)	(3)
	Non-Disclosure Share	Non-Disclosure Share	Non-Disclosure Share
Brand Founded After 2000	0.016*** (0.004)	0.017*** (0.006)	0.014*** (0.005)
# Brand Account Followers (Log-Transformed)	0.008*** (0.002)	0.009*** (0.002)	0.010*** (0.002)
Complier Dummy × After-2019 Dummy			-0.321*** (0.101)
Year FE	Yes	Yes	Yes
Industry FE	No	Yes	Yes
Joint F-stat Industry FE		923	826.3
Observations	2,105	2,105	2,105
Brands	268	268	268

Table 4: **Non-Disclosure & Brand Characteristics.** The unit of observation is a brand/year combination. Standard errors are clustered at the brand level.

5.1 Evolution over Time

Table 3 shows the distribution of brand-level non-disclosure rates by year. For the vast majority of brands, non-disclosure rates are very high throughout the sample period, and while there is some heterogeneity across brands, the distribution is quite tight. The average brand fails to disclose between 99 and 96 percent of their sponsored content. Even the 5th percentile brand in 2021 fails to disclose over 85 percent of their sponsored posts. Disclosure is going up throughout the sample period, with the distribution of non-disclosure probabilities shifting left. At all points in the distribution, non-disclosure falls between 2014 and 2021. Non-disclosure for the 5th percentile brand drops by nearly 10 percentage points.¹² Overall, the improvement is modest, with the average non-disclosure rate dropping by only 3.3 percentage points from 0.989 to 0.956 between 2014 and 2021. Therefore, despite stricter disclosure rules and enforcement over time, the change in the disclosure behavior among the wide set of brands in our sample is small.

5.2 Brand Characteristics

To study which brand characteristics correlate with disclosure, we regress the non-disclosure share on year fixed effects and various brand characteristics. In column (1) we include the number of followers of a brand’s account as well as a dummy for whether the brand was founded after 2000.¹³

¹²We condition our sample on brands whose Twitter accounts were created before 2014, but some brands do not have any sponsored activity (disclosed or undisclosed) in the first years of our sample. Therefore, we observe a slight change in the number of brands in Table 3. When using a balanced set of brands, results are the qualitatively very similar.

¹³We used Pitchbook and Crunchbase to determine the founding years of brands. When there were inconsistencies between both sources, we retrieved the founding year from the brand’s official website or Wikipedia.

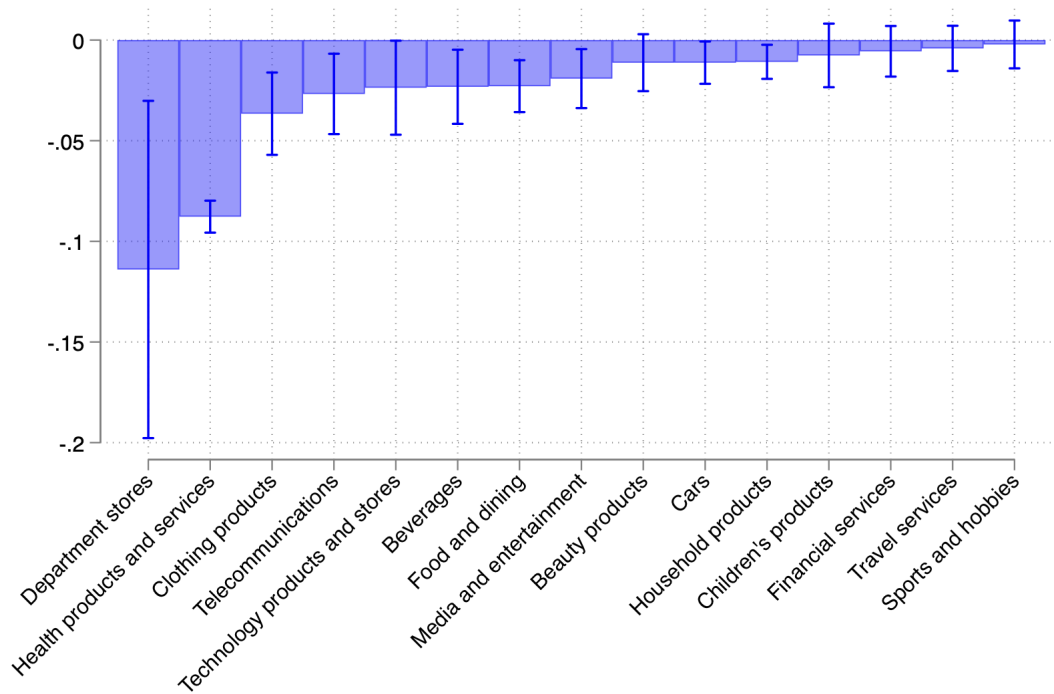


Figure 2: **Industry Differences in the Share of Undisclosed Posts.** The graph plots the estimated values of the industry fixed effects from the regression results reported in column (3) of Table 4. The omitted category is “Home design and decoration”, the category with the largest fixed effect.

We find that brands with a stronger presence on social media, as measured by their follower count, as well as younger brands are characterized by a significantly higher share of non-disclosed posts.¹⁴ We include category fixed effects using the 16 brand categories defined in Lovett et al. (2014) in column (2) of Table 4. Results for the two brand characteristics look similar .

To analyze category difference in more detail, we plot the distribution of category fixed effects in Figure 2. The figure shows that most categories have high non-disclosure levels that are quite similar to the baseline category with the highest fixed effect, “home design and decoration”. There are two main exceptions: “department stores,” and “health products and services,” which have significantly higher disclosure rates than the other categories.¹⁵ Clothing product brands also have a lower fixed effect estimate, though it is not statistically different from most other categories. Overall, we interpret the brand level results to indicate that more traditional brands, i.e. older brands with less social media presence in industries such as department stores, are more likely to disclose sponsored content.

Finally, we exploit a discrete change in the regulation of sponsored content. In December 2019,

¹⁴In Table A4 in the appendix we report results from a regression with separate decade dummies and find that most of the difference in disclosure behavior is due to young brand founded after 2000 disclosing less.

¹⁵We note that the second category includes only Walgreens.

	Organic	Sponsored	
		Undisclosed	Disclosed
Followers	14,148	27,200	49,193
Likes per Follower	0.054	0.049	0.009
Replies per Follower	0.008	0.005	0.001
Retweets per Follower	0.019	0.017	0.004

Table 5: **Average Reach and Engagement for Different Types of Posts.**

the FTC mandated that the disclosure hashtag needed to appear at the beginning of a post. As we show in Appendix D, most sponsored posts tended to include the disclosure hashtags towards the end of the post. We find that only 10 brands (3.7% of our sample) display a sharp shift in disclosure position after 2019. We refer to these brands as “compliers”. In column (3) of Table 4 we add an interaction of a complier dummy with a post-2019 dummy to our regression specification. We find that the brands that start displaying the sponsored hashtag earlier in the post after 2019 also start disclosing a larger share of content. The effect is large in magnitude relative to the general time trend and disclosure differences related to brand characteristics. In line with our other brand-level results, the complier brands tend to be older brands in more traditional industries such as BestBuy, Footlocker, and Disney. In Appendix E we provide additional details on how we identify compliers, as well as their characteristics and behavior over time.

5.3 Followers and Engagement

In this section, we analyze how different types of posts vary in terms of the accounts from which they originate and in terms of their engagement. In Table 5 we document that the average number of followers of accounts posting disclosed sponsored posts is almost twice as large as the follower numbers of accounts associated with undisclosed sponsored posts. Organic posts tend to come from accounts with the lowest follower numbers among the three types of posts. This pattern is intuitive and lends support to our classification strategy as we would expect undisclosed sponsored posts to come from accounts with a larger reach than organic posts. In terms of engagement we find that undisclosed posts garner more engagement per follower across all three engagement metrics relative to disclosed sponsored posts. While this might suggest that disclosure leads to lower engagement, consistent with the findings by Karagür et al. (2022), we need to be careful with a causal interpretation because these posts come from different accounts.

We also analyze differences in the content of posts (length of the post, number of hashtags, etc.), but find only minor differences between disclosed and undisclosed sponsored posts. We provide more details on these additional results in Appendix D.

6 Conclusion

In this paper, we quantify the importance of undisclosed sponsored content on Twitter based on a unique data set of over 100 million posts and a novel classification method. We find that undisclosed sponsored posts are ubiquitous with 96% of all sponsored content being undisclosed. The share of undisclosed content decreases only slightly over time despite stronger regulation and only a small share of brands responded to a regulatory change that mandated disclosure at the beginning of a post. These results highlight that tightening regulation has had a very modest impact on the disclosure of sponsored content. We also find that young brands with a larger social media following are less likely disclose sponsored content. These kind of brands will likely rely more heavily on influencers and therefore disclosure rates might remain low in the future.

References

- EASA (2018). *EASA Best Practice Recommendation on Influencer Marketing*. <https://www.easa-alliance.org/wp-content/uploads/2018/04/EASA-BPR-ON-INFLUENCER-MARKETING-2023.pdf>. Accessed in 2023.
- Ershov, Daniel and Matthew Mitchell (2020). “The effects of influencer advertising disclosure regulations: Evidence from instagram”. In: *Proceedings of the 21st ACM Conference on Economics and Computation*, pp. 73–74.
- Fainmesser, Itay P and Andrea Galeotti (2021). “The market for online influence”. In: *American Economic Journal: Microeconomics* 13.4, pp. 332–72.
- FTC (Dec. 2015a). *Enforcement policy statement on deceptively formatted advertisements*. <https://www.ftc.gov/news-events/news/press-releases/2015/12/ftc-issues-enforcement-policy-statement-addressing-native-advertising-deceptively-formatted>. Accessed in 2023.
- (Dec. 2015b). *Native Advertising: A Guide for Businesses*. <https://www.ftc.gov/business-guidance/resources/native-advertising-guide-businesses>. Accessed in 2023.
- (2017a). *FTC Staff Reminds Influencers and Brands to Clearly Disclose Relationship*. <https://www.ftc.gov/news-events/news/press-releases/2017/04/ftc-staff-reminds-influencers-brands-clearly-disclose-relationship>. Accessed in 2023.
- (2017b). *FTC’s Endorsement Guides*. <https://www.ftc.gov/business-guidance/resources/ftcs-endorsement-guides>. Accessed in 2023.
- (2019). *Disclosures 101 for Social Media Influencers*. <https://www.ftc.gov/business-guidance/resources/disclosures-101-social-media-influencers>. Accessed in 2023.
- Geysler, Werner (2018). *Influencer Contract Template*. <https://influencermarketinghub.com/influencer-contract-template/>. Accessed in 2023.
- (2019). *The State of Influencer Marketing 2019 : Benchmark Report*. <https://influencermarketinghub.com/influencer-marketing-2019-benchmark-report/>. Accessed in 2023.
- Gong, Shiyang, Juanjuan Zhang, Ping Zhao, and Xuping Jiang (2017). “Tweeting as a marketing tool: A field experiment in the TV industry”. In: *Journal of Marketing Research* 54.6, pp. 833–850.
- He, Sherry, Brett Hollenbeck, and Davide Proserpio (2022). “The market for fake reviews”. In: *Marketing Science* 41.5, pp. 896–921.
- Huang, Yufeng and Ilya Morozov (2022). “Video Advertising by Twitch Influencers”. In: *Available at SSRN 4065064*.
- Hughes, Christian, Vanitha Swaminathan, and Gillian Brooks (2019). “Driving brand engagement through online social influencers: An empirical investigation of sponsored blogging campaigns”. In: *Journal of Marketing* 83.5, pp. 78–96.
- Karagür, Zeynep, Jan-Michael Becker, Kristina Klein, and Alexander Edeling (2022). “How, why, and when disclosure type matters for influencer marketing”. In: *International Journal of Research in Marketing* 39.2, pp. 313–335.

- Kim, Seungbae, Jyun-Yu Jiang, and Wei Wang (2021). “Discovering undisclosed paid partnership on social media via aspect-attentive sponsored post learning”. In: *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pp. 319–327.
- Lanz, Andreas, Jacob Goldenberg, Daniel Shapira, and Florian Stahl (2019). “Climb or jump: Status-based seeding in user-generated content networks”. In: *Journal of Marketing Research* 56.3, pp. 361–378.
- Li, Nan, Avery Haviv, and Mitchell J Lovett (2021). “Digital Marketing and Intellectual Property Rights: Leveraging Events and Influencers”. In: *Available at SSRN 3884038*.
- Lovett, Mitchell, Renana Peres, and Ron Shachar (2014). “A data set of brands and their characteristics”. In: *Marketing Science* 33.4, pp. 609–617.
- Luca, Michael and Georgios Zervas (2016). “Fake it till you make it: Reputation, competition, and Yelp review fraud”. In: *Management Science* 62.12, pp. 3412–3427.
- Mayzlin, Dina, Yaniv Dover, and Judith Chevalier (2014). “Promotional reviews: An empirical investigation of online review manipulation”. In: *American Economic Review* 104.8, pp. 2421–2455.
- Michaelsen, Frithjof, Luena Collini, Cécile Jacob, Catalina Goanta, Sara Elisa Kettner, Sophie Bishop, Pierre Hausemer, Christian Thorun, and Sevil Yesiloglu (2022). “The impact of influencers on advertising and consumer protection in the Single Market”. In: *Study requested by the IMCO committee*.
- Mitchell, Matthew (2021). “Free ad (vice): Internet influencers and disclosure regulation”. In: *The RAND Journal of Economics* 52.1, pp. 3–21.
- Pei, Amy and Dina Mayzlin (2022). “Influencing social media influencers through affiliation”. In: *Marketing Science* 41.3, pp. 593–615.
- Rajaram, Prashant and Puneet Manchanda (2020). “Video influencers: Unboxing the mystique”. In: *arXiv preprint arXiv:2012.12311*.
- Silva, Márcio, Lucas Santos de Oliveira, Athanasios Andreou, Pedro Olmo Vaz de Melo, Oana Goga, and Fabrício Benevenuto (2020). “Facebook ads monitor: An independent auditing system for political ads on facebook”. In: *Proceedings of The Web Conference 2020*, pp. 224–234.
- theinfluencers.com.au (2021). *How Much Creative Control Should You Give Influencers? [Hint: You’ll Be Surprised]*. <https://www.theinfluencers.com.au/articles/why-brands-should-be-giving-influencers-creative-freedom>. Accessed in 2023.
- Valesia, Francesca, Davide Proserpio, and Joseph C Nunes (2020). “The positive effect of not following others on social media”. In: *Journal of Marketing Research* 57.6, pp. 1152–1168.
- Yang, Jeremy, Juanjuan Zhang, and Yuhan Zhang (2021). “First law of motion: Influencer video advertising on tiktok”. In: *Available at SSRN 3815124*.

APPENDIX

A Brand Selection

We select brands using the following procedure: We first generate a list of candidate brands that have a large following on Twitter and that have at least some amount of disclosed influencer activity. We start by collecting all English language original tweets containing the hashtags "#ad" or "#sponsored" tweeted between 2021/01/01 and 2021/12/31, which yields a total of 2,753,580 tweets. Next, we generate a list of accounts by extracting all mentions (defined as @ followed by the name of a Twitter account) from these posts. To identify brands from the set of all mentioned accounts, we first select only verified accounts, since well-known brand accounts are verified by Twitter.¹⁶ We also exclude accounts that joined Twitter after the start of our sample, i.e. in 2014 or later. After these steps, we end up with a list of 16,604 accounts. We then select "important" accounts by retaining accounts with at least 500,000 followers and at least 10 sponsored tweet mentions in 2021. Both of these thresholds correspond to roughly the 80th percentile of their respective distributions (among verified accounts mentioned in sponsored tweets), and lead us to retain 774 accounts.

Next, we manually screen each account by analyzing its Twitter or Google business profile to check if it corresponds to a brand account or a personal account (e.g., celebrity, online influencer). In total, 568 (73.4%) are brand accounts. We retain 426 of these brands which are headquartered in the US, and 56 that are headquartered abroad but are highly active in the US (i.e., Nintendo, Chanel). We exclude several categories of accounts which do not constitute traditional brands, such as non-commercial entities (e.g., NASA), sports leagues and clubs, non-profit agencies, and news and media organizations.

Finally, we excluded an additional 8 brands. Office365: its private account setting prevents data collection via the Twitter API; Etsy and eBay: the number of posts is substantially higher than for other brands, making it difficult to collect and store all posts; Puma Football: data corruption; HBO Max: the product was officially launched only in 2020, but the Twitter handle appears to have been created earlier; Nordstrom, Amazon, and Macy's: all three brands experience a very large and temporary activity spike in 2016-2017. The final list contains 268 brands.

B Bigram Extraction

Before extracting the bigrams displayed in Table 1, we first implement several pre-processing steps. We convert the text to lowercase and remove elements such as special characters, numbers, hashtags, URLs, and punctuation. We also replace common Twitter abbreviations such "lol" or "wtf" with the corresponding full phrases and remove stop words (such as "and", "to", "in", ...). We then compute the frequency of all bigrams for each brand separately and maintain the top 500 most

¹⁶<https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts>.

	Organic	Sponsored	
		Undisclosed	Disclosed
# Hashtags (excl. #Ad and #Sponsored)	1.06	1.42	1.41
Text length	112.81	118.10	128.49
Brand Mention Position	0.48	0.48	0.45
Disclosure Position	n/a	n/a	0.63
Disclosure in the First Half Dummy	n/a	n/a	0.29

Table A1: **Characteristics of Different Types of Posts.** Brand Mention Position and Disclosure Position are calculated by dividing the position at which the brand / disclosure appears by the length of the post.

frequent bigrams. Finally, we aggregate these frequencies across all brands to obtain our final list of bigram occurrence frequencies.

C Text-Processing Pipeline for Post Classification

We apply several pre-processing steps to the text of posts before applying our classification algorithm. We remove disclosure hashtags (#ad and #sponsored) because we want the classifier to recognize sponsored posts based on their content and not the disclosure itself. We also remove all brand mentions as well as the '#' and '@' symbols for all remaining hashtags and mentions. We replace any link addresses within the text with the term "http" so that the presence of a link can be used for classification. We remove the sponsored and organic bigrams that were used to form our sample of "true organic" posts so that these words are not "re-used" by the classifier. Finally, we remove stop words and special characters.

D Post Characteristics: Additional Results

In the Table A1 we analyze difference in the content of different types of posts. We note that none of these characteristics was used for classification. We find that the only dimension along which disclosed and undisclosed posts differ substantially is the length of the post with disclosed post being somewhat longer. In terms of the number of hashtags used and the position within the post at which the brand is mentioned, the two types of sponsored posts are very similar.

We also investigate the position of the disclosure hashtag used in sponsored posts. We compute the position of the disclosure hashtag as the position of the hashtag symbol relative to the length of the posts, so that the position is equal to 0.5 if the hashtag appears exactly in the middle of the post. We find that the disclosure hashtag appears on average relatively late in the post with an average disclosure position of 0.63. The disclosure hashtag appears in the first half of the post in only 29% of all posts.

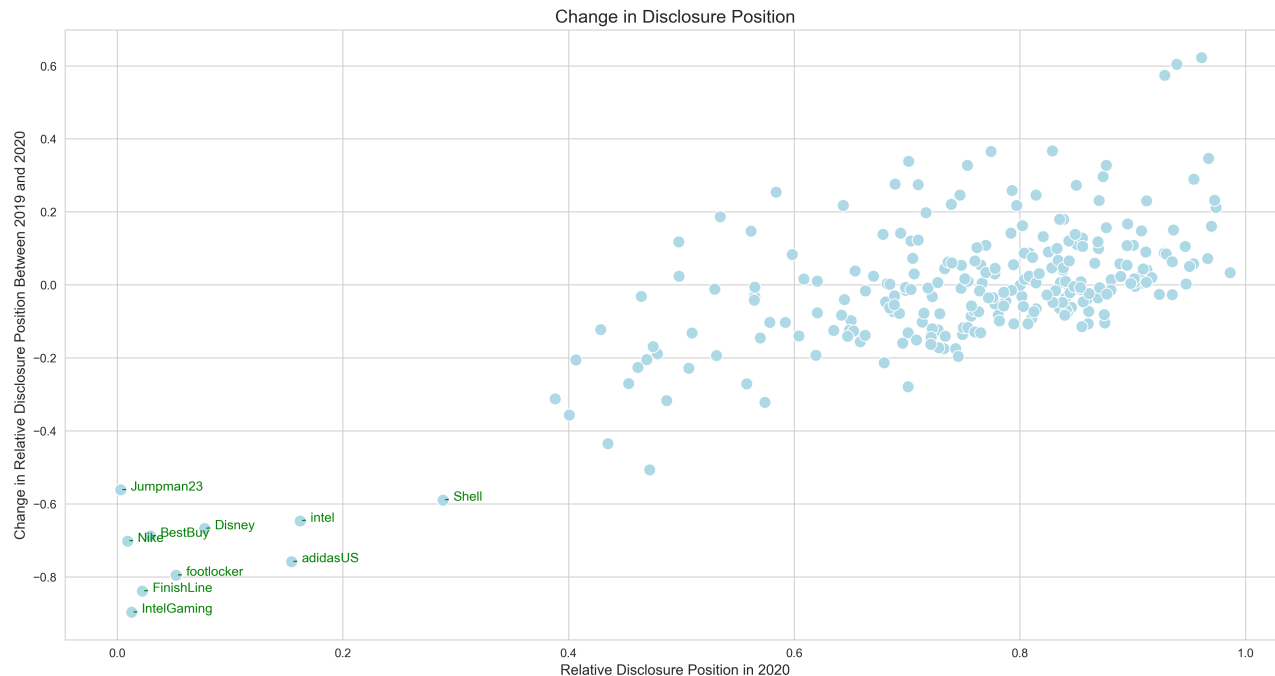


Figure A1: Change in Disclosure Position After the 2019 Regulatory Change.

E Disclosure Regulation & Compliers

As we discussed in Section 2, the FTC mandated the disclosure hashtag to appear at the beginning of a post in December 2019. To analyze whether some brands changed their behavior in response to this change in regulation, we compute the relative position of the disclosure hashtag by dividing the location of the disclosure hashtag by the total length of the post (as described in Appendix D). We plot the change in disclosure position in between 2019 and 2020 against the disclosure position level in 2020 in Figure A1. We would expect complier brands to have a disclosure position close to zero in 2020 and potentially a large change in disclosure position relative to the previous year. The scatterplot in Figure A1 shows a cluster of brands in the bottom-left corner which have an average disclosure position close to the beginning of the post as well as a large change in behavior relative to 2019. We label the 10 brands in this cluster as “compliers”. These brands are labeled in green in the graph. The set of compliers contains mostly larger, older, and more traditional brands such as BestBuy, Disney, and Footlocker.

We also run a regression of disclosure share on year and industry fixed effects as well as complier dummies interacted with each year of the sample. This regression specification mirrors the one in column (4) of Table 2, but allows us to analyze in more detail when the complier brands change their disclosure behavior. Results are reported in column (2) of Table A4 showing that compliers change their behavior in 2020 and continue to disclose a large share of sponsored posts in 2021.

Sample Construction:				
	# Posts	Training	Hold-out	
Sponsored	667,774			
Non-disclosed	100,846,515			
True organic (based on ChatGPT)	13,280	5,381	1,318	
True sponsored (based on ChatGPT)	6,699	5,381	1,318	
Maybe organic	100,833,235			
Total Brand-related Posts	101,514,289			

Classifier Performance:				
	Accuracy	Precision	Recall	AUC-ROC
Random Forest Classifier	65.50%	66.04%	63.88%	71.57%

	Prob. Classified as ...	
	Organic	Sponsored
True Organic	0.671	0.329
True Sponsored	0.361	0.639

Classification Results:		
	Organic	Sponsored
Maybe Organic Posts	0.574	0.426
Adjusted Prediction	0.686	0.314
Including True Organic Posts	0.686	0.314

Table A2: **Classification Results for ChatGPT-based Classifier.**

F ChatGPT Sponsorship Detection

We implement an alternative classification approach where we construct a training sample of sponsored and organic posts using OpenAI’s ChatGPT. For 20,000 randomly selected posts without disclosure hashtags,¹⁷ we query ChatGPT about whether they are sponsored or organic. Then, we train a classifier on this data set and use it to label the remaining posts. We re-iterate that our preferred classification approach assumes that disclosed and undisclosed sponsored posts are based on the same dictionary and that the set of posts we isolate based on the top bigrams constitute organic posts. By contrast, we now use ChatGPT’s pre-trained model to identify organic and sponsored posts from the set of posts without disclosure and thus avoid both assumptions. Instead, we rely on ChatGPT’s ability to correctly label posts that we can then use as a training sample.¹⁸

In more detail, we proceed as follows: Using the ChatGPT 3.5 API (text-davinci-3), we query

¹⁷We drop a small number of posts because 12 posts are empty after pre-processing and 9 posts fail to generate a valid response from ChatGPT.

¹⁸An alternative approach would be to directly let ChatGPT label all posts without a sponsorship disclosure. This approach is not feasible because our sample contains over 100 million posts without disclosure.

Label Based on...		
Our Classifier	GPT-Based Classifier	Share of Total Posts
Organic	Organic	0.482
Organic	Sponsored	0.309
Sponsored	Organic	0.064
Sponsored	Sponsored	0.145

Table A3: **Comparison of GPT to our Preferred Classifiers for Posts without Disclosure.** The table is based on all posts without disclosure except for those that were categorized as “true organic” posts and therefore used in the training sample of either classifier.

GPT with the prompt “Decide whether the following Tweet is sponsored with more than 50% probability: {tweet}”, where {tweet} is replaced with the text of a tweet. The responses we receive from ChatGPT are either “Yes, this tweet is likely sponsored,” or “No, this Tweet is not sponsored with more than 50% probability.” These responses are then transformed into a dummy variable. We set the query parameters to minimize randomness in the responses (“temperature” = 0). We also do not include any additional conditions to the prompt, which attempt to place ChatGPT in a role (i.e., “As a customer:...” or “As a regulator:...”). After obtaining the “GPT ground truth” for our sample of posts, we use these to train a classifier. To keep this alternative classification approach as comparable as possible with our approach, we employ a random forest classifier and balance the observation across classes in the training sample. We apply the same mis-classification correction to the GPT-based classifier as we do for the main classifier (see Section 4.2). We present the classification results for this classifier in Table A2 which can be directly compared to our main classification results in Table 2.

We find that 31.4% of all posts without disclosure are classified as sponsored by the GPT-based classifier compared to 17% for our main classifier. The larger number of undisclosed sponsored posts leads to a share of undisclosed posts among all sponsored posts of 98.0% compared to 96.1% when using our classifier. Therefore, the alternative classification validates our main finding that a large share of sponsored posts is undisclosed. If anything, our classification leads to a more conservative, i.e. smaller, estimate of the share of undisclosed sponsored posts. We also assess the degree to which the two classifiers agree in Table A3. Our preferred classification approach and the GPT-based classifier agree on approximately 63% of labels (0.48 organic/organic + 0.15 sponsored/sponsored). The main disagreements come from the GPT-based classifier labeling substantially more undisclosed posts as sponsored. 31% of posts in this sample are labeled as organic by our classifier and sponsored by the GPT classifier. Only a small share of posts is labeled as sponsored by our approach but not by ChatGPT.

Finally, we assess the ability of both classifiers to correctly label disclosed sponsored posts. This set of posts is the only set of posts for which their correct label is known (because posts without disclosure could be either organic or undisclosed sponsored posts). Therefore, disclosed sponsored

posts allow us to assess the performance of different classifiers by analyzing how many posts are correctly labeled by each classifier. We find that our approach classifies 92% of disclosed sponsored posts correctly, whereas the GPT-based classifier only labels 73% of posts correctly. Due to its superior performance we regard our classifier as preferable to the classification based on ChatGPT.

G Additional Tables

Dependent Variable	(1) Non-Disclosure Share	(2) Non-Disclosure Share
Brand Founding Decade = 1950s	0.000 (0.013)	
Brand Founding Decade = 1960s	-0.037* (0.020)	
Brand Founding Decade = 1970s	-0.010 (0.015)	
Brand Founding Decade = 1980s	-0.001 (0.007)	
Brand Founding Decade = 1990s	-0.001 (0.008)	
Brand Founding Decade = 2000s	0.012** (0.006)	
Complier \times (Year = 2015)		-0.041 (0.043)
Complier \times (Year = 2016)		-0.051 (0.060)
Complier \times (Year = 2017)		-0.059 (0.062)
Complier \times (Year = 2018)		-0.075 (0.075)
Complier \times (Year = 2019)		-0.061 (0.046)
Complier \times (Year = 2020)		-0.389*** (0.099)
Complier \times (Year = 2021)		-0.262** (0.118)
Industry FE	Yes	Yes
Year FE	Yes	Yes
Other Controls	Yes	Yes
Observations	2,105	2,105
Brands	268	268

Table A4: **Additional Regression Results: Decade-specific Dummies & Yearly Effects for Compliers.** The unit of observation is a brand/year combination. Standard errors are clustered at the brand level. We include all brands founded between 2000 and 2014 as part of the “Founding Decade = 2000” category.