How Much Influencer Marketing is Undisclosed? Evidence from Twitter*

Daniel Ershov
University College London
& CEPR

Yanting He
Imperial College London

Stephan Seiler
Imperial College London
& CEPR

This draft: February 20, 2025

We study the disclosure of influencer posts on Twitter across a large set of brands based on a unique data set of over 100 million posts and a novel classification method to detect undisclosed sponsorship. Using our preferred empirical specification we find that 96% of sponsored posts are not disclosed. This result is robust to a series of specification tests and even a lower bound classification still yield an undisclosed share of 82%. 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 young brands with a large Twitter following. Using an online survey we find that many consumers are not able to identify sponsored content without disclosure. Our findings highlight a potential need for further regulatory scrutiny and suggest that researchers studying influencers must account for non-disclosed sponsored content.

Keywords: Social Media, Influencer Marketing, Advertising Disclosure, Consumer Protection

*We thank Brett Hollenbeck and Ilya Morozov, seminar participants at Bocconi, Cardiff University, ESADE, Goethe University Frankfurt, Helsinki GSE, Imperial College London, University of Mannheim and Yelp, and conference participants at EMAC, Marketing Science, the CESifo digitization area meeting, the UK Digital Economics Workshop and the YES|Marketing conference 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. Because consumers might find it difficult to distinguish paid influencer posts from genuinely organic content, regulators in many countries now require any paid content to be disclosed. However, due to the novelty of this marketing channel, the evolving regulatory framework, and the potentially negative effects of disclosure on content engagement, the extent to which influencers adhere to this regulation is not clear.[1] In this paper, we use a novel data set of over 100 million brand-related posts from 268 brands on Twitter and a new text-based classification approach to identify undisclosed sponsored content. We quantify the overall importance of undisclosed influencer posts, evaluate whether consumers can detect undisclosed commercial content, track the evolution of undisclosed posts over time, and characterize the brands from which they originate.

Our analysis is relevant for regulators who worry that undisclosed influencer marketing might mislead consumers. For example, the Federal Trade Commission (FTC) states: "if there's a connection between an endorser and the marketer that a significant minority of consumers wouldn't expect [...], that connection should be disclosed clearly and conspicuously". We find that a large share of 96% of sponsored content is undisclosed and that a substantial share of participants in an online survey are not able to identify commercial content when it lacks disclosure. We also show that although regulation tightened over our sample period from 2014 to 2021, the share of undisclosed content decreased only slightly. Finally, we find that undisclosed sponsored posts tend to originate from younger brands with a large Twitter following indicating that future compliance rates may remain low. In totality, our findings suggest that current efforts to monitor disclosure and enforce compliance have been insufficient, pointing to the need for further regulatory scrutiny.

Our findings are also more broadly relevant to marketers and researchers analyzing influencer marketing or sponsored user-generated content, suggesting the need to carefully deal with non-disclosure. This is obviously true when studying the effects of disclosure on engagement (Karagür et al. (2022), Bairathi and Lambrecht (2024)), consumer choice and welfare. Additionally, this is also true when studying the effectiveness of this channel, the interaction between influencers and brands, or the impact of platform and public policies on sponsored content (e.g. Ershov (2024)). A number of papers studying these topics distinguish between sponsored and organic posts solely based on disclosure hashtags (e.g., Cheng and Zhang (2024); Tian et al. (2024)). Our results suggest this approach both misses a large share of sponsored content and may incorrectly label actual sponsored content as organic, potentially biasing results. Our analysis is also closely related

[1] Evidence on the effects of disclosure on engagement is mixed. Several papers show that disclosure of commercial relationships between influencers and brands can reduce influencer trustworthiness (Karagür et al. (2022)), consumer sentiment (Liu et al. (2024)), and can lead to loss of followers (Cheng and Zhang (2024)). However, Cao and Belo (2023) and Zhang et al. (2023) show evidence suggesting that disclosure of sponsored content increases engagement for some influencers.

to the literature on advertising disclosure in traditional advertising markets (Sahni and Nair (2020)) and the literature on fake reviews (Mayzlin et al. (2014), Luca and Zervas (2016), and He et al. (2022)), which similarly to our study aims to identify hidden commercial content online.

On the methodological side, we provide a new approach for detecting hidden sponsored posts. By definition, the presence of hidden content implies that we do not have access to a clean training sample of truly organic and undisclosed sponsored posts. The nature of the classification problem therefore requires us to build a training sample from available data that is a valid under a set of assumptions. Specifically, we assume that disclosed and undisclosed sponsored posts employ the same language and that organic posts constitute the majority class among posts without disclosure. The latter assumption allows us to isolate a training sample of organic posts based on the most frequently occurring terms in posts without disclosure. Based on these assumptions, we train a text-based classifier on the content of sponsored and organic posts in our training sample. We probe the underlying assumptions through a series of validation checks, by classifying posts using an alternative approach based on ChatGPT, and by employing an intentionally biased classification approach that provides a lower bound for the share of undisclosed sponsored posts. ChatGPT predicts a similar share of undisclosed posts as our baseline classifier, and the lower-bound approach predicts a share of 82%. Taken together, we see these results as evidence that a large share of sponsored content in our data is undisclosed.

A second methodological innovation of our approach is an adjustment procedure for misclassification error. We show that the classification error of the algorithm, which is known in the training sample, can be used to adjust the classification results in a new sample of posts. This adjustment procedure can be applied to any classification problem and therefore has value beyond our specific setting.

Both methodological innovations set our paper apart from other papers in the literature, which either did not address the problem of a contaminated training sample (Kim et al. (2021), Ershov (2024), Bertaglia et al. (2024)), or relied on manual labeling to generate a training sample (e.g., Silva et al. (2020), Bairathi and Lambrecht (2024)). Manual labeling is less scalable and influenced by the decisions of a small group of individuals, which, as suggested by our online survey results, can be unreliable. Instead, our approach rests on a transparent set of assumptions, making it possible to evaluate any potential biases. Moreover, none of the existing papers in the literature dealing with hidden sponsorship adjust for classification error. Finally, we note that, while some elements of our approach are based on specific institutional details, many aspects are likely relevant in other related settings. The adjustment procedure is applicable to any classification problem and the idea of using the most frequently occurring terms to isolate posts from the majority class applies to any settings where hidden content makes up a small share of all content.

Prior evidence on the amount of undisclosed sponsored content on social media is limited and comes from a small set of academic studies and sweeps conducted by European regulators. These generally show that non-disclosure of advertising is common, but there is variation in the exact non-disclosure rates they find. Mathur et al. (2018) analyze posts with affiliate links on YouTube and Pinterest from March 2018, and find non-disclosure rates of 93.0% and 95.4%, respectively.[2] Bertaglia et al. (2024) find a non-disclosure rate of 88% for a small and selected sample of 100 US Instagram influencers. Using data from a sample of German Instagram influencers under stringent regulations, Ershov (2024) find non-disclosure rates of between 40% and 60%. In an audit of 576 influencers, the EU Commission found that only 20% of influencers systematically disclosed sponsorship status (Commission (2024)). Other small scale sweeps conducted by authorities in France, Spain, the UK, and Ireland find non-disclosure rates between 50% and 80%.[3] Our study is the first large-scale investigation of sponsored content that spans a large set of brands and a long time period of 8 years. Our results confirm the high non-disclosure rates found in previous studies with more limited scope.

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 (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,b)). 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

[2] Affiliate links are an observable subset of all commercial content.

[3] See https://commission.europa.eu/live-work-travel-eu/consumer-rights-and-complaints/influencer-legal-hub_en

and advertisers (FTC (2017a)). Updated guidelines and information packages were released in 2017 and 2019 (FTC (2017b, 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 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 Online Appendix A1. 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 773,393 posts with a sponsorship disclosure such as #ad or #sponsored[7] 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 two consecutive words) in both types of posts in Table 1.[8] We find that the set of top bigrams for

[4]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).

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

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

[7]In Online Appendix A.2 we provide details on how we define a valid sponsorship disclosure based on existing FTC rules.

[8]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

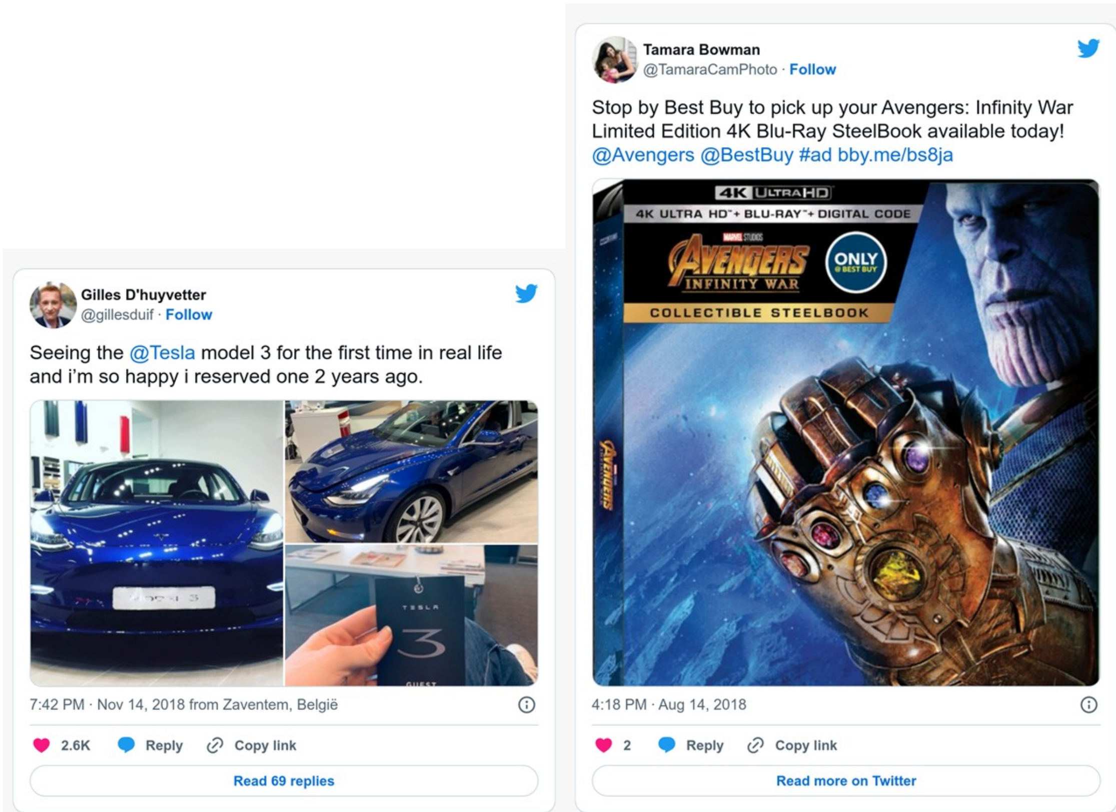

Figure 1: **Example: Post without Disclosure (left) & Disclosed Sponsored Post (right).**

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

Given the weak enforcement of disclosure regulations throughout our sample period, and the lack of guidelines prior to 2016, there are likely undisclosed sponsored posts in our sample. Identifying these posts is our main empirical challenge. An ideal data-set for this purpose would include a large random sample of posts without disclosure that we know were paid for (i.e., hidden sponsored posts), and a large random sample of posts without disclosure that were not paid for (i.e., genuine organic posts). Then, we could train a ML classifier to distinguish the language of sponsored and organic

in posts for fewer than 20 brands. In Online Appendix A.3 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	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.**

posts without disclosure, and apply this classifier to the rest of our sample. However, because we are trying to detect *hidden* sponsored content and due to the legal requirement to disclose, no such data-set exists.[9] Therefore, we instead train a classifier on a training sample constructed from available data. We discuss our construction of the training sample and the assumptions under which this sample accurately captures the underlying language of the two types of posts below.

To illustrate our approach, it is helpful to consider the illustrative distribution of organic and undisclosed-sponsored post language in panel (a) of Figure 2. For simplicity we consider a single dimension of differentiation - the degree to which posts contain commercial language, which is likely a key differentiator between organic and hidden sponsored posts. The two types of posts are represented by the black and gray distribution respectively. A classifier trained on a representative sample of posts, will try to detect differences between the two types of posts and classify them based on a cut-off indicated by the vertical line. We need to deal with two issues: (i) if our training sample is non-representative, the training samples of organic or sponsored posts will not coincide with the true distributions in Figure 2, (ii) some posts will be mis-classified due to the overlap in distributions. We discuss both issues below.

[9] The only possibility of obtaining such a data-set would be from information gathered by regulators that reveals ex-post whether some posts without disclosure were sponsored. Due to the current lack of enforcement described in Section 2, no such data exists.

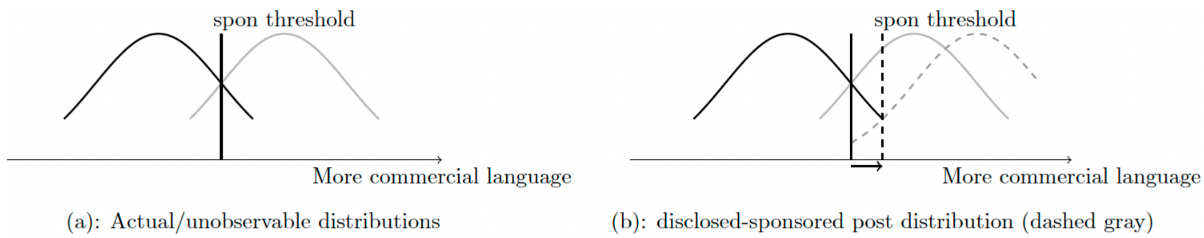

Figure 2: **Illustrative Distribution of Organic and Sponsored Posts.** Organic and hidden sponsored posts are represented by the black and gray distribution respectively. Panel (a) depicts the true distributions of the two types of posts. Panel (b) depicts the case where the training sample of sponsored posts (dashed gray) is shifted right-wards towards more commercial language.

4.1 Training Sample: Sponsored Posts

We use disclosed-sponsored posts to stand-in for undisclosed-sponsored posts in the training sample. This provides us with an unbiased sample of hidden sponsored language if the language of disclosed and undisclosed sponsored posts is the same. We believe this is a reasonable assumption - as described in Section 2, brands want commercial language in sponsored content to drive purchase behavior. Moreover, since the probability of detection and punishment for non-compliance with regulations is low influencers and brands should have little incentive to alter the language of posts based on their disclosure.

Nonetheless, given that disclosure is a choice variable, there may still be a selection effect leading to biased language in our training sample's sponsored posts. We believe the most likely scenario in our setting is that disclosed sponsored posts contain more commercial language than undisclosed sponsored posts, which is consistent with theory work on influencer marketing. If consumers on average find organic posts more persuasive/ higher quality than sponsored posts (as in Mitchell (2021), Ershov (2024), and Pei and Mayzlin (2022)'s analytic models), and if consumers form beliefs about posts' sponsorship status based on their language (as in Ershov (2024)), then there are stronger incentives to not disclose sponsored posts with more organic language.[10] If our sponsored training sample contains posts with more commercial language than hidden sponsored posts, the training sample of sponsored posts will be shifted rightwards as illustrated in panel (b) of Figure 2. Such a shift in the training sample will shift the classification cut-off rightwards which will lead to fewer posts being classified as sponsored.

4.2 Training Sample: Organic Posts

To identify a subset of organic training 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,

[10] We note that this assumption is based on institutional features and an implicit model of influencer behavior from prior literature, but it is not directly testable in our data. There are plausible drivers that lead to a bias in the other direction (influencers might be less likely to add a disclosure when the content is obviously commercial) or lead to a bias that is difficult to sign (larger brands might invest more in content production and disclose more often).

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.[11] We compile a list of organic training 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. Although this approach likely filters out hidden sponsored posts, it possibly selects a non-representative subset of organic posts whose language is more "organic" than the average organic post.[12] Such a non-representative set of training posts will lead to an overstatement of the number of undisclosed sponsored posts because it will shift the distribution of organic posts in the training sample (the black distribution in Figure 2) left and therefore the classification cut-off leftwards.

While it is not feasible to directly test the assumptions underlying the formation of the training samples, we provide a robustness check based on an alternative classifier and a classification approach where both sponsored and organic training samples are shifted rightwards in Section 4.5. The latter approach provides a lower bound to the number of sponsored posts due to the direction of the overall bias.

4.3 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 which arises if the distributions of organic and hidden sponsored posts overlap (as in Figure 2). To fix ideas, consider the case where organic posts have a 5% chance of being 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

[11] Our classifier results label 19.6% of posts without disclosure as sponsored, confirming this assumption ex-post.

[12] By construction, we are excluding all posts containing the most frequent sponsored bigrams, i.e. with very commercial language, from our organic training sample.

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.

4.4 Classification Approach and Results

Our classification results are based on a random forest classification algorithm.[13] To train the algorithm, we take all 773,393 disclosed-sponsored posts and randomly select a sample of equal size from the organic posts which we identified based on the procedure laid out in Section 4.2.[14] 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 organic training posts in the hold-out sample, 91.8% are correctly classified as organic. We achieve a similar success rate with regards to sponsored training 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 4.3 that corrects for the known mis-classification error of our model[15] and add the sample of organic training posts

[13] 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). We also implement an approach where we first transform words into embeddings using the pre-trained GloVe algorithm (Pennington et al. (2014)) and then apply a random forest classifier to the average embedding values of each post. This classifier yields a share of sponsored posts among the "maybe organic" posts of 20.3% which is very close to the rate of 19.6% obtained from our main classifier.

[14] We apply several pre-processing steps to all posts which we describe in Online Appendix A.4. Importantly, 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 organic posts.

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

Sample Construction:				
	# Posts	Training	Hold-out	
Disclosed Sponsored	773,393	618,079	155,314	
Non-disclosed	100,737,164			
Organic Training Sample	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 State: Organic	0.918	0.082		
True State: Sponsored	0.108	0.892		
Classification Results:				
	Organic	Sponsored		
Maybe Organic Posts	0.760	0.240		
Adjusted Prediction	0.804	0.196		
Including Organic Training Sample	0.812	0.188		

Table 2: **Classification of Undisclosed Sponsored 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.[16] Therefore, 96.1% of influencer commercial activity is undisclosed.

4.5 Robustness: Alternative Training Samples

To probe the robustness of our classification results we report results from two alternative classifications based on different training samples. First, we implement a classifier where we change the organic training sample by selecting a random set of posts from all posts without disclosure. As such, we are mixing organic and sponsored posts in the organic training data. According to our framework in Figure 2, this shifts the organic training sample to the right. This approach will lead to a clear downward bias in the number of undisclosed sponsored posts because both organic and

[16] An alternative application of the classifier could use a continuous measure of sponsorship, based on the posterior probability of being sponsored, instead of the binary sponsored/non-sponsored classification. However, then it wold be impossible to apply the classification error correction procedure, which depends on the existence of the binary classes.

sponsored training sample are shifted to the right, i.e. they both have language that is "too sponsored" as compared to the true underlying distribution. This classifier therefore provides a lower bound to the share of undisclosed sponsored posts. We find that even this conservative approach yields an undisclosed share of 81.7%.[17]

Second, we implement a classifier based on a training sample that is formed by letting ChatGPT classify a set of posts without disclosure as organic or sponsored. Importantly, this approach does not rely on the assumption used to form the training sample for our main classifier. We find that the ChatGPT-based approach has a high agreement rate with our classifier and the predicted undisclosed sponsorship rate is equal to 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.

Taken together we view the lower bound of 81.7% undisclosed share and the very similar classification results of our main classifier and ChatGPT as evidence that a large share of sponsored posts are undisclosed. We provide additional details on the two alternative classifiers in Online Appendix B.

4.6 Other Validation Checks

We further test the validity of our classifier in several ways: First, we look at the correlation between our classification and post sentiment, which is not used as an input into classification. Intuitively, 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.[18] We also analyze how our classifier performs on a publicly available dataset containing tweets about Brexit. As expected, we classify the vast majority of such posts as organic. Finally, we investigate which features of the language of posts are most predictive of sponsorship status and find that a set of obviously commercial terms are strong predictors of whether a post is sponsored. We report more detailed descriptions of the above validation checks in Online Appendix C.

4.7 Human Coding of Sponsored Posts

We also attempted to use human coders on MTurk to classify posts as sponsored or organic. To assess coders' ability to recognize sponsored posts, we show them a randomly selected sample

[17] This kind of biased training sample is exactly the approach that has been used by several prior papers as their main classifier (Kim et al. (2021) and Ershov (2024)).

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

of disclosed sponsored posts with disclosure hashtags removed. Interestingly, we find that the performance of survey participants is relatively poor. 34% of respondents could not correctly identify posts as sponsored. Additional details are in Online Appendix D.

We take away two lessons from this exercise. First, the low rates of correct classification caution against the use of non-trained human coders as the basis for the classification of hidden sponsored posts.[19] Second, we interpret the survey results as evidence that social media users are not able to easily detect hidden sponsored content.[20] The degree to which consumers are able to detect sponsored content is important for understanding the possible welfare consequences of undisclosed posts and our findings suggest that the large amount of hidden sponsored posts might mislead consumer. We note that to fully capture welfare implications one would need to also study how disclosure affects consumers' purchase behavior, which is outside of the scope of this paper.

5 Determinants of Undisclosed Content

Our data contains a broad set of brands from a wide variety of industries, ranging from 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 which brands are more likely to disclose sponsored content.[21] We conduct 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.[22]

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 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.[23] Overall, the improvement is modest, with the average

[19] Manual classification by graduate RAs has been used in several papers (for example Silva et al. (2020) and Bairathi and Lambrecht (2024)), but would likely be prohibitively expensive for the purpose of labeling tens of thousands of posts.

[20] We required survey participants to be active Twitter users.

[21] We provide a complementary set of results regarding how the characteristics of the accounts and posts vary across types of posts in Online Appendix E.

[22] 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.

[23] 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

Year	Mean	Percentiles					Number of
		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.

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.[24] 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.[25] 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 differences in more detail, we plot the distribution of category fixed effects in Figure 3. 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.[26] 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

change in the number of brands in Table 3. When using a balanced set of brands, results are qualitatively similar.

[24] 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.

[25] In Table A5 in the online 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 brands founded after 2000 disclosing less.

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

	(1)	(2)	(3)
Dependent Variable	Non-Disclosure Share	Non-Disclosure Share	Non-Disclosure Share
Brand Founded After 2000	0.018***	0.020***	0.014***
	(0.004)	(0.005)	(0.004)
# Brand Account Followers	0.007***	0.008***	0.010***
(Log-Transformed)	(0.002)	(0.002)	(0.002)
Law-Abiding Dummy			-0.046
			(0.049)
Law-Abiding Dummy			-0.389***
× After-2019 Dummy			(0.117)
Year FE	Yes	Yes	Yes
Industry FE	No	Yes	Yes
Joint F-stat Industry FE		127.8	152.4
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.

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 Online Appendix E, 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 "law abiding". In column (3) of Table 4 we add a law-abiding dummy and an interaction of the law-abiding 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 law-abiding brands that comply with the regulatory change tend to be older brands in more traditional industries such as BestBuy, Footlocker, and Disney. In Online Appendix F we provide additional details on how we identify law abiding brands, as well as their characteristics and behavior over time.

6 Conclusion

In this paper, we quantify the extent 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

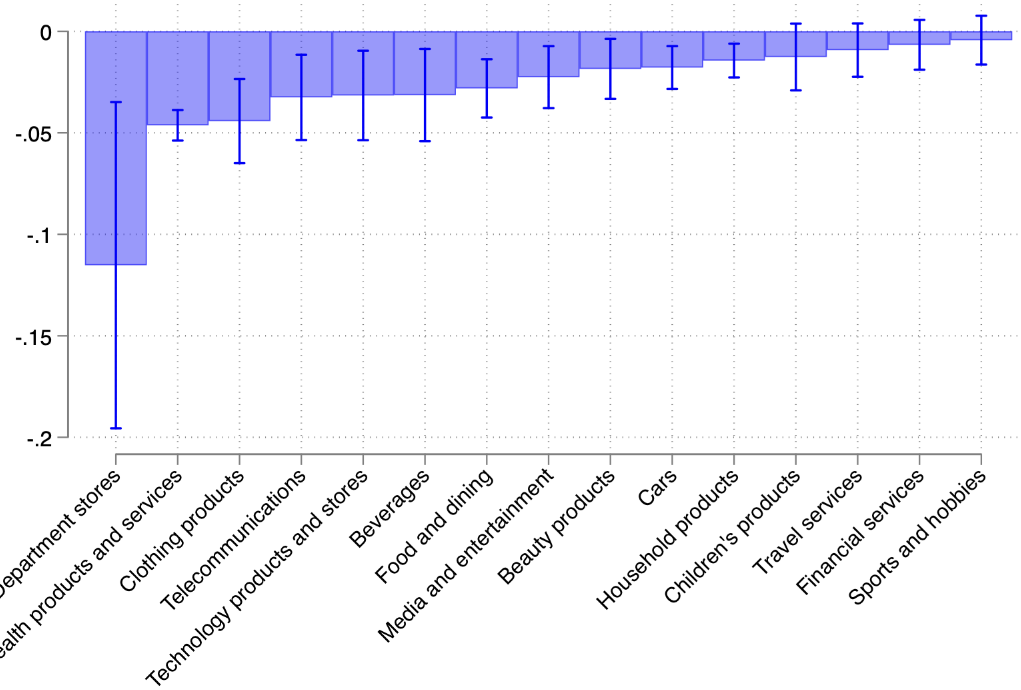

Figure 3: **Industry Differences in the Share of Undisclosed Posts.** The graph is plots the estimated values of the industry fixed effects from the regression results reported in column (3) of Table 4. The omitted category is "Home design and decoration", the category with the largest fixed effect.

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 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.

Our findings are relevant for regulators who are concerned that non-disclosure will lead to consumers making suboptimal purchase decisions because they perceive undisclosed influencer advertising as "authentic" advice. Our main findings show that undisclosed content is prevalent and our surveys results suggest there is likely a sizeable minority of consumers who cannot correctly infer commercial relationships in undisclosed-sponsored posts, consistent with FTC concerns. However, we do not have the necessary data to causally link non-disclosure to consumer purchase decisions which requires random variation in disclosure status as in Sahni and Nair (2020). As such, we view our findings and methodology as an important first step towards establishing the welfare impact of undisclosed-sponsored content, and leave further exploration of consumption effects and welfare to future research.

Our findings are also relevant to other research on influencer markets, which often requires knowledge of undisclosed-sponsored content. For example, researchers wishing to evaluate the effectiveness of influencer advertising relative to organic content would need to first separate organic from undisclosed-sponsored content. Similarly, studies that analyze the impact of regulatory changes on influencer behavior need to account for hidden sponsored posts to provide an exhaustive analysis.

Declarations

Funding and Competing Interests: All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. The authors have no funding to report.

References

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

Bertaglia, T., C. Goanta, G. Spanakis, and A. Iamnitchi (2024): "Influencer Self-Disclosure Practices on Instagram: A Multi-Country Longitudinal Study," *Working Paper*.

Cao, Z. and R. Belo (2023): "Effects of explicit sponsorship disclosure on user engagement in social media influencer marketing," *Forthcoming, MIS Quarterly*.

Cheng, M. and S. Zhang (2024): "Reputation burning: Analyzing the impact of brand sponsorship on social influencers," *Management Science*.

Commission, E. (2024): "Investigation of the Commission and consumer authorities finds that online influencers rarely disclose commercial content," https://ec.europa.eu/commission/presscorner/detail/en/ip_24_708. Accessed in 2024.

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, D. (2024): "The effects of advertising disclosure regulations on social media: Evidence from instagram," *The RAND Journal of Economics*.

FTC (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.

——— (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, W. (2018): "Influencer Contract Template," https://influencermarketinghub.com/influencer-contract-template/. Accessed in 2023.

He, S., B. Hollenbeck, and D. Proserpio (2022): "The market for fake reviews," *Marketing Science*, 41, 896–921.

Karagür, Z., J.-M. Becker, K. Klein, and A. Edeling (2022): "How, why, and when disclosure type matters for influencer marketing," *International Journal of Research in Marketing*, 39, 313–335.

Kim, S., J.-Y. Jiang, and W. 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*, 319–327.

Liu, X., C. Sun, Z. Shi, A. Ghose, F. Xiong, and X. Li (2024): "The Effect of Voice AI on Digital Commerce," *Information Systems Research (forthcoming).*

Lovett, M., R. Peres, and R. Shachar (2014): "A data set of brands and their characteristics," *Marketing Science*, 33, 609–617.

Luca, M. and G. Zervas (2016): "Fake it till you make it: Reputation, competition, and Yelp review fraud," *Management Science*, 62, 3412–3427.

Mathur, A., A. Narayanan, and M. Chetty (2018): "Endorsements on Social Media: An Empirical Study of Affiliate Marketing Disclosures on YouTube and Pinterest," *Proceedings of the ACM on Human-Computer Interaction*, 2, 1–26.

Mayzlin, D., Y. Dover, and J. Chevalier (2014): "Promotional reviews: An empirical investigation of online review manipulation," *American Economic Review*, 104, 2421–2455.

Michaelsen, F., L. Collini, C. Jacob, C. Goanta, S. E. Kettner, S. Bishop, P. Hausemer, C. Thorun, and S. Yesiloglu (2022): "The impact of influencers on advertising and consumer protection in the Single Market," *Study requested by the IMCO committee.*

Mitchell, M. (2021): "Free ad (vice): Internet influencers and disclosure regulation," *The RAND Journal of Economics*, 52, 3–21.

Pei, A. and D. Mayzlin (2022): "Influencing social media influencers through affiliation," *Marketing Science*, 41, 593–615.

Pennington, J., R. Socher, and C. D. Manning (2014): "GloVe: Global Vectors for Word Representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Sahni, N. S. and H. S. Nair (2020): "Sponsorship disclosure and consumer deception: Experimental evidence from native advertising in mobile search," *Marketing Science*, 39, 5–32.

Silva, M., L. Santos de Oliveira, A. Andreou, P. O. Vaz de Melo, O. Goga, and F. Benevenuto (2020): "Facebook ads monitor: An independent auditing system for political ads on facebook," in *Proceedings of The Web Conference 2020*, 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.

Tian, Z., R. Dew, and R. Iyengar (2024): "Mega or micro? Influencer selection using follower elasticity," *Journal of Marketing Research*, 61, 472–495.

Zhang, J., S. Lee, and Y. Tan (2023): "The Effects of the FTC Policy and Affiliation Disclosures on Product Review Video Engagement: Evidence from YouTube," *Available at SSRN 4383852.*