CESifo WORKING PAPERS

10743
2023

Original Version: November 2023

This Version: April 2024

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

CESifo *Working Paper No. 10743*

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
d.ershov@ucl.ac.uk

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

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

This draft: April 22, 2024
We thank Brett Hollenbeck and Ilya Morozov, seminar participants at ESADE, Goethe University Frankfurt, Imperial College London and Helsinki GSE, Yelp, and conference participants at the UK Digital Economics Workshop and the CESifo digitization area meeting for helpful comments. 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 leverages private individuals to promote brands, for which they are compensated. 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 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 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 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 (Geyser (2019)). Moreover, the platform has been an early adopter of influencer advertising, thus allowing us to study its evolution over a relatively long time period - from 2014 to 2021. Twitter allows us to collect particularly rich data through their data-API[2] 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 enforcement of disclosure regulation over time, the share of undisclosed content only decreases

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

[2] Until 2023, the Twitter data-API allowed researchers 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.

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

slightly from 98.4% in 2014 to 94.4% 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 US regulations do 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 empirical 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 social media influencer markets (e.g., Lanz et al. (2019) and Valsesia et al. (2020)). More closely related to our work are Ershov and Mitchell (2020) who study the impact of disclosure regulation on influencers' disclosure decisions and content production as well as Karagür et al. (2022) and Bairathi and Lambrecht (2024) who study the impact of disclosure on engagement.[4] To the best of our knowledge, none of the previous papers detecting hidden sponsorship deal with a contaminated training sample or adjust for classification error, which constitute the main methodological contributions of this paper. We also contribute to the literature on influencer marketing by basing our analysis on a large and comprehensive brand-level data set rather than a sample of influencers as in Ershov and Mitchell (2020) or Bairathi and Lambrecht (2024). This data allow us to quantify non-disclosure rates for a representative set of important brands and to study the determinants of disclosure as well as the impact of regulatory changes.

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, which should raise regulatory concerns.

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

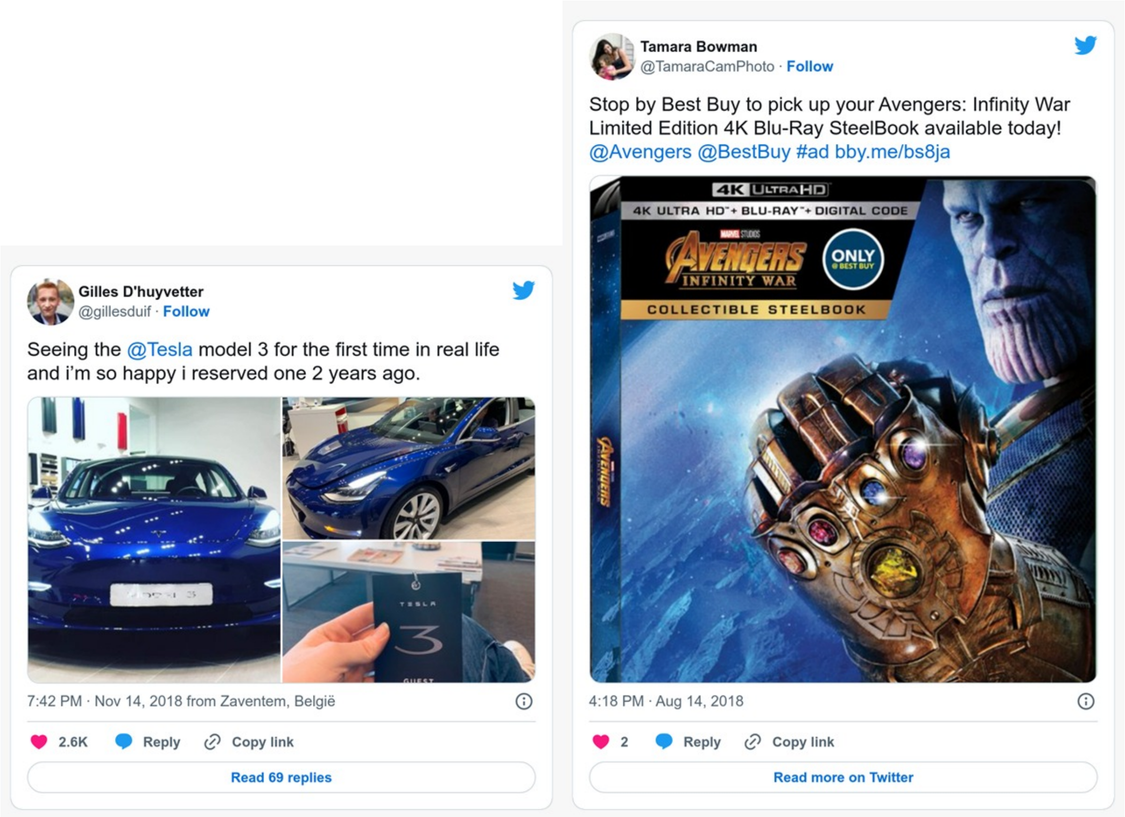
[4] There is also a related literature in computer science on the identification of undisclosed sponsored content, which includes Kim et al. (2021) and Silva et al. (2020).

or negative language (Geyser (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 active in the US 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.[5]

[5] The 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).



Gilles D'huyvetter
@gillesduif · Follow

Seeing the @Tesla model 3 for the first time in real life and i'm so happy i reserved one 2 years ago.

7:42 PM · Nov 14, 2018 from Zaventem, België

Tamara Bowman
@TamaraCamPhoto · Follow

Stop by Best Buy to pick up your Avengers: Infinity War Limited Edition 4K Blu-Ray SteelBook available today! @Avengers @BestBuy #ad bby.me/bs8ja

4:18 PM · Aug 14, 2018

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

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.[6] We first find a set of brands that have 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.[7] 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 101,510,557 posts: 773,393 posts with a sponsorship disclosure such as #ad or #sponsored[8] and 100,737,164 posts without disclosure. Although secondary to 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 contain different 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

[6] This research API has been discontinued in early 2023.

[7] Because we use a low threshold of at least 10 sponsored posts we are likely to capture most brands that engage in sponsored activity.

[8] In Appendix B we provide details on how we define a valid sponsorship disclosure based on existing FTC rules.

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	holiday season	dont miss
oh my god	limited edition	coming soon
bring back	get free	brand new
social media	dont miss	weve got
years ago	twitter party	tell us
looking forward	holiday gift	happy birthday
every time	giving away	behind scenes
feel like	blog today	stay tuned
anyone else	get ready	make sure
would like	supplies last	new york
dont know	try new	take look
well done	love new	limited edition
new video	perfect gift	modified tweet
shake my head	dont forget	new year

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

two consecutive words) in both types of posts in Table 1.[9] We find that the set of top bigrams for sponsored posts predominantly contain commercial language such as “available via” or “use code”, whereas posts without disclosure are more likely to contain casual and conversational language. 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. We also note that the non-commercial nature of the top bigrams in posts without disclosure suggests that organic posts constitute the majority of posts without disclosure.

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.

[9] Whenever 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 C we provide additional detail on how we compute the list of top bigrams.

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 misclassified by a content-based classifier. In the next two subsection 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.[10]

We compile a list of "true organic" posts by selecting posts that do not contain any of the top 20 sponsored bigrams and that contain at least one of the top 20 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.

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

[10] Our classifier results label 19% 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 $Pr(O|S^*)$. Using the fact that $Pr(S^*) = 1 - Pr(O^*)$ and re-arranging the equation above gives us the formula for the share of organic posts that accounts for mis-classification:

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

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.[11] To train the algorithm, we take all 773,393 disclosed-sponsored posts and randomly select a sample of equal size from the 3,776,387 true organic posts, which we identified based on the procedure laid out in Section 4.1.[12] 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 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 96%. 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, 91.8% are correctly classified

[11] Random forest performs best among a set of different classifiers. It yields an accuracy (AUC) of 0.905 (0.964) compared to XGBoost’s 0.874 (0.947) and Naive Bayes’ 0.842 (0.914).

[12] We apply several pre-processing steps to all posts which we describe in Appendix D.

Sample Construction:				
	# Posts	Training	Hold-out	
Sponsored	773,393	618,079	155,314	
Non-disclosed	100,737,164			
True organic (based on top bigrams)	3,776,387	618,079	155,314	
Maybe organic	96,960,777			
Total Brand-related Posts	101,510,557			
Classifier Performance:				
	Accuracy	Precision	Recall	AUC-ROC
Random Forest Classifier	90.52%	91.62%	89.20%	96.40%
	Prob. Classified as ...			
	Organic	Sponsored		
True Organic	0.918	0.082		
True Sponsored	0.108	0.892		
Classification Results:				
	Organic	Sponsored		
Maybe Organic Posts	0.760	0.240		
Adjusted Prediction	0.804	0.196		
Including True Organic Posts	0.812	0.188		

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

as organic. We achieve a similar success rate with regards to sponsored posts, of which 89.2% are correctly classified.

We then apply the classifier to the remaining posts without a sponsorship disclosure. We find that 76% of posts are classified as organic whereas 24% 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[13] and add the sample of true organic posts which yields a rate of hidden sponsored posts of 18.8% among all posts without disclosure. This rate of non-disclosure translates into a total of 18.97 million undisclosed sponsored posts compared to only 773,393 disclosed sponsored posts. Therefore, 96.1% of influencer commercial activity is undisclosed.

4.4 Robustness Checks

We test the robustness of our classifier in a variety of ways: First, we look at the correlation between our classification and post sentiment, which is not used as an input into classification. Intuitively,

[13] We input the observed mis-classification rates into Equation (1) which yields $Pr(O^*) = (0.760 - 0.108)/(0.918 - 0.108) = 0.804$.

influencers are unlikely to express negative sentiments about brands or products they are paid to promote. After using a sentiment classifier on our posts, we find a strong positive correlation between predicted sponsorship probability and positive sentiment. Second, we look at duplicate posts (i.e., posts with the same words) which are more likely to be sponsored as they likely reflect campaigns where brands directed a precise script to multiple influencers. We find duplicate posts with positive sentiment to have a higher average sponsorship probability than unique posts with similar sentiment.[14]

Third, we use an alternative classification method based on ChatGPT. We find that it has a high agreement rate with our classifier, and results in similar predicted undisclosed sponsorship rate of 97.4%. These results suggest that both classifiers detect similar patterns in the content of sponsored and organic posts, and that our classifier is more conservative than ChatGPT.[15] Finally, we repeat our classification approach but use a random sample of posts without disclosure labeled as "true organics." Since some of these organic training posts could be undisclosed sponsored posts, this classifier will under-predict the prevalence of sponsored posts in the data, and can provide a lower bound to the share of undisclosed sponsored posts. We find that even at the lower bound, 81.7% of all sponsored posts are undisclosed.

Appendices E, F, and G provide details on these robustness checks.

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.[16]

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 98 and 94 percent of their sponsored content. Even the 5th percentile brand in 2021 fails

[14] Coordinated complaints about customer service sometimes result in duplicate negative posts which are unlikely to be sponsored.

[15] We also experimented with human labelling of sponsored posts using MTurk - see Appendix H.

[16] We 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.

	Mean Across	Percentiles					Number of
Year	Brands	P5	P25	Median	P75	P95	Brands
2014	0.984	0.902	0.989	0.997	1.000	1.000	259
2015	0.982	0.923	0.988	0.996	0.999	1.000	260
2016	0.980	0.919	0.984	0.995	0.998	1.000	263
2017	0.972	0.872	0.978	0.990	0.996	1.000	264
2018	0.965	0.890	0.963	0.985	0.994	0.999	267
2019	0.962	0.870	0.959	0.980	0.992	0.998	267
2020	0.950	0.839	0.957	0.981	0.991	0.998	267
2021	0.944	0.795	0.941	0.971	0.987	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.

to disclose close to 80 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 over 10 percentage points.[17] Overall, the improvement is modest, with the average non-disclosure rate dropping by only 4 percentage points from 0.984 to 0.944 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.[18] 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.[19] 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

[17] We 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 qualitatively very similar.

[18] We 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.

[19] In Table A5 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.

	(1)	(2)	(3)
Dependent Variable	Non-Disclosure Share	Non-Disclosure Share	Non-Disclosure Share
Brand Founded After 2000	0.018***	0.020***	0.015***
	(0.004)	(0.005)	(0.005)
# Brand Account Followers	0.007***	0.008***	0.009***
(Log-Transformed)	(0.002)	(0.002)	(0.002)
Complier Dummy			-0.432***
× After-2019 Dummy			(0.095)
Year FE	Yes	Yes	Yes
Industry FE	No	Yes	Yes
Joint F-stat Industry FE		127.8	159.4
Observations	2,115	2,115	2,115
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.

are two main exceptions: "department stores," and "health products and services," which have significantly higher disclosure rates than the other categories.[20] 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, the FTC mandated that the disclosure hashtag needed to appear at the beginning of a post. As we show in Appendix J, most sponsored posts tended to include the disclosure hashtags towards the end of the post. We find that only 9 brands (3.4% 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 to disclosure differences based on 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 I we provide additional details on how we identify compliers, as well as their characteristics and behavior over time.

[20] We note that the second category includes only Walgreens.

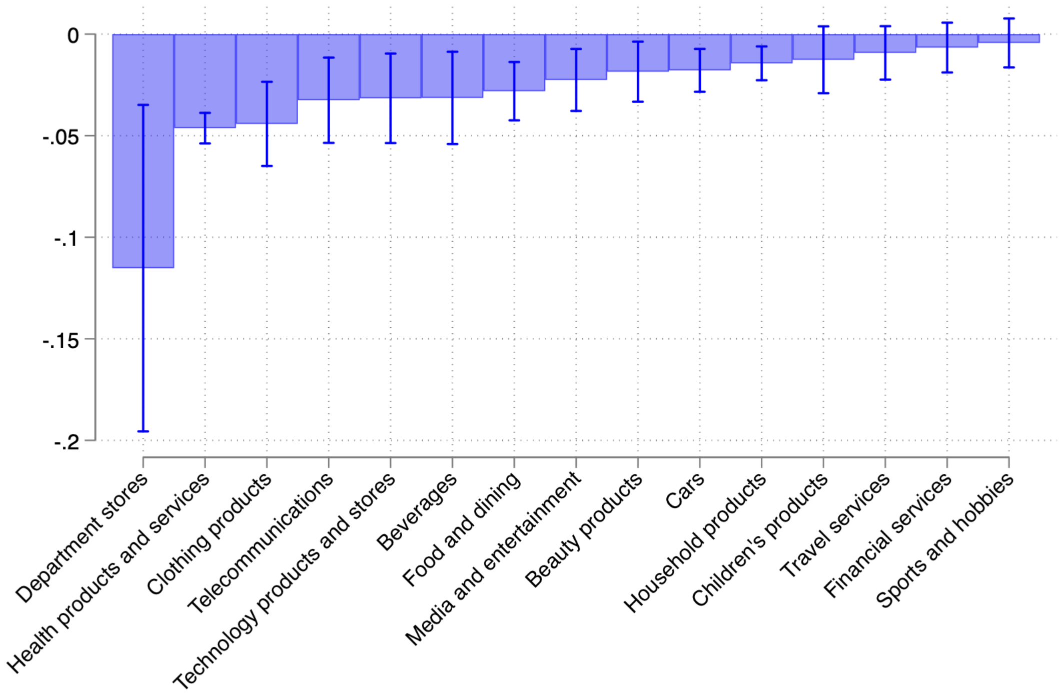


Figure 2: **Industry Differences in the Share of Undisclosed Posts.** The graph is plots the estimated values of the industry fixed effects from the regression results reported in column (3) of Table 4. The omitted category is "Home design and decoration", the category with the largest fixed effect.

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 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 J.

		Sponsored	
	Organic	Undisclosed	Disclosed
Followers	14,506	23,372	59,723
Likes per Follower	0.056	0.042	0.011
Replies per Follower	0.008	0.004	0.001
Retweets per Follower	0.020	0.015	0.005

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

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 to disclose sponsored content. These kinds of brands will likely rely more heavily on influencers and therefore disclosure rates might remain low in the future.

References

Bairathi, Mimansa and Anja Lambrecht (2024). "Influencer Marketing: Sponsorship Disclosure and Authenticity". Working Paper.

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.

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.

Geyser, 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*.

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.

Valsesia, 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.[21] 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 Sponsorship Disclosure Indicators

Our methodology for classifying posts with a sponsorship disclosure is based on official FTC guidance regarding how influencer are supposed to disclose their financial relationships with brands (FTC, 2019). A first set of indicators of disclosure are individual hashtags such as #ad, #sponsored, #paidpartnership and #brandedcontent. We include all hashtags related to the terms "ad",

[21] https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts.

“sponsored”, “gifted”, “paid”, “partnership”, “promotion” and “branded”.[22] A second set of valid disclosures consists of combining a brand’s name with sponsorship-related keywords, such as #nikesponsored, and prefixes such as “sponsoredby” or “paidby” followed by the brand’s username (e.g., #sponsoredbynike). Third, we capture disclosures through phrases such as “sponsored by @nike”. [23]

We note that #ad or #sponsored have emerged as the primary indicators of disclosure. Of all sponsored post defined based on the list of disclosures above, 86% contain one of these two hashtags. Nevertheless, a non-trivial amount of posts use alternative disclosures which are consistent with FTC guidelines and which we are able to capture with our expanded list.

C Bigram Extraction

Before extracting the bigrams displayed in Table 1, 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 frequent bigrams. Finally, we aggregate these frequencies across all brands to obtain our final list of bigram occurrence frequencies.

D 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.

E Validation Checks: Sentiment & Duplicate Posts

In this section, we analyze whether our classifier is able to predict sponsorship status for posts that are likely to be sponsored. We assess the prevalence of undisclosed sponsorship along two

[22] Full list of hashtags: #ad, #adv, #advert, #advertising, #advertisement #adpartner, #branded, #brandedcontent, #brandedpost, #brandedpromotions, #brandedtweet #gifted, #paid, #paidad, #paidads, #paidpartner, #paidpartnership, #paidpost, #paidpromotion, #paidtweet #promotionalpartnership, #spon, #spons, #sponsored, #sponsoredby, #sponsoredcontent, #sponsoredpartners, #sponsoredpost, #sponsoredseries, #sponsoredtweet #sponsorshipdeal, #sp, #undersponsorship.

[23] Specifically, we find all posts that contain "paid by", "sponsored by", "partnering with", "advertised by", "endorsed by", "paid partnership with", "gifted by", "affiliate with", "courtesy of", "brand rep for", and "promoted by" followed by ’@’ [brand Twitter username].

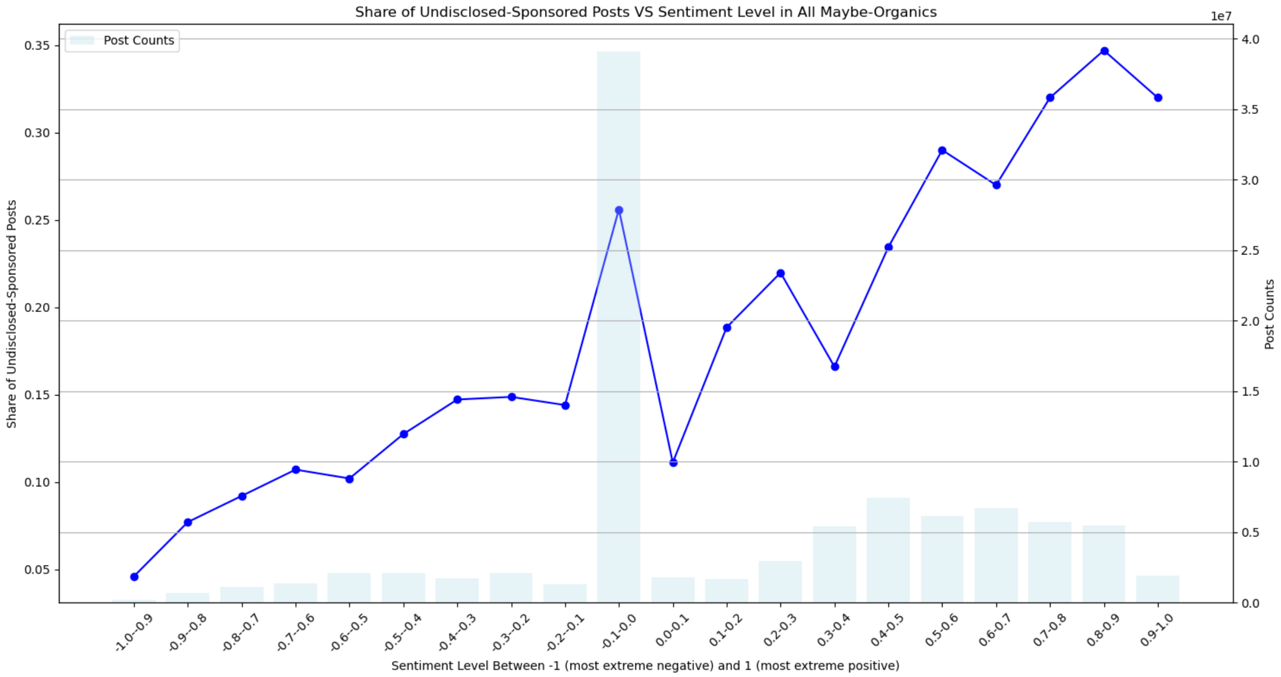


Figure A1: **Share of Undisclosed Sponsored Posts as a Function of Sentiment.** The graph plots the share of undisclosed sponsored posts for sets of posts with different sentiment levels (units are displayed on the left-hand axis). The bar chart displays the count of posts for each sentiment bracket, with the number of posts (in millions) displayed on the right-hand axis.

dimensions that were not used as an input to the classifier and can therefore serve as validity checks. First, we analyze whether the sponsorship status varies with the sentiment of a post. Second, we analyze duplicate posts, i.e. identical tweets that were posted multiple times. We would expect positive sentiment to be more prevalent in sponsored posts. Moreover, duplicate posts often represent influencer posts where brands communicate an exact script to multiple influencers and hence should be characterized by a higher probability of sponsorship status.

In Figure A1 we plot the share of sponsored posts for different values of the sentiment score, which varies between -1 and 1 and is calculated using "Vader" (Valence Aware Dictionary and sEntiment Reasoner), a sentiment analysis model specifically designed for analyzing social media texts. For very negative posts with a sentiment score close to -1 we find a low share of sponsorship status of below 5%, whereas for very positive posts the share goes up to around 30-35%.[24] This patterns, of negative posts being unlikely to be sponsored, and positive sentiment posts being substantially more likely to be sponsored, is consistent with the institutional background of brands providing clear guidance to influencers on what they can and cannot say in promotional posts. Industry guidelines encourage brands to provide influencers with a list of terms that they *do not* wish the influencer to use, which usually includes any negative language (`https://influencermarketinghub.com/influencer-contract-template/`).

[24] For simplicity we use the unadjusted share of sponsored posts.

One noteworthy data point is posts with a neutral sentiment score of zero, which are very common in the data (see the overlaid histogram in Figure A1). Such posts do not contain any words that the Vader algorithm considers to be of positive or negative valance. Interestingly, we observe a slight spike in the share of sponsored posts around zero compared to the otherwise mostly monotonic relationship between sentiment and the share of undisclosed posts. This pattern likely occurs because some sponsored posts contain short calls to action (e.g. "buy this product by clicking on the link") which do not contain words that are considered positive or negative.

We repeat this exercise for a subsample of duplicate posts. We find that duplicate posts are generally more likely to be sponsored than non-duplicate posts, which is consistent with the idea that they are more likely to represent coordinated promotional campaigns run by brands.[25] For duplicate posts with a high sentiment score (above 0.7), we find a large share of sponsored posts of around 45% compared to 30-35% for the full sample of all posts displayed in Figure A1.[26]

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 50,000 randomly selected posts without disclosure hashtags, 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.[27]

In more detail, we proceed as follows: Using the ChatGPT 3.5 API (GPT-3.5-turbo-instruct), we query GPT with the prompt "Decide whether the following Tweet is sponsored with more than 50% probability: {tweet}. Always say 'Yes' if the Tweet is sponsored with more than 50% probability, and 'No' if the Tweet is not sponsored with more than 50% probability." {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:..."). However, we replace any Twitter IDs (e.g., @PUMA) with a generic @twitter_id

[25] We note that duplicates with positive sentiment can arise for reasons other than hidden sponsorship such as posts about viral trends, community support initiatives, reactions to influencer activities, celebrations of cultural or sporting events, educational or awareness campaigns, and sharing of positive news.

[26] Duplicates with positive sentiment are more likely to constitute sponsored content, whereas duplicate posts with negative sentiment sometimes occur due to coordinated complaints about customer service.

[27] An 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.

Sample Construction:				
	# Posts	Training	Hold-out	
Sponsored	773,393			
Non-disclosed	100,737,164			
True organic (based on ChatGPT)	34,747	12,281	2,972	
True sponsored (based on ChatGPT)	15,253	12,281	2,972	
Maybe organic	100,737,164			
Total Brand-related Posts	101,510,557			
Classifier Performance:				
	Accuracy	Precision	Recall	AUC-ROC
Random Forest Classifier	71.37%	71.85%	70.26%	78.37%
	Prob. Classified as ...			
	Organic	Sponsored		
True Organic	0.725	0.275		
True Sponsored	0.297	0.703		
Classification Results:				
	Organic	Sponsored		
Maybe Organic Posts	0.600	0.400		
Adjusted Prediction	0.709	0.291		
Including True Organic Posts	0.709	0.291		

Table A1: **Classification Results for ChatGPT-based Classifier.**

when feeding tweets to the GPT. We do this for two reasons: (i) we remove brand names from our baseline classification, and would like to maintain consistency between the different approaches, and (ii) unlike our baseline classifier, GPT's labelling of organic or sponsored is based on broader language. As such, GPT is more likely to classify any discussion of specific brands as sponsored.[28]

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 misclassification 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 A1 which can be directly compared to our main classification results in Table 2.

We find that 29.1% of all posts without disclosure are classified as sponsored by the GPT-based classifier compared to 18.8% for our main classifier. The larger number of undisclosed sponsored

[28] We query GPT about the reasons for its sponsored labelling for a small sample of posts, in addition to reasons like level of detail and enthusiasm about specific products and the use of promotional hashtags, it mentions specific brand names. This suggests that in GPT's model, a brand name increases the probability that a given post is sponsored. When we do not remove brand names, GPT predicts an additional 10% of posts to be sponsored.

Label Based on...		
Our Classifier	GPT-Based Classifier	Share of Total Posts
Organic	Organic	0.526
Organic	Sponsored	0.234
Sponsored	Organic	0.074
Sponsored	Sponsored	0.166

Table A2: **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.

posts leads to a share of undisclosed posts among all sponsored posts of 97.4% 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 A2. Our preferred classification approach and the GPT-based classifier agree on approximately 70% of labels (0.53 organic/organic + 0.17 sponsored/sponsored). The main disagreements come from the GPT-based classifier labeling substantially more undisclosed posts as sponsored. 23% 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 89% of disclosed sponsored posts correctly, whereas the GPT-based classifier only labels 83% of posts correctly. Due to its superior performance, we regard our classifier as preferable to the classification based on ChatGPT.

G Lower Bound

In this section, we derive a lower bound for the share of undisclosed sponsored posts. One fundamental problem when detecting hidden sponsored content is the fact that we do not have access to a sample of organic posts. This happens because posts without a sponsorship disclosure could be either organic posts or hidden sponsored posts. Our main classifier addresses this issue by selecting a subset of posts that are highly likely to be organic based on the most common organic and sponsored bigrams. A potential concern is that our approach, while eliminating undisclosed sponsored posts from the organic training sample, also selects a non-representative set of organic posts - i.e., posts with extreme organic language. Since the classifier is more likely to label posts that do not use similar language to our labeled organic posts as sponsored, this could lead us to overestimate the amount of hidden sponsored content.

A way to test the robustness of our main approach is by using a random sample of posts without disclosure as the organic training sample. This approach avoids selecting a non-representative sample of organic posts due to the random selection. However, the presence of undisclosed sponsored posts introduces bias: Because this biased organic training sample looks more similar to sponsored posts than a representative sample of organic posts, a classification based on such a sample will lead to an underestimation of the number of hidden sponsored posts. Due to the direction of the bias, this biased classification approach can provide a lower bound to the share of undisclosed sponsored posts.

We implement a classifier based on a biased organic training sample as described above, but keep all other elements of the classifier constant (such as the classification algorithm and the sponsored training sample). In terms of prediction errors, we find that this new classifier performs worse than our preferred classifier. This is to be expected, because the biased training sample makes it harder for the classifier to distinguish organic and sponsored posts. In terms of the classification results, we find that a smaller share of 3.4% of posts without disclosure are classified as sponsored which translates into a share of undisclosed posts among all sponsored posts of 81.7%. Therefore, even a conservative lower bound estimate suggests there is a very large amount of undisclosed sponsored content.

H MTurk Sponsorship Detection

Another possible approach for generating a "ground truth" of sponsored and organic posts for classification is to use human labeling - i.e., provide a sample of posts to a group of labelers and use these labels to train the classification algorithm. However, it is challenging in practice to appropriately elicit correct classifications.

In test runs on Amazon's MTurk, a top crowd-working website, we asked workers to classify a random sample of 10 disclosed-sponsored posts. We removed the #ad or #sponsored from 4 of the 10 posts, but left the hashtags for the remaining 6 posts. Since different workers may have heterogeneous prior beliefs about the propensity of sponsored content online or what such content may look like, we asked 10 workers to label each post. We provided 0.1$ for each label from each worker, a payment that MTurk suggests for "medium complexity tasks."[29] For each post, we asked "Is this Tweet a sponsored post / advertisement? Did the user receive a reward / payment for posting it?" With the classification options of "Sponsored advertising" and "Non-sponsored." We show the 10 tweets and a summary of the results in Table A3.

We found that more than 5 MTurkers correctly classified each one of the disclosed sponsored posts with hashtags. However, only two of these posts were unanimously classified correctly. The

[29] In trial runs with lower payments, classification results were worse.

Tweet	Hashtags Still Included?	Correctly Classified
#ad We're spicing up our morning routine with three new kinds of cereal from @GenMillsCereal that we purchased from @walmart. Head on over to the blog today to see how we do mornings in our home, and which new cereal is our favorite! #NewYearNewCereals	Yes	6/10
Wow! There are some great deals on a ton of unlocked phones @BestBuy. All the best brands too. You gotta check it out before it's all over Saturday! Save up to $150 on select models. #ad	Yes	10/10
Get Milk-Bone®, Milo's Kitchen®, Pup-Peroni® and Canine Carry Outs® treats @Target. #HowloweenHouseParty #Sponsored https://t.co/VQVdChOszM via @ripplestreetfun	Yes	10/10
Hiring? Write A #Job #Ad That Excites: https://t.co/PxxM6xWocW @hubspot #Business #Ecommerce #Marketing #PR #SmallBiz #SMM #Startups #Tech	Yes	8/10
Win a copy of WWE 2K18 for @Xbox One: https://t.co/S0vZHy8ceN #ad	Yes	7/10
All the hearts for this fuzzy robe from @Walmart - It's perfect for an At-Home Valentine's Day! https://t.co/kfKWyBm1cg \\ #sponsored #walmart #walmartfashion	Yes	9/10
Want to change your life? It only takes 10. #Just10 #WMT @Walmart	No	6/10
Get that healthy glow with this incredible radiance serum @ultabeauty for 25% off! Get the deal: https://t.co/jOsP4zqzpM #SarahScoopSaves #shopping #ultabeauty #deals	No	6/10
I think you can't go wrong with any #recipe featuring @Snickers #WhenImHungry	No	1/10
You need to see this @Walmart farmer's advice to make 30 second grape jelly! http://t.co/8etIMRJa4E #TeamWalmartProduce	No	5/10

Table A3: **MTurk Classification Trial Run Results.**

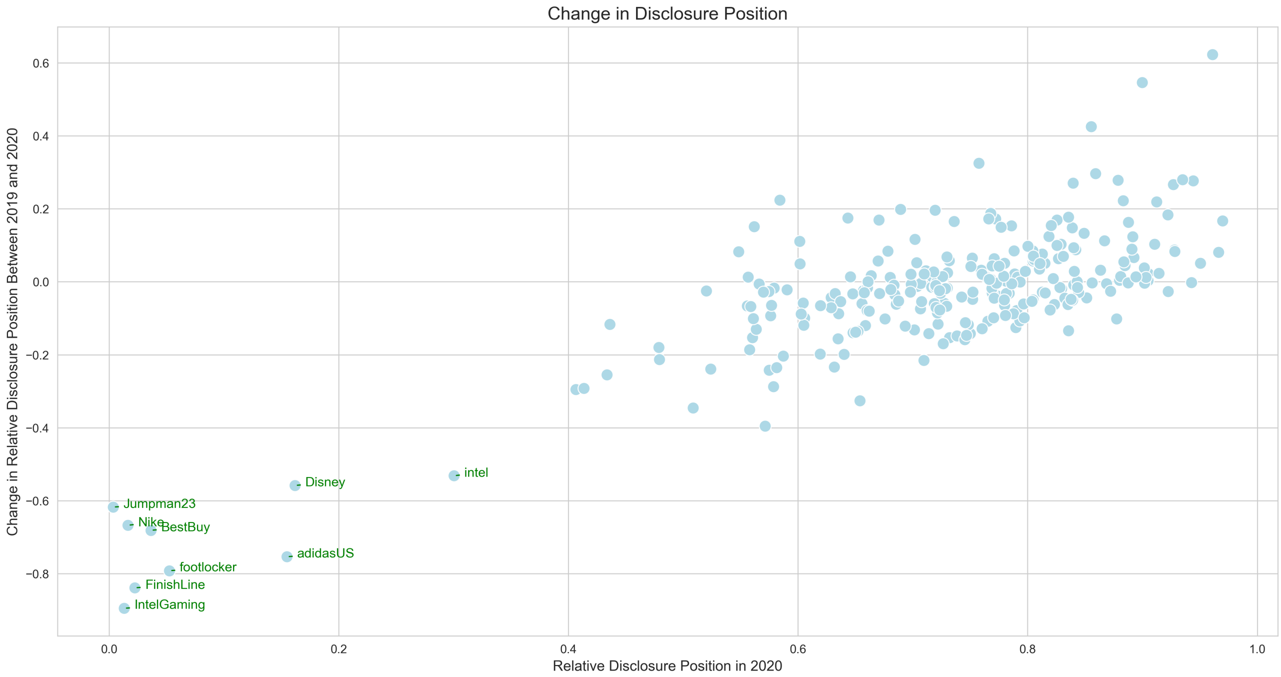


Figure A2: **Change in Disclosure Position After the 2019 Regulatory Change.**

other disclosed-sponsored posts with hashtags were correctly classified by 6/10, 7/10, 8/10, and 9/10 MTurkers. The 6/10 post is particularly concerning. The post, in its entirety is: "#ad We're spicing up our morning routine with three new kinds of cereal from @GenMillsCereal that we purchased from @walmart. Head on over to the blog today to see how we do mornings in our home, and which new cereal is our favorite! #NewYearNewCereals." Notably, the #ad appears first in the post.

Of the 4 disclosed-sponsored posts where the hashtags were removed, 3 were correctly classified by at least 5/10 MTurkers. However, only slim majorities correctly classified the posts in each case (5/10, 6/10 and 6/10). The fourth post was only classified correctly by 1/10 MTurkers.

To sum up, while the MTurk classification generally appears to be correct on average, it contains substantial noise.[30] It may be possible to elicit better responses by expanding the number of workers classifying each post, by increasing the compensation for each worker, or by doing both. However, doing this would substantially inflate the costs of classification, making it infeasible for researchers. At the rates above (10 workers paid 0.1$ per post), classification costs 1$ per post. Collecting a sample of 50,000 posts (the sample size we use for the ChatGPT-based classifier in Appendix F) would therefore be prohibitively expensive.

I 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. We plot the change in disclosure position between 2019 and 2020 against the disclosure position level in 2020 in Figure A2. We would expect complier brands to have a disclosure position close to zero in 2020 and a large change in disclosure position relative to the previous year. The scatterplot in Figure A2 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 9 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 A5 showing that compliers change their behavior in 2020 and continue to disclose a large share of sponsored posts in 2021.

J Post Characteristics: Additional Results

In the Table A4 we analyze differences 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 post, 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 0.64. The disclosure hashtag appears in the first half of the post in only 27% of all posts.

[30] Ershov and Mitchell, 2020 also used MTurk to label a small sample of Instagram captions as sponsored or organic and had similar findings regarding the noise in the classification.

		Sponsored	
	Organic	Undisclosed	Disclosed
# Hashtags (excl. Disclosure Hashtag))	1.05	1.40	1.54
Text Length	113.90	112.64	129.91
Brand Mention Position	0.49	0.46	0.44
Disclosure Position	n/a	n/a	0.64
Disclosure in the First Half Dummy	n/a	n/a	0.27

Table A4: **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.

K Additional Tables

	(1)	(2)
	Non-Disclosure	Non-Disclosure
Dependent Variable	Share	Share
Brand Founding Decade = 1950s	0.005	
	(0.012)	
Brand Founding Decade = 1960s	-0.039**	
	(0.018)	
Brand Founding Decade = 1970s	-0.007	
	(0.015)	
Brand Founding Decade = 1980s	0.002	
	(0.008)	
Brand Founding Decade = 1990s	0.000	
	(0.008)	
Brand Founding Decade = 2000s	0.015**	
	(0.006)	
Complier × (Year = 2015)		-0.014
		(0.025)
Complier × (Year = 2016)		-0.037
		(0.046)
Complier × (Year = 2017)		-0.076
		(0.072)
Complier × (Year = 2018)		-0.086
		(0.084)
Complier × (Year = 2019)		-0.069
		(0.052)
Complier × (Year = 2020)		-0.508***
		(0.085)
Complier × (Year = 2021)		-0.367***
		(0.125)
Industry FE	Yes	Yes
Year FE	Yes	Yes
Other Controls	Yes	Yes
Observations	2,115	2,115
Brands	268	268

Table A5: **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.