The RAND Journal of Economics

ARTICLE OPEN ACCESS

The Effects of Advertising Disclosure Regulations on Social Media: Evidence From Instagram

Daniel Ershov[1] | Matthew Mitchell[2]

[1]UCL School of Management | [2]Graduate Department of Management, University of Toronto, Toronto, Canada

Correspondence: Daniel Ershov (d.ershov@ucl.ac.uk)

Funding: 'Daniel Ershov would like to acknowledge the support from ANR under grant ANR-17-EUR-0010 (Investissements d'Avenir program).'

ABSTRACT

We study the effects of advertising disclosure regulations in social media markets. Using data from a large sample of Instagram influencers in Germany and Spain and a difference-in-differences approach, we empirically evaluate the effects of German strengthening of disclosure regulations on post content and follower engagement. We measure whether posts include suggested disclosure terms and use text-based approaches (keywords, machine learning) to assess whether a post is sponsored. We show substantial adoption of disclosure but also a 12% increase in sponsored content and an increase in the share of undisclosed-sponsored content consumers are exposed to. We also find reductions in engagement, suggesting that followers were likely negatively affected.

1 | Introduction

This article studies social media influencers to better understand government regulation of online information dissemination. Consumers in many online markets rely on advice or consume content from intermediaries without compensating them directly. Examples include blogs, popular social media users ("influencers"), or larger providers of search information like Google and Amazon.[1] How intermediary content or advice quality might be impacted by compensation is one of several key digital-market related policy concerns.[2] In the case of Google, search results might be steered to Google owned properties that earn them revenues.[3] In the case of a smaller influencer on social media like Instagram or TikTok, advice might be affected by payment received from a sponsored product. Sponsorship in these markets is common as influencers are compensated to post content about specific products or services. By some estimates, the influencer economy is valued in the billions of dollars/euros, with top influencers receiving as much as $1 million per sponsored post (CNBC.com).

In recent years, concerns about hidden sponsored content and misleading online advertising led to regulatory scrutiny of this market, and how to regulate influencers large and small is an important policy question. A growing number of countries including Germany, the United Kingdom, and the United States instituted disclosure regulations on social media posts (i.e., ASA.org.uk). Under a disclosure regime, influencers have to identify content with a "#ad" or an equivalent statement if they were compensated for it. In some countries, such as Germany, failure to comply has resulted in fines for influencers and advertisers (ISLA.com). Nonetheless, unlike similar regulations in other sectors such as finance, disclosure regulations online are more imperfect. The nature of the content and the regulated individuals means that legislation leaves room for interpretation by enforcement agencies (e.g., who is an "influencer"? what does "compensation" mean?) and results in imperfect compliance.

Economic theory presents two competing views on the effects of disclosure regulations. Drawing on theories of buyer–seller transactions, regulatory agencies (i.e., FTC in the United States,

or ASA in the United Kingdom) view disclosure regulations are welfare increasing. In this view, more information is better, as consumers are less likely to unknowingly engage with (and purchase) low-quality sponsored content/products. However, the intuition behind this view is primarily based on models where content supply is fixed, and recent articles such as Inderst and Ottaviani (2012), Fainmesser and Galeotti (2021), Pei and Mayzlin (2021), and Mitchell (2021) suggest it might be incomplete. These articles show that in settings where advice is not compensated directly, regulations affecting the compensation channels for advice might have adverse effects. The total amount of sponsored content produced might increase in equilibrium after disclosure regulations, and market welfare could fall.[4]

Our research question is: What effects do advertising disclosure regulations have on content and engagement in user generated social media platforms? To empirically evaluate the effects of disclosure regulations, we collect Instagram data from the 2010s for a random sample of 12,000 *local* German and Spanish Instagram influencers using CrowdTangle.com. The German regulatory environment became substantially stricter toward the end of 2016: In October 2016, German state media authorities clarified that existing requirements for advertising disclosure applied to social media and provided guidelines for compliance. This was quickly followed by legal cases and fines against non-compliant influencers in 2017. By comparison, Spain had no existing guidelines or regulations about social media advertising disclosure.

For each influencer in our sample, we observe a full history of public posts, including post text, the number of likes, the number of comments, and a partial history of the number of followers.[5] We face a measurement challenge using this data: while we easily detect *disclosed and sponsored* posts using a list of disclosure words, we do not directly observe *undisclosed-sponsored* posts. The second type of post is likely particularly popular in Spain and Germany prior to the regulatory change.

We detect sponsored posts by applying natural language processing and classification algorithms on the text of posts. The algorithms separate sponsored content from non-sponsored content, independent of disclosure, allowing us to study how disclosure regulation impacts disclosed ads and undisclosed ads. We use two main approaches: (1) a manual rule-based approach that labels a post as sponsored if it includes certain keywords associated with commercial intent (i.e., "promotion," "promo code," "context," or a brand name);[6] (2) a supervised machine learning (ML) approach that labels posts with language similar to the language of disclosed-sponsored posts as sponsored. We take a random sample of 300,000 posts from post-regulation Germany to train a Stochastic Gradient Descent (SGD) classifier.[7] We address a novel challenge of different languages used by influencers and potential changes in language over time by transforming posts from "word-space" into multi-lingual "embedding/meaning-space," a popular approach in natural language processing.[8] We also use a combination of the two approaches, labeling a post as sponsored if both the manual and an ML algorithm classifies it as such. After applying the classification methods, we have a sponsored/non-sponsored and a disclosed/undisclosed label for each post. We identify a substantial amount of sponsored content in Germany and in Spain, including *undisclosed-sponsored* content.

To isolate the causal effect of regulatory changes on sponsorship and engagement, we use a difference-in-differences methodology, comparing influencers in Germany to influencers in Spain before and after the regulatory change in Germany. We use a Coarsened Exact Matching strategy to restrict our sample to comparable influencers in the two countries. We are left with a sample of approximately 600 German and 600 Spanish influencers.[9] Our difference-in-differences regressions on this sample control for influencer and time fixed effects, as well as other country and influencer time-varying characteristics. We look at a number of influencer/month level outcomes: how much do influencers disclose (does the regulation actually work?) and how much sponsored content (either disclosed or undisclosed) they post. We also look at engagement—whether stronger disclosure regulations affect the average number of likes, comments, and followers, and whether the ratio of engagement between undisclosed-sponsored and non-sponsored posts changes.

We introduce a theoretical model to help interpret our empirical findings, highlighting the roles of organic, disclosed-sponsored, and undisclosed-sponsored posts, and taking into account our measurement exercise. In the model, an influencer chooses to post an organic or sponsored post and the degree of sponsored/commercial language used in the post. The influencer earns higher payoffs from sponsored posts with more commercial language, but their revenues also depend on follower attention, and followers prefer to pay attention to organic posts. Regulations that disclose a fraction of sponsored posts can increase follower attention to more sponsored language, as they trust language in undisclosed-sponsored posts that slip through the disclosure filter more.[10] Although influencers in the model endogenously adjust their language, we show that our measurement exercise and the comparison of sponsorship rates in regulated and unregulated markets will correctly identify changes in equilibrium influencer incentives.

Our difference-in-differences estimates show that disclosure regulations affect the type of content influencers post online. Results from the SGD classifier, the manual classifier, and the combination of classifiers show a statistically significant increase in the share of sponsored content posted by German influencers after disclosure requirements became stronger. The magnitude of changes is substantial. Relative to a baseline pre-treatment SDG predicted mean sponsorship rate of 38 percentage points, sponsored shares increase by approximately 4.6 percentage points (12%). The share increases are due to increases in the number of sponsored posts since the number of total posts per influencer does not change. Disclosure increases after the regulatory change, but there is still a substantial number of posts that are not disclosed: We show that the sponsorship rate among undisclosed posts increases. Timing tests show that effects are not driven by differences in pre-policy outcome trends (see Figures 1 and 2).

We also show that the regulatory change affects content engagement. The average number of likes per post increases for undisclosed-sponsored posts relative to non-sponsored posts. Through the lens of the model, this suggests increased consumer trust in more sponsored language. Finally, we show that both the mean number of likes and the mean number of comments that influencers in Germany receive falls after the regulatory change. The decreases in engagement are quantitatively large:

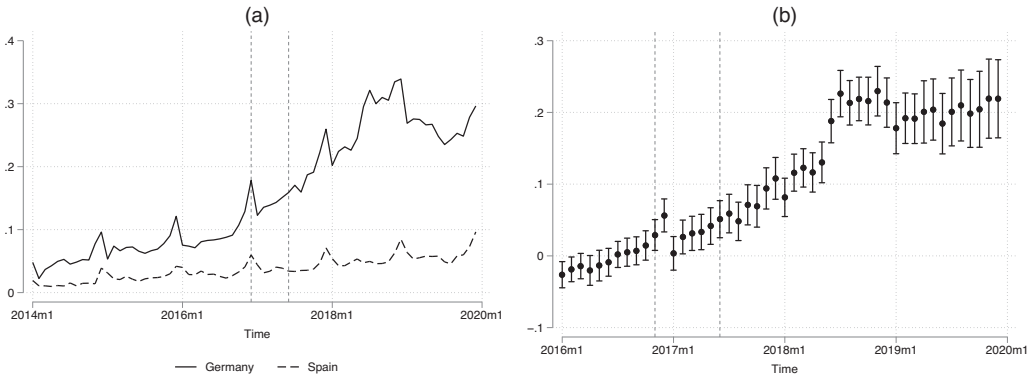

FIGURE 1 | Disclosed post shares in Germany and Spain. Each line in (a) shows the total number of posts labeled as "disclosed" advertising over the total number of posts in month t in Germany or Spain ($\frac{N \text{ Disclosed Posts}}{N \text{ Posts}}$). A post is labeled as disclosed if it includes one of the disclosure words from Appendix A3.1. A CEM-matched sample of influencers is used. (b) Estimates from a difference-in-differences regression with heterogeneous monthly treatment effects and with placebo effects starting in January 2016 (Equation 3). The baseline periods for this regression are 2014 and 2015. 95% confidence intervals are shown. The first dashed vertical line represents the initial changes to German disclosure regulations in November 2016 (see Section 2.1). The second dashed vertical line represents the first fines handed out to German influencers in mid 2017.

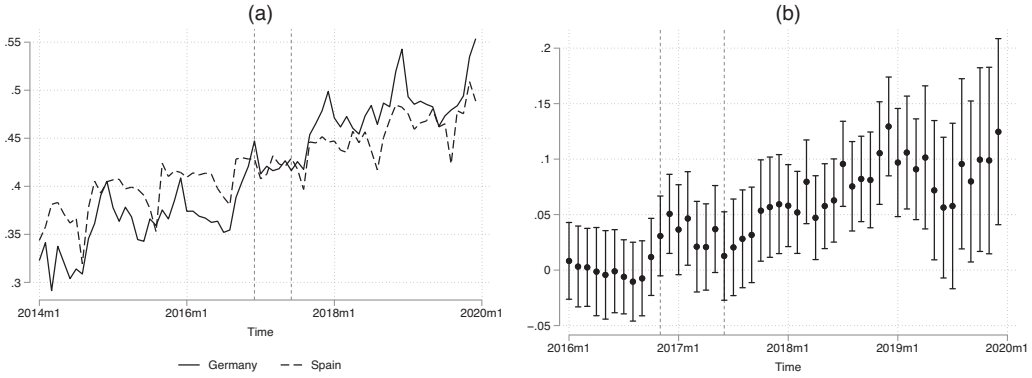

FIGURE 2 | SGD-predicted sponsored post shares in Germany and Spain. Each line in (a) shows the average share of predicted sponsored posts of the total number of posts ($\frac{N \text{ Predicted Sponsored Posts}}{N \text{ Posts}}$) between Germany and Spain in month t according to the SGD classifier. A CEM-matched sample of influencers is used. (b) Estimates from a difference-in-differences regression with heterogeneous monthly treatment effects and placebo effects starting in January 2016 (Equation 3). The baseline periods for this regression are 2014 and 2015. 95% confidence intervals are shown. The first dashed vertical line represents the initial changes to German disclosure regulations in November 2016 (see Section 2.1). The second dashed vertical line represents the first fines handed out to German influencers in mid 2017.

On average, likes fall by over 50% relative to the baseline mean. Engagement is an imperfect proxy of attention in our model, so these results are suggestive, but they are consistent with the notion that followers are made worse off by the disclosure rules.

The contributions of this article are three-fold. This is the first empirical article looking at the effects of changes in online advertising disclosure regulations on the equilibrium amount of advertising in a market where there is no direct compensation between the advisor and advisee. Theoretical predictions are counter-intuitive (i.e., stricter regulations increase ads) and have not been tested empirically. There is also widespread skepticism in the popular press about the effectiveness of such regulations (TheGuardian.com). Previous empirical literature such as Sahni and Nair (2020) focused on the demand response of consumers to the disclosure of advertising. We show that while disclosure regulations have an effect on actual disclosure, they also influence content production, possibly adversely.

We also present a novel mechanism for why sponsorship increases, drawing on the imperfection of online regulations and a distinction between disclosed- and undisclosed-sponsored posts. Our findings are relevant for the broader question of regulation of online markets, and even platforms like Google Search. Google Search has a mix of "authentic" (organic) results and sponsored content. Some sponsored links on Google are disclosed advertisements, but some are links to Google-owned products ("Google Shopping" or YouTube) *within* the organic results. Such links also compensate Google, potentially without consumer knowledge. Many other e-commerce websites and platforms featuring product reviews also have a mix of "authentic" and paid-for/fake reviews or discussions (e.g., Godes and Mayzlin 2009; Mayzlin 2006; He, Hollenbeck, and Proserpio 2022). The trade-offs we identify for influencers may apply to larger platforms as well. Our findings suggest that forcing platforms to disclose only one form of advertising (or revenue-generating content) may increase the amount of *total* advertising that consumers are

exposed to (including more hidden paid-for content). This is a key concern for policy-makers and regulators.

Our article also contributes methodologically to the empirical economics literature that uses text as data (e.g., Gentzkow, Shapiro, and Taddy 2019; Hansen, McMahon, and Prat 2018; Ash, Chen, and Naidu 2024). Our combined use of multi-lingual embeddings and supervised machine learning classification is novel to this literature. Such methods allow for straightforward comparisons of changes in text meaning and expand the range of possible future analysis to be done on large cross-country text-based datasets.

The article proceeds as follows. Section 1.1 gives an overview of the related literature. Section 2 describes the industry background and the regulatory change we study. Section 3 presents the raw Instagram data we use in the article and discusses the classification of sponsored content. Section 4 presents our theoretical model. Section 5 describes the estimation approach, presents some descriptive evidence, the main difference-in-differences regression estimates, and discusses various robustness checks. Section 6 concludes.

1.1 | Related Literature

To our knowledge, there is no existing empirical research on the effects of advertising regulations on content production online/on social media.[11] There are several existing theoretical studies on this topic.[12] Fainmesser and Galeotti (2021) set up a static matching model with many followers and influencers. There is asymmetric information between followers and influencers: Influencers can provide sponsored or authentic content to the followers, and followers are not aware of the content type until they "consume." Followers decide on who to follow based on the degree of authenticity of the influencers; sponsored content (which is foreseen in equilibrium) brings less value to followers. There are also matching frictions due to follower search costs. Influencers differ from one another vertically—some provide better content than others. Influencers with higher quality are more likely to have more followers and also more sponsored content. In fact, the biggest influencers in this model oversupply sponsored content in equilibrium. Mandatory disclosure policies in this world make sponsored content less costly for followers. This can increase sponsored content because followers are now less sensitive to the composition of organic versus sponsored content because they can ignore sponsored content. At the same time, there is a loss of followers in equilibrium because of reduced content quality. Overall, this model predicts that sponsored content increases and total welfare falls in the market after transparency. Their model is silent on undisclosed-sponsored content, which we study, as they assume that all sponsored content is disclosed.

Mitchell (2021) sets up a dynamic mechanism design model between a follower (the principal) and an influencer (the agent). The influencer receives "ideas" at some Poisson rate and can perform one of two actions: (1) post something "authentic," which gives her zero payoffs and the follower positive payoffs, or (2) post something "sponsored," which gives her non-zero payoffs and the follower zero payoffs. Posting authentic content is costly because it foregoes sponsorship. The follower chooses whether to follow the influencer or not based on the observed history of actions and the follower's beliefs about the influencer's future behavior. In equilibrium, the influencer rotates between periods of building up reputation by providing authentic content and periods of cashing in via sponsored content. The key for the influencer's strategy is not to provide sponsored content for too long so that the relationship does not break down permanently. Mitchell (2021) mimics disclosure regulations through a counterfactual that lowers the influencer's returns for posting sponsored content. Because this lowers the return to sponsored content, it also reduces the return to improving the relationship with followers by providing organic content. This can lead to more or less sponsored content in equilibrium. Mitchell (2021) also does not focus on undisclosed, sponsored posts.

Pei and Mayzlin (2021) also study recommendations by influencers. In their article, the influencer faces an explicit informational model in persuading a potential consumer. They focus on the equilibrium informativeness of the influencer's signal and how that determines the way in which firms optimally communicate with consumers. Critical in their model is how statements from an influencer lead to information for followers. In Pei and Mayzlin's (2021) model, some form of credible commitment to what is and is not endorsed (like an FTC rule) is necessary for the market to function. In our model, conditional on post type, informativeness is exogenous; it is implicit in the followers' desire to engage with organic content more than sponsored content. This means we are envisioning forces that Pei and Mayzlin (2021) model operating behind the scenes, rather than modeling them directly. However, the forces in the two models are similar. The equilibrium amount of sponsored content in our model determines followers' engagement, because it impacts the informativeness of the average post, even though informativeness conditional on post type is exogenous. In our model, influencers are small, and therefore, the actions of influencers cannot, individually, impact the average informativeness of the messages.

Focused on a different application, Inderst and Ottaviani (2012) study a static model of regulating advice, especially in financial markets. In their model, the reason for the adviser to want to give some good advice is exogenous, but the nature of the static relationship is modeled in much more detail. Disclosure can reduce welfare because it undoes the information value that advisers sometimes have.

Prior empirical literature has studied the impact of disclosure regulations on paid intermediaries—for example, in the market for insurance advice (Bhattacharya, Illanes, and Padi 2019) and for financial advice (Anagol, Cole, and Sarkar 2017). In these markets, however, there is direct compensation between the intermediaries and consumers. Changing disclosure rules may then have different effects. Unlike these markets, we also have clear indicators of disclosure and can see when disclosure does not happen. Lack of compliance and incomplete disclosure is critical in our market and in most online markets but is less of a focus in financial markets. On the other hand, we have less information about outcomes; we have only indirect measures via follower engagement. Our focus, however, is on the supply of sponsored advice.

Using a field experiment, Sahni and Nair (2020) vary disclosure for a collection of restaurant ads and find that the disclosed ads led to a greater response from consumers. This is consistent with the signaling value of paid advertising. Anecdotal evidence, survey-based measures, and our text-based analysis show that influencers fail to fully comply with disclosure regulations, suggesting that disclosure does not provide an unambiguous positive effect. Their setting is similar to ours in that it is an online advertisement; a key difference in our setting is that the post comes from an intermediary. Information-providing intermediaries have long worked to avoid the appearance of being advertisement-driven,[13] presumably to highlight the informativeness of their message. These intermediaries typically monetize their advice through subscriptions; the small volume of each individual piece of advice being given in our context makes such an arrangement difficult, and therefore, other channels of monetization are necessary.

The literature on advertising has long highlighted the informative content of ads as another reason, beyond signaling, for ads to be effective. For instance, Horstmann and MacDonald (2003) study the information content of explicit advertisements. Our setting highlights a tension between informativeness and signaling for intermediaries that provide information and advertisements side by side.

Our work is also related to a large literature on sponsored (or firm-created) word of mouth, which influencers contribute to. One early model of sponsored word of mouth is Mayzlin (2006). In that model, an exogenous stock of organic posts from real consumers is mixed with posts directly from firms, which are effectively like sponsored content. We extend this idea to an explicitly language-based model that maps to our empirical application, where there is a trade-off for influencers between organic and sponsored content. Empirical work has studied word of mouth. Godes and Mayzlin (2004) study word of mouth on discussion boards, which are thought to be largely free of modern influencer motives. Godes and Mayzlin (2009) use a field experiment to measure the impact of paid content on word of mouth. They highlight the difference between word of mouth generated by customers, who have actual knowledge of the product, and agents that are hired for the word of mouth without prior experience with the product. Understanding better the heterogeneity in influencers and what they produce is a critical question that is not directly addressed by our research.

2 | Background

2.1 | Background: Advertising Disclosure Regulations in Germany

Social media advertising regulations are not standardized across EU countries.[14] There are existing national and EU-wide advertising disclosure regulations that apply to traditional media such as newspapers and television. The Unfair Commercial Practices Directive (UCPD) from 2005 specifically regulates potentially misleading omissions such as ambiguity about transactional relations between a commercial "trader" and an advertiser (Ducato 2020). The problem is that most influencers cannot be simply defined as "traders"—a travel influencer posting pictures of herself on trips does not obviously have commercial interests.[15] Since 2008, there have also been some "best practices recommendations" on social media advertising provided by the European Advertising Standards Alliance (EASA), a collection of national European self-regulatory organizations (EASA-Alliance.com). These are non-binding, and each national body is free to pick and choose which guidelines apply.

In different countries, influencer marketing is regulated by consumer watchdogs, advertising authorities, or competition authorities. Jurisdiction is based on influencer residence—influencers who live in Italy are subject to Italian regulations. Below, we describe changes to the German regulatory environment. To the best of our knowledge, there are no online advertising disclosure regulations in Spain beyond the baseline non-binding EU regulations or any changes in the regulatory environment during our sample period.

In October 2016, Die Medienanstalten, a consortium of 14 German state media authorities responsible for the licensing and supervision of media, released a set of "clarifications" for advertising on social media (Osborne-Clarke.com). The clarification emphasized that existing laws governing the disclosure of paid advertising on traditional media also apply to social media markets and to influencers. In Germany, these laws (i.e., the German Marketing Law - UWG - also known as the Unfair Copmetition Law) are enforced by competition authorities. The October 2016 release by Die Medienanstalten also provided guidelines for compliant disclosure of advertising on social media: labeling any posts where the influencer has been remunerated by a brand, including free products, as an ad with a visible hashtag.

The role of Die Medienanstalten is comparable to the FCC in the United States or OfCom in the United Kingdom.[16] To the best of our understanding, there were no changes to actual advertising laws in Germany as a result of the October 2016 statement. However, this clarification appears to have triggered a wave of legal activity against German influencers in 2017 and 2018. Under German advertising laws, both the advertising platform (the influencer) and the advertiser are financially liable (mediawrites.law). Among other examples, a German YouTube fitness influencer was fined over 10k EUR for failing to disclose a video as advertising in June 2017 (ISLA.com). Also in 2017, a court in Hagen fined an Instagram fashion influencer and forced her to start adding "#ad" to posts that were paid for by fashion brands. Court decisions directly cited Die Medienanstalten's October 2016 statement and disclosure guidelines.

Even after the "clarification," there was still disagreement about the interpretation of existing laws and the extent to which they apply to different influencers and posts. Various courts had different rulings. In 2018, a court in Berlin ruled that if the purpose of an influencer is merely to keep followers updated about trends, even posts not directly linking to brands can have commercial intent and should be labeled as advertising (Ducato 2020). At the same time, a lower court in another case had an even stricter interpretation of the law than what Die Medienanstalten suggested. This interpretation held that any post by an influencer who has previously used their account for commercial gain should be considered a commercial post and labeled as an ad. This interpretation was overturned by an upper court of appeals.

Despite the remaining ambiguity, to the best of our understanding, the release of Die Medienanstalten's guidelines and subsequent legal activity created a regulatory environment where many influencers had legitimate concerns of legal action and fines for non-disclosure of advertising content. By comparison, the release of a similar document in December 2015 by the FTC in the United States resulted in several formal complaints against large advertisers (Warner Brothers in 2016, CS:GO Lotto in 2017) and a single financial settlement with another popular advertiser, Teami, for $1 million in 2020 (ftc.gov). The FTC also issued several warning letters to several celebrities. "Clarifications" of existing advertising disclosure laws similar to Germany's were also done in Italy and France in 2018, but these resulted in little legal activity that was limited to several top influencers.[17] Spain, which we use as our "control" group, maintained a lax social media advertising disclosure regime throughout our sample period. More generally, we believe that the German regulatory environment, which combines a clear risk of non-disclosure with ambiguous interpretations of compliance guidelines, reasonably captures the intensity of disclosure regulations for a broad set of online intermediaries. For example, Google knows that it is required to disclose paid search results, but disclosure requirements for organic search links that it earns revenues from (e.g., links to YouTube, Google Flights, or hotel links) are more ambiguous.

3 | Data

3.1 | Data Description

We collect data from CrowdTangle, which describes itself as "a public insights tool owned and operated by Facebook" (CrowdTangle.com).[18] Our raw data are at the post level. We observe a full history of posts for each influencer in our sample. For each post, we observe the text of the post, the user name of the influencer, the date of the post, the number of likes, the number of comments, and some post characteristics (i.e., is it an image or a video). We do not record the image associated with the post.

Our sample consists of randomly selected 6000 German and 6000 Spanish influencers provided by HypeAuditor, a leading online influencer marketing firm (Hypeauditor.com). Each influencer in this list has at least 10,000 followers by May 2019 (the date when HypeAuditor selected the data). In the raw data, we observe posts from 2010 until 2020.[19] Each influencer is local to their country—they live in Germany or Spain and a majority of their followers are from their country of residence.[20] This is important to make sure that influencers are only affected by laws of the country in question, rather than laws in other countries.[21] Spanish followers' conception of the world is also not being changed by regulation in Germany, since most Spanish followers are not reading German posts.[22]

We collect additional country–year/month-specific data. Germany and Spain are different in many respects, which could affect the amount of advertising posted by influencers. We collect quarterly data on per-capita income and population from the OECD. We also proxy for the time-varying popularity of Instagram in each country using monthly Google Trends search query volumes for the keyword "Instagram" from Germany and Spain between 2014 and 2020.[23]

Our raw data do not provide us with a specific identifier for sponsored posts or for posts that are disclosed as sponsored. [24]

3.2 | Detecting Disclosure

We detect disclosed-sponsored posts by searching caption text for words that were recommended by German regulators to disclose advertising online, such as #ad, #ambassador, and their German equivalents. To be as conservative as possible, we include additional words that come from national and international advertising guidelines. We also translate all recommended disclosure words into Spanish and search Spanish posts for them as well. A full list of words is in Appendix A3.1.

3.3 | Detecting Sponsorship

While we can uncover disclosed-sponsored posts as per Section 3.2, undisclosed-sponsored posts are by definition hidden.[25] Such posts are likely popular in Spain, where there are no regulatory disclosure incentives. There is also reason to suspect that some posts in Germany after the regulatory change are sponsored but undisclosed.[26]

We propose two approaches for identifying sponsored posts using text (post caption) data:[27] (1) A "manual" approach using a list of pre-determined keywords, which generally denote commercial activity/sponsorship. (2) A supervised "automatic" approach using machine-learning classifiers.

In the first approach, we use translations of English, Spanish, and German words. These include references to coupons, contests, or discount codes, as many sponsorships allow influencers to offer discounts for products. Relevant keywords also include any links to outside websites (anything that ends with ".com," ".de," ".es"), references to shopping ("shop []," "compra []," etc), references to products, or to availability (i.e., "out now"). A full description of the words we use is in Appendix A3.2. The manual method assumes that if certain words are present, the message is sponsored. It also assumes that the set of words that denote sponsorship is known to the researchers.

In the second approach, we train supervised ML classification algorithms on labeled data and project the trained algorithms on non-labeled data. A randomly selected 300,000 German post sample from 2018 serves as our training data.[28] Since we know that disclosed posts are sponsored, the algorithms look for language associated with disclosure.

There are several concerns with our text-based approach to detecting sponsorship, related to the language German influencers use for organic and sponsored content. First, it is not clear that ML classifiers are able to accurately separate undisclosed-sponsored content from disclosed-sponsored content in our training data. The only labels we have are for disclosed-sponsored posts, and the classifiers predict disclosure. However, influencers likely do not fully comply with regulations. In that case, similar posts could be disclosed and undisclosed, making accurate prediction challenging.

Second, there is the possibility that ML classifiers will not be able to accurately separate undisclosed-sponsored and organic posts. This could happen if influencers choose to use different language for disclosed-sponsored and undisclosed-sponsored posts to mislead or confuse consumers. For example, they could add organic language to undisclosed posts to balance their commercial nature. In that case, ML classifiers could mislabel undisclosed-sponsored posts as organic.

Finally, even if ML classifiers accurately separate sponsored and non-sponsored content in the training data—the "post" period for Germany—they may not perform as well in the "pre" period in Germany or in Spain. This is both because language naturally changes over time and because influencers can make purposeful, strategic changes to the language they use in response to the changing regulatory environment. Disclosed-sponsored language from the "post" period in Germany may not be representative of sponsored language in unregulated markets (the "pre"-period Germany or Spain). A change in the predicted number of sponsored posts could, therefore, be caused by a changing language rather than changing sponsorship.

We attempt to address the first concern by tuning classifiers' parameters to focus on correctly labeling true-positive outcomes (i.e., actual disclosed-sponsored posts) while being more forgiving of false-positive outcomes.[29] In this application, false-positive outcomes are undisclosed-sponsored posts. Therefore, our classifiers effectively attempt to distinguish between all posts that have similar language to disclosed-sponsored posts and all posts that do not.

We address the second concern in part by using a flexible measure of language. We transform each post into a *multilingual embedding space*, where the post is represented by a 300-dimensional continuous vector in linguistic-meaning space.[30] Rather than attempting to measure which specific words or phrases predict disclosure, our classifiers more flexibly measure what linguistic meaning predicts disclosure (e.g., commercial meaning). Even with some obfuscation by influencers, undisclosed-sponsored posts should likely maintain some "sponsored" meaning, due to advertisers having some control over sponsored post content.[31] This suggests they would be nearer to disclosed-sponsored posts in embedding space than to organic posts.[32]

We address the concern about the stability of language by using a SGD classifier as our main ML classifier. This classifier has both good prediction quality and a stable prediction rate for disclosed-sponsored posts, which suggests that it is less sensitive to language changes over time.[33] We also show that conditional on the SGD classification, the average embeddings in the organic and undisclosed-sponsored classes are not moving over time in Germany or Spain.[34] This suggests that language is reasonably stable over time. What changes is the composition of posts—that is, the number of posts assigned to each class.

Finally, in the main analysis, we also use a combination of the two approaches, labeling a post as sponsored only if both the ML classifier and the manual approach classify it as sponsored. We do this because the ML classifier potentially disciplines our broad manual keyword selections. Conversely, the manual classifier potentially disciplines the ML classifier—restricting the set of sponsored posts to those that have some words denoting commercial intent. The manual classifier should also not be as sensitive to changes in language over time, since it uses a preselected set of commercial-language keywords that do not come from any particular time period or country.

Nonetheless, we acknowledge remaining concerns that our classification process will not fully be able to accurately distinguish changes in language over time from changes in sponsorship. As we discuss below in Section 4, we are confident in identifying the direction, but not the precise magnitude, of changes in sponsorship. This would only be fully addressable with a dataset containing a set of known labeled undisclosed-sponsored posts and organic posts, which we do not observe.

3.4 | Matching and Influencer–Month Summary Statistics

After the classification procedure described in Sections 3.2 and 3.3, we merge post-level data with monthly country-level data. Post-level data include dummy variables of whether each post is classified as sponsored based on each one of the classifiers described above, as well as a dummy variable of whether each sponsored post is disclosed as sponsored. We then aggregate the merged data to the influencer/month level. We restrict our time period to be from 2014 to 2020. Although some influencers in our sample have been active since 2010, the vast majority were not. In our regression analysis, we further restrict the data to consider only monthly observation of influencers with more than two posts in a month. With two posts or one post, many of our outcomes are too noisy and results could be driven by outliers.[35]

We further restrict our sample by matching German influencers to observationally similar Spanish influencers using 2015 data. We use Coarsened Exact Matching (CEM), a matching method commonly used in economics literature to balance covariates.[36] After matching, we are left with approximately 600 influencers from Germany and a similar number from Spain.[37] Additional details, including covariate-balance tests, are in Appendix A6. Summary statistics for the CEM-matched sample are in Table A4.[38]

4 | Model

This section introduces a model of influencer and follower behavior. The model allows us to characterize market equilibria and relate our measurement of sponsored content to an equilibrium object, to better explain how to interpret what we measure.

A post is a collection of words $\omega \subset \Omega$, where Ω is the dictionary of all possible words.[39] To model the contrast between sponsored and organic posts, we assume that these posts are generated by a mixture between two distributions, $i \in \{A, B\}$ with common support. Distribution f_i generates a draw of a collection of words ω. A post's language is described by a mixture α where $f_\alpha = \alpha f_B + (1-\alpha) f_A$.

We model post-formation and attention as a two-stage game with private information. First, an influencer picks whether to

post sponsored or organic content, and the language α to use. Their return to sponsored content and the choices are private information. Given α, there is a draw of words ω. In the second stage, followers observe ω and draw inferences about the type of post to choose attention. Attention is costly and followers prefer to pay attention to organic posts, other things equal. The initial discussion does not explicitly discuss regulations, but in Section 4.3, regulation is taken up explicitly.

4.1 | Influencers

In the first stage of the game, influencers privately draw a value γ, distributed uniformly on the unit interval, that describes how valuable sponsorship is at that moment. They then decide, also privately, whether to make a sponsored or organic post and the language to use. Both choices are driven by a trade-off between revenue and follower attention.[40] The influencer's payoff, fixing the type of post, is proportional to attention $x(\omega) \geq 0$ by followers given words ω. The determination of $x(\omega)$ will come from the followers and is described in Section 4.2 below. A sponsored post is worth, for fixed attention, $V(\alpha, \gamma)$ times as much as an organic post if it uses a mixture α of the B distribution when the draw is γ. Normalize higher γ to have a higher sponsored return, so the derivative $V_2 > 0$. Assume that $V_1 > 0$: that is, for a given level of attention, using the words from the B distribution generates more return, when you are a sponsor you should say words like "sale." In the absence of this assumption, there would be no tension between language for the poster. Further, let $V_{11} < 0$ for concavity. The sponsor, therefore, impacts but does not determine the language of posts, even when they are sponsored.

The language choice for a sponsored post solves

$$\max_{\alpha} V(\alpha, \gamma) \int x(\omega) f_{\alpha}(\omega) d\omega$$

When expected attention is higher for the A distribution, as will occur in equilibrium (see proof in Appendix A7.2), organic posts simply maximize attention by choosing the A distribution. The influencer chooses whether to make a sponsored post by comparing its return $V(\alpha, \gamma) \int x(\omega) f_{\alpha}(\omega) d\omega$ to $\int x(\omega) f_A(\omega) d\omega$. We can summarize the optimal decisions in terms of the relative expected attention of A versus B words, $\theta = \frac{\int x(\omega) f_A(\omega) d\omega}{\int x(\omega) f_B(\omega) d\omega}$.

Proposition 1. *For $\theta \geq 1$, the optimal $\alpha(\gamma, \theta)$ for a sponsored post is decreasing in θ and solves*

$$\frac{V_1(\alpha, \gamma)}{V(\alpha, \gamma)} = \frac{\theta - 1}{\alpha + (1 - \alpha)\theta}$$

The influencer posts organic content if it draws $\gamma < \gamma(\theta)$ where $\gamma(\theta)$ solves

$$V(\alpha, \gamma(\theta)) = \frac{\theta}{\alpha + (1 - \alpha)\theta} \tag{1}$$

Proof of Proposition 1 is in Appendix A7.1.

From (1) it follows that higher θ leads to a higher cutoff $\gamma(\theta)$ for posting sponsored content for any attention $g(\omega)$. An important implication is that sponsored content becomes more similar to organic content (α decreases in θ) as sponsored content decreases ($\gamma(\theta)$ increases in θ). Intuitively, when followers are more likely to engage with words from the B distribution, the influencer has an incentive both to post more sponsored content and to make sponsored content draw more heavily from the B distribution. The equilibrium determination of γ depends on follower attention, which we model next.

4.2 | Consumers/Followers

The follower chooses how much attention to pay to a post based on their beliefs about its type given ω. Their return to each unit of attention is h for organic posts and l for sponsored posts, where $h > l > 0$: followers prefer organic content. This conforms to the idea that informativeness is not constant across messages, as modeled in Pei and Mayzlin (2021). The convex cost of attention $x(\omega)$ on a post of words ω is $c(x)$; since attention is costly, the follower prefers attention to likely organic posts. Let the follower's posterior probability of a post being organic be $g(\omega)$. Given a set of words, they choose attention $x(\omega)$ to solve

$$x(\omega) = arg \max_{x} (g(\omega)h + (1 - g(\omega)l)x - c(x)$$

Attention is increasing in g; let $x(0) > 0$, since even posts that are disclosed (and therefore presumably known to be sponsored) get some attention, as measured by engagement.[41]

4.3 | Equilibrium, Regulation, and Measurement

4.3.1 | Equilibrium

Equilibrium requires that attention and sponsorship are optimized, given one another. Let the belief about a post of words ω given the equilibrium posting cutoff γ for influencers be $g(\omega|\gamma)$. Then, the relative attention to the two distributions, which determines θ, is

$$\Theta(\gamma) = \frac{\int x(g(\omega|\gamma)) f_A(\omega) d\omega}{\int x(g(\omega|\gamma)) f_B(\omega) d\omega}$$

Definition 1. A cutoff γ^* and attention ratio θ^* is an equilibrium if $\gamma^* = \gamma(\theta^*)$ and $\theta^* = \Theta(\gamma^*)$

4.3.2 | Regulation

To introduce regulation into the model, assume that along with sponsorship level γ, with probability ρ, the sponsored post is disclosed using words never present in the A distribution. If a post is disclosed, there is no reason to choose words from the A distribution, so for those posts, it is natural to assume $\alpha = 1$. This is also consistent with the notion that disclosed posts often are very "obviously" sponsored even outside of disclosure itself, for instance in the #ad post by Kylie Jenner in Figure A1. When a sponsored post, if chosen, is not disclosed, the same condition describes the cutoff $\gamma(\theta)$ above which posts are sponsored. If a post must be disclosed, however, the cutoff is higher, since the

return to the sponsored post is lower when it is disclosed. We define this higher cutoff as $\gamma^d(\theta)$.

Holding language fixed, the influencer faces a trade-off under a disclosure regime: Disclosed-sponsored posts receive low attention, reducing the productivity of sponsored posts. At the same time, disclosure can also increase trust in the sponsored language (lower θ), increasing the productivity of undisclosed-sponsored posts. The net effect is ambiguous (which is especially true when influencers can also endogenously adjust their language), but this trade-off intuitively explains the forces that can lead to more sponsored posts in the regulated equilibrium as compared to the unregulated equilibrium.

4.3.3 | Measurement

Our measurement exercise seeks to uncover how θ^* varies across locations and policies, since it determines the incentive to post sponsored content. Suppose we measure sponsored content imperfectly via some $h(\alpha)$, which is an increasing function of the post's chosen language α. Then, the total measured sponsored content under θ^* is

$$M(\theta^*) = \gamma(\theta^*)h(0) + \int_{\gamma>\gamma^*(\theta)} h(\alpha(\gamma,\theta^*))d\gamma$$

Since both $\alpha(\gamma,\theta)$ is decreasing and $\gamma(\theta)$ is increasing in θ, higher θ makes sponsored posts harder to detect and posts less likely to be sponsored:

Proposition 2. *$M(\theta^*)$ is decreasing*

M can wrongly classify non-sponsored posts as sponsored (overcounting for $h(0)$) and does not always correctly classify sponsored posts as such (undercounting for $h(\alpha) < 1$) and the fractions of these change as equilibrium changes. However, it is always the case that a higher measure of sponsorship indicates lower θ, which in turn implies, by (1), more actually sponsored content. This is true even though language varies as θ^* varies.

Under regulation, there is a distinction between disclosed-sponsored and undisclosed-sponsored posts (and of course, organic posts). This measured share is

$$M^r(\theta^*) = \frac{\rho\gamma^d}{1-\rho+\rho\gamma^d}h(0) + \frac{1-\rho}{1-\rho+\rho\gamma^d}M(\theta^*)$$

Since $h(0) < M(\theta^*)$,

Proposition 3. $M^r(\theta^*) < M(\theta^*)$

The proposition highlights another force against finding more measured sponsored content among the undisclosed posts in the regulation period: That some sponsored content that would have been posted is replaced by organic posts when disclosure is required. The measurement of more content as sponsored among undisclosed posts under regulation points to lower θ^* in that period, that is, a higher incentive to post sponsored content based on attention.

4.4 | Model Summary and Implications

To summarize, the model centers around an equilibrium statistic (θ) that measures follower trust in organic words relative to sponsored ones. Proposition 1 shows that this equilibrium statistic is related to both influencer language choice in sponsored posts and the volume of sponsored posts, in a monotone way. Proposition 2 then describes how even imperfect measurement of changes in the amount of sponsored content—such as our shares of posts classified as sponsored (see Section 3.3)—can recover changes in equilibrium θ and the true amount of sponsored content. Finally, Proposition 3 shows that increases in undisclosed sponsored content after regulation point to the equilibrium statistic making sponsored posts more favorable.

There is a distinction between attention and engagement, where the former could include many activities whereas the latter involves specific actions including likes and comments. Engagement is a subset of all attention and therefore is an indirect measure of attention. We use the term attention here despite the fact that in the empirics, the only analog is measures of engagement. We think of the model and the mechanism as applying to more general forms of attention but will focus on engagement when we attempt to measure the impact of policy on followers directly.

5 | Estimation Methodology and Results

5.1 | Estimation Methodology

Changes in the regulatory environment in Germany but not in Spain at the end of 2016 suggest a difference-in-differences estimation strategy to identify the effects of stronger disclosure regulations. We compare influencers in a country where disclosure regulations were strengthened (Germany) to influencers in a country where disclosure regulations have not been implemented (Spain) before and after the changes.

We aggregate our outcomes at the influencer and month level.[42] We model outcome Y_{it} (i.e., share of sponsored posts) for influencer i at month t as:

$$Y_{it} = \alpha(\text{Germany}_i \times \text{Treated Period}_t) + \beta X_{it} + \delta_i + \delta_t + \epsilon_{it} \quad (2)$$

where Germany_i is a variable equal to 1 for all German influencer observations, and Treated Period_t is a variable equal to 1 for all observations after November 2016.[43] X_{it} are a set of influencer/time varying controls, such as account age and country characteristics (i.e., popularity of Instagram, GDP per capita). δ_i and δ_t are influencer and year/month fixed effects that absorb country and Treated Period_t fixed effects. We include a number of country/time-varying controls (X_{it}) to try to capture country–time-specific shocks, such as the popularity of Instagram and GDP per capita (which may influence consumption or advertising behavior).[44]

We also test for anticipation effects or diverging pre-regulation influencer behavior in the two countries with a formal timing test.

We estimate the regression:

$$Y_{it} = \sum_{t=\text{Jan 2016}}^{t=T} \alpha_t(\text{Germany}_i \times \text{D}_t) + \beta X_{it} + \delta_i + \delta_t + \epsilon_{it} \quad (3)$$

where instead of having a single Treated Period$_t$ dummy, we have a set of month dummies (D_t) ranging from 10 periods before any regulatory changes in Germany (January 2016) to the final period in our sample (December 2019, which we define as $t = T$). α_t estimates the monthly differences in Germany relative to Spain in month t as compared to a baseline period—2014 and 2015.

5.2 | Descriptive Evidence and Timing Regressions

5.2.1 | Disclosure in Germany and Spain

Figure 1 shows the percentage of all posts disclosed as advertising in Germany and Spain during our sample period in panel (a) and results from a difference-in-differences regression with heterogeneous timing effects in panel (b). The regression in panel (b) estimates Equation (3). There are two vertical lines in each panel of the graph. The first line represents the initial change in the regulatory environment in Germany in October 2016. The second vertical line represents the beginning of regulatory enforcement in Germany through fines to influencers in 2017 (see Section 2.1 for more details). Panel (a) shows that disclosure in Germany increases dramatically around changes in the regulatory environment. There are no similar changes in disclosure in Spain over the same period. Panel (b) shows that the differences in disclosure between the two countries are not systematically statistically different at the 95% confidence level until after changes in the regulatory environment.

5.2.2 | Sponsorship in Germany and Spain

We plot monthly shares of predicted sponsored posts in Germany and Spain in panel (a) of Figure 2. For each country, we compute the average share of predicted sponsored posts of the total number of posts ($\frac{\text{N Predicted Sponsored Posts}}{\text{N Posts}}$) in month t. In panel (b), similar to Figure 1, we plot the results of estimating Equation (3) with the share of SGD-predicted sponsored posts as the dependent variable.

Figure 2 suggests that there is a change in sponsorship rates between Germany and Spain after the strengthening of disclosure regulations in Germany (the two vertical dashed lines represent changes in disclosure regulations and their enforcement). Sponsorship in Germany increases relative to sponsorship in Spain after regulations are strengthened. This corresponds to an absolute increase in sponsorship in Germany, consistent with the theoretical model, and suggests that content in Germany is changing in response to changes in the regulatory environment.[45] Notably, comparing panel (a) in Figure 1 and panel (a) in Figure 2, sponsorship and disclosure are very different between the two countries. While Spain has substantially less disclosure than Germany, especially after regulations are introduced, it has as much (if not more) sponsorship. This is consistent with the underlying notion that without substantial legal incentives to disclose sponsored content, most such content in Spain is undisclosed. This comparison also suggests that we are uncovering something novel about the underlying text data through our classification, rather than simply re-stating changes in disclosure.

Panel (b) of Figure 2 shows that differences in sponsorship rates between Germany and Spain in the first 10 months of 2016 are not statistically different at the 95% confidence level as compared to 2014 and 2015. However, starting from November/December 2016, sponsorship rates in Germany are statistically significantly higher than in Spain.

5.3 | Average Effects

This section shows Average Treatment Effect estimates from the difference-in-differences regression. Table 1 shows the effects of intensified disclosure regulations on advertising/sponsorship rates and on the rates of sponsorship among posts that are not disclosed as sponsored. Table 2 shows the effects of changes in disclosure regulations on other outcomes at the influencer/month level, such as the mean number of posts. Table 3 shows the effects of disclosure regulation on follower engagement measures: the average number of likes received by an influencer, the average number of comments, and the average number of followers. Table 4 shows the effects of changes in disclosure regulations on the ratio of the average number of likes for undisclosed-sponsored posts over the average number of likes for non-sponsored posts.

Table 1 shows the first set of results. The outcome variables in the table are all shares. In the top panel, the outcome variable is the number of predicted sponsored posts over the number of total posts for an influencer/month observation. In the bottom panel, the outcome variable is the number of predicted sponsored-undisclosed posts over the total number of undisclosed posts for an influencer/month observation. Each column uses a different approach to label posts as sponsored. Column (1) uses SGD, column (2) uses the manual approach, and column (3) uses a combination of the SGD and manual approach: A post is only labeled as sponsored if both the SGD classifier and the manual approach classify them as sponsored.[46] All regressions control for influencer and time fixed effects, as well as flexible influencer account age controls. These controls allow for cohort effects depending on when the influencers became active on Instagram. We also include country-level controls—population, GDP per capita, and a Google Trends search intensity for the term "Instagram" as a control for Instagram's popularity in Germany and Spain. Standard errors are clustered at the influencer level.

Estimates in this table show a statistically significant increase in the share of sponsored posts in Germany after the strengthening of disclosure regulations. This result holds for the ML classifier, the manual classifier, and for the combination. The changes in sponsored post shares are large in magnitude. Pre-treatment means are 38, 45, and 22 percentage points for the SGD, manual, and combined classifiers, respectively. The increases in sponsorship after the regulatory change are between 2 and 4.6 percentage points. At a minimum, the relative increase is 4.5%, and at a maximum, it is approximately 20%. This is consistent with the prediction of Fainmesser and Galeotti (2021) that sponsorship

TABLE 1 | Influencer/month DiD estimates—sponsored share.

	(1)	(2)	(3)
Outcome:	**Predicted sponsored shares**		
Classifier:	**SGD L1**	**Manual**	**SDD L1 + Manual**
Germany × Treated Period	0.046***	0.019**	0.045***
	(0.008)	(0.009)	(0.008)
Pre-treatment mean	0.382	0.452	0.216
Observations	67,235	67,235	67,235
R-squared	0.522	0.556	0.571
Outcome:	**Predicted sponsored shares | Non-disclosure**		
Classifier:	**SGD L1**	**Manual**	**SDD L1 + Manual**
Germany × Treated Period	0.025***	0.004	0.019**
	(0.008)	(0.010)	(0.008)
Pre-treatment mean	0.373	0.444	0.207
Observations	65,984	65,984	65,984
R-squared	0.474	0.519	0.515
Country controls	YES	YES	YES
Influencer FE	YES	YES	YES
Year–month FE	YES	YES	YES
Account age FE	YES	YES	YES
Account age × First-account-year FE	YES	YES	YES

Note: Sample includes influencer/month-level observations from January 2014 to December 2019 with at least two posts in a month. Influencers in the sample are CEM-matched as described in Section 3.4. The dependent variable in each regression in the top panel is the number of posts that were labeled as sponsored for influencer i in month t as a share of the total number of posts made by influencer i in month t. The dependent variable in the bottom panel is the number of undisclosed posts that were labeled as sponsored over the total number of undisclosed posts. Each column uses a different classification approach to label posts as sponsored. In the data used for column (1), posts are labeled as sponsored by an SGD classifier. In the data used for column (2), posts are labeled as sponsored using the manual approach. In column (3), posts are labeled as sponsored if both an ML classifier and the manual approach classify them as sponsored. "Germany × Treated Period" is a dummy equal to 1 for all German influencer observations after November 2016 and 0 otherwise. All regressions include influencer and time fixed effects, as well as account age fixed effects and account age × first-account year fixed effects. Country controls include quarterly GDP per capita, quarterly population, and monthly measures of Instagram popularity based on Google Trends results. Standard errors are clustered at the influencer level. ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.

unambiguously rises with disclosure regulation, and a risk that can occur in Mitchell (2021) and the model presented in Section 4.

An important question about the results in the top panel of Table 1 is what happens to *undisclosed-sponsored* content. In our model as well as in Fainmesser and Galeotti (2021), disclosed-sponsored posts do not receive much attention, which improves welfare for a given level of sponsorship. Outcome variables in the bottom panel of Table 1 are shares of predicted sponsored posts conditional on non-disclosure: the number of undisclosed posts predicted as sponsored over the number of total undisclosed posts for an influencer/month observation. Point estimates in this panel are smaller than in the top panel, but estimates in columns (1) and (3) show a statistically significant increase in the share of *undisclosed* sponsored/advertising content consumers are exposed to.[47] This evidence suggests that disclosure regulations, which were designed to limit undisclosed-sponsored content, had the unintended consequences of actually increasing the share of undisclosed-sponsored content visible to consumers (as well as the amount of sponsored content more generally). Interpreting these results through the lens of our model suggests that as sponsorship (and undisclosed sponsorship) increases in response to regulations (θ from Section 4 falls), consumer welfare could decrease, if influencer language (α in Section 4) does not change drastically.

Table 2 shows difference-in-differences estimates with additional outcomes related to post content. Column (1) looks at disclosure rates—the number of disclosed posts as a share of the total number of posts for an influencer/month. This regression confirms the descriptive evidence in Figure 1 and shows that disclosure rates in Germany had statistically significant increases after disclosure regulations were strengthened. Disclosure increases by nearly 10 percentage points, on average. These changes were very large relative to the mean pre-regulation disclosure rate in the sample, 5%. The increase in disclosure reflects two different factors: the disclosure of existing sponsored content and the increase in disclosed and sponsored content. However, as Table 1 shows, there was also an increase in the proportion of undisclosed sponsored content consumers in Germany are exposed to.

TABLE 2 | Influencer/month DiD estimates—additional post content outcomes.

	(1)	(2)
Outcome:	**Disclosed share**	***N* posts**
Germany × Treated Period	0.091***	0.974
	(0.007)	(0.777)
Pre-treatment mean	0.0509	19.29
Country controls	YES	YES
Influencer FE	YES	YES
Year–Month FE	YES	YES
Account age FE	YES	YES
Account age × First-account-year FE	YES	YES
Observations	67,235	67,235
R-squared	0.576	0.579

Note: Sample includes influencer/month-level observations from January 2014 to December 2019 with at least two posts in a month. Influencers in the sample are CEM-matched as described in Section 3.4. "Germany × Treated Period" is a dummy equal to 1 for all German influencer observations after November 2016 and 0 otherwise. All regressions include influencer and time fixed effects, as well as account age fixed effects and account age × first-account year fixed effects. Country controls include quarterly GDP per capita, quarterly population, and monthly measures of Instagram popularity based on Google Trends results. Standard errors are clustered at the influencer level. ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.

Column (2) in Table 2 shows that the number of posts per month for influencers in Germany relative to influencers in Spain does not change. This suggests that increases in sponsored post shares in Table 1 reflect changes in the number of sponsored posts rather than the number of posts that influencers are posting or other strategies.[48]

Table 3 looks at influencer/month-level outcomes related to aggregate follower engagement. Columns (1) and (2) look at the average number of likes and comments that posts by influencer i in month t receive. Column (3) looks at the mean number of followers. The results suggest that the number of likes and comments falls after regulation. The decrease is both statistically significant and quantitatively large. Relative to a baseline pre-treatment mean of 770 likes per post, the average number of likes in Germany after regulations falls by over 480 (over 50%).[49] This is consistent with the intuition from Section 4 that regulations causing an increase in sponsorship will reduce attention, on average. Fainmesser and Galeotti (2021) and Mitchell (2021) generate theoretical predictions consistent with this result.

Decreasing engagement may indicate a decrease in average consumer welfare in Germany after the strengthening of disclosure regulations. However, as discussed in Section 4.4, there is a distinction between engagement and attention, and attention may encompass more than just likes or comments. Although there cannot be engagement with zero attention, there may be many followers who pay attention to post content but who do not like or comment. As such, through the lens of our model from Section 4, there is no direct connection between changes in engagement and consumer welfare.[50] We therefore take this evidence to be suggestive.

Column (3) shows that there is no statistically significant change in the average number of followers that an influencer has after the regulatory environment becomes stricter. However, the number of observations in this regression is small since most influencer/month observations in our sample do not have an observable number of followers for each month.[51]

The model in Section 4 generates predictions about the effects of regulations on follower beliefs and attention across sponsored and non-sponsored language. In particular, sponsored language could be trusted more conditional on a post being undisclosed. This in turn suggests that while aggregate engagement may fall due to the increase in advertising, the trust that followers have in undisclosed sponsored posts could increase *relative* to the non-sponsored posts. Our data do not have information about the beliefs of followers.[52] As well, as discussed above, engagement is a noisy measure of attention. Nonetheless, we test the predictions using engagement data. We first estimate influencer–month regressions, where the dependent variable is the monthly ratio of mean sponsored-undisclosed post likes over mean non-sponsored post likes.

Estimates of these regressions are in Table 4. As in previous tables, each column represents a different classifier used to define sponsored posts. Although the estimates are somewhat noisy, they broadly show that mean engagement for undisclosed-sponsored posts increases relative to non-sponsored posts after the strengthening of disclosure regulations in Germany. SGD classifier estimates in column (1) suggest that relative to a pre-treatment mean engagement ratio of 0.9, the ratio in the post-regulation period is approximately 0.96. This is also the case for the manual classifier in column (2): From a pre-treatment ratio of 0.728, the regulatory changes increase the ratio to approximately 0.9. It is consistent with the model's predictions of increasing follower trust in sponsored posts that "slip" through the disclosure filter.

The model also predicts that posts disclosed as advertising will have very "sponsored language" (high α in Section 4) and, as a result, will receive lower attention and engagement as compared to undisclosed-sponsored and non-sponsored posts. This is the case for the Kylie Jenner posts in Figure A1, and we observe such patterns in the general data. Mean likes for disclosed-sponsored posts during the regulated period in Germany are much lower than mean likes for non-sponsored posts, and are also lower than mean likes for sponsored undisclosed posts. The SGD-predicted mean like ratio for sponsored and disclosed posts over non-sponsored posts is 0.4 (compared to a ratio of nearly 1 for sponsored undisclosed posts). The manual predicted mean like ratio for sponsored and disclosed posts is 0.55 (compared to a ratio of approximately 1.1 for sponsored undisclosed posts). We show the average relative like ratios for sponsored-disclosed and sponsored-undisclosed posts relative to non-sponsored posts in Germany after disclosure regulations in Table 5.

We further confirm these influencer-level estimates by estimating a series of post-level difference-in-differences regressions

TABLE 3 | Influencer/month DiD estimates—engagement.

Outcome:	(1) Mean *N* likes	(2) Mean *N* comments	(3) Mean *N* followers
Germany × Treated Period	−483.217***	−22.663***	−4693
	(157.693)	(7.232)	(8275)
Pre-treatment mean	769.1	17.10	76,790
Country controls	YES	YES	YES
Influencer FE	YES	YES	YES
Year–month FE	YES	YES	YES
Account age FE	YES	YES	YES
Account age × First-account-year FE	YES	YES	YES
Observations	67,235	67,235	14,165
R-squared	0.637	0.251	0.906

Note: Sample includes influencer/month-level observations from January 2014 to December 2019 with at least two posts in a month. Influencers in the sample are CEM-matched as described in Section 3.4. "Germany × Treated Period" is a dummy equal to 1 for all German influencer observations after November 2016 and 0 otherwise. All regressions include influencer and time fixed effects, as well as account age fixed effects and account age × first-account year fixed effects. Country controls include quarterly GDP per capita, quarterly population, and monthly measures of Instagram popularity based on Google Trends results. Standard errors are clustered at the influencer level. ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.

TABLE 4 | Influencer/month DiD estimates—relative engagement (likes).

Outcome:	(1)	(2) Mean *N* Likes for sponsored-undisclosed posts / Mean *N* likes for non-sponsored posts	(3)
Classifier:	**SGD L1**	**Manual**	**SGD L1 + Manual**
Germany × Treated Period	0.056*	0.154**	0.012
	(0.028)	(0.072)	(0.016)
Pre-treatment mean	0.906	0.728	0.729
Country controls	YES	YES	YES
Influencer FE	YES	YES	YES
Year–month FE	YES	YES	YES
Account age FE	YES	YES	YES
Account age × First-account-year FE	YES	YES	YES
Observations	63,298	59,627	64,722
R-squared	0.054	0.058	0.114

Note: Sample includes influencer/month-level observations from January 2014 to December 2019 with at least two posts in a month. Influencers in the sample are CEM-matched as described in Section 3.4. The dependent variable in each regression is a ratio of the mean number of likes of posts that were labeled as sponsored and undisclosed for influencer *i* in month *t* over the mean number of likes of posts that were labeled as non-sponsored. A full list of words used to detect disclosure is in Appendix A3.1. Each column uses a different classifier to label posts as sponsored. "Germany × Treated Period" is a dummy equal to 1 for all German influencer observations after November 2016 and 0 otherwise. All regressions include influencer and time fixed effects, as well as account age fixed effects and account age × first-account year fixed effects. Country controls include quarterly GDP per capita, quarterly population, and monthly measures of Instagram popularity based on Google Trends results. Standard errors are clustered at the influencer level. ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.

in Appendix A8. We segment the sample by disclosure and sponsorship status and compare the number of likes that similar posts receive before and after regulations. In the post-level regressions, we condition on the popularity of the influencer, which abstracts from the aggregate engagement effects of regulations. Estimates in Table A7 in the Appendix show that conditional on influencer popularity, likes for non-sponsored posts weakly fall after disclosure regulations, while likes for sponsored undisclosed posts do not change. These results support the mechanism in the theoretical model, suggesting that followers trust sponsored content relatively more after regulation.

Because attention may encompass more than just likes and comments, there is no direct connection between the changes in engagement and consumer welfare. We take this evidence to therefore be suggestive. A more complete model that tried to

TABLE 5 | Relative post engagement in Germany in treated period.

Classifier:	(1) SGD L1	(2) Manual	(3) SGD L1 + Manual
Average $\frac{\text{Mean } N \text{ likes spon-undisc.}}{\text{Mean } N \text{ likes non-spon.}}$	0.96	1.11	0.81
Average $\frac{\text{Mean } N \text{ likes spon-disc.}}{\text{Mean } N \text{ likes non-spon.}}$	0.42	0.55	0.39

Note: Sample includes posts from Germany from November 2016 to December 2019 with at least two posts in a month. Influencers in the sample are CEM-matched as described in Section 3.4. Each cell shows a ratio of either (i) the mean monthly number of likes of posts that were labeled as sponsored and undisclosed over the mean monthly number of likes of posts that were labeled as non-sponsored, or (ii) the mean monthly number of likes of posts that were labeled as sponsored and disclosed over the mean monthly number of likes of posts that were labeled as non-sponsored. A full list of words used to detect disclosure is in Appendix A3.1. Each column uses a different classifier to label posts as sponsored.

measure welfare effects would need additional elements, such as follower heterogeneity.

6 | Discussion and Conclusion

We show that advertising disclosure regulations on social media have real effects. Influencers in Germany increase both the number of posts that are labeled as disclosed and disclosure rates of sponsored posts after disclosure regulations are substantially strengthened in late 2016. This is an important empirical finding in and of itself, given widespread popular skepticism about such regulations (TheGuardian.com). Consistent with previous theoretical work (Fainmesser and Galeotti 2021; Mitchell 2021), we also show that there are potentially adverse effects to such regulations. The number and percentage of sponsored posts increase at the influencer level, and the share of sponsored content among undisclosed posts increases. Overall, our findings suggest that in markets with no direct compensation mechanisms, regulations that distort indirect compensation mechanisms can have large and unanticipated effects on the supply of sponsored content.

We present a novel approach to detecting sponsored content, including undisclosed sponsored content, using natural language processing. We use *multilingual embeddings* and a supervised ML approach that labels posts with language similar to the language of disclosed-sponsored posts as sponsored. We train our ML model on data after the strengthening of regulations in Germany. This raises concerns about the ability of our classifier to accurately distinguish between sponsored and non-sponsored content in the presence of strategic obfuscation by influencers. As well, there may be limitations in our ability to apply the classifier to content from other time periods and countries (i.e., the "pre" period in Germany and Spain) given potential changes in language. We show that our main ML classifier (SGD) performs well and that its performance is stable over time. Nonetheless, although we are confident that we correctly identified the *direction* of changes in sponsorship, there may be some measurement error in our estimated magnitudes. Potential avenues for future research could include using NLP classifiers in conjunction with MTurk, or with large language models (LLMs), to better identify the commercial content of posts independently of changes in disclosure regulations.

Our findings are relevant for regulators of online markets and platforms by helping understand the responses of intermediaries to regulation. Online platforms and services such as Google Search, Spotify, and Amazon mix explicitly sponsored content, "authentic" content, and content that is not sponsored directly but that benefits the platform. Our findings on increasing sponsorship, including increased hidden sponsorship, suggest that forcing platforms to disclose one channel of advertising or paid content (such as paid reviews) may increase the total amount of advertising that consumers are exposed to. This is a key concern for policymakers and regulators.

There are questions about the welfare implications of these results. If we choose to interpret the number of likes per influencer as a revealed preference measure of consumer utility in this market, our findings suggest that consumer welfare falls after regulation. We also find changes in relative engagement for undisclosed-sponsored and organic posts that are consistent with possible falling consumer welfare. At the same time, it is not clear how to account for likes as a measure of welfare if consumers are deceived about post content in the pre-disclosure period. Influencers may also be endogenously changing the type or quality of sponsored content in response to regulations.

Evaluating the overall welfare effects of the policy would require incorporating additional assumptions into the model and a different empirical approach than the one we currently use. Both are outside the scope of the current article. Nonetheless, these are open avenues for future research. A more fully specified model, together with the data used in this article, could also allow for direct estimation of parameters governing influencer and follower behavior. With these parameters in hand, it should be possible to evaluate the effects of counterfactual regulation schemes of the kind proposed by Fainmesser and Galeotti (2021) and Mitchell (2021).

Acknowledgments

We would like to thank Editor Allan Collard-Wexler and three referees for very helpful comments, as well as seminar participants at CEMFI-UC3M, Tilburg, Toronto, TSE, Mannheim, UCLA-Anderson, the QVMS, the 2019 Israeli IO Day, and ACM-EC 2020. We would also like to thank CrowdTangle.com for giving us access to Instagram data, HypeAuditor.com for providing us with a list of German and Spanish influencers, and Camille Portes for excellent research assistance. A previous draft of this article circulating under a similar title used a different sample of influencers and its extended abstract was published in the conference proceedings of ACM-EC 2020. Daniel Ershov would like to acknowledge the support from ANR under grant ANR-17-EUR-0010 (Investissements d'Avenir program).

Endnotes

[1] Recent evidence highlights that online advice can have real effects (Alatas et al. 2024, Müller and Schwarz 2023).

[2] See the Stigler Center report on Digital Platforms (ChicagoBooth.edu), EU Commission Report on Competition Policy in the Digital Era (Europa.eu), and the UK Competition Authority report on "unlocking digital competition" (https://www.gov.uk/government/publications/unlocking-digital-competition-report-of-the-digital-competition-expert-panelGov.uk) for recent summaries of a broad range of policy concerns.

[3] Google search mixes "organic" search results and sponsored links. Google earned more than $130 billion USD from advertising in 2018 (AndroidAuthority.com). An equally important channel is links to its own properties such as YouTube, maps, news, or shopping from its search engine, which has led to regulatory action in several jurisdictions and a 2.4 billion Euro fine from the EU Commission (Europa.eu).

[4] See additional discussion of the literature in Section 1.1.

[5] We do not capture post images.

[6] See Appendix A3.2 for a full list of keywords.

[7] Other classifiers such as Naive Bayes or Random Forest produce similar results (see Appendix A9).

[8] Each post is represented by a 300-dimensional continuous vector (Arora, Liang, and Ma 2017). Posts similar to one another in meaning, even if they use different language/words, are close to each other in that space (Joulin et al. 2018).

[9] Results from the non-matched sample are similar and are available in Appendix A13. Results from a propensity score matching approach are in Appendix A14 and are also similar.

[10] Previous literature, such as Fainmesser and Galeotti (2021) and Mitchell (2021), propose alternative mechanisms that also generate more sponsorship following disclosure regulation.

[11] There is an emerging empirical literature in economics and marketing studying the behavior of influencers. For examples, see Hinnosaar and Hinnosaar (2024), Yang, Zhang, and Zhang (2021), and Hughes, Swaminathan, and Brooks (2019).

[12] There is also a legal literature on advertising disclosure regulations. This literature deals with the many practical issues of legally defining what influencers are, what is advertising, and the jurisdictions that different authorities have to enforce regulations. Recent works include Ducato (2020) and Goanta and Ranchordas (2020) among others.

[13] For instance, see ConsumerReports.com.

[14] It is not subject to the GDPR or other European laws. Consumers choose to follow influencers and the advertising does not involve the collection of personal data outside of agreements that influencers sign (sideqik.com).

[15] Exceptions to these rules could be influencers who primarily sell their own line of products or influencers who are "brand ambassadors" and who have longer-term contractual relations with brands.

[16] Influencer advertising in the United States is regulated by the FTC (see more details in this section) and in the United Kingdom by the Advertising Standards Authority (ASA) and the Competition and Markets Authority (CMA). However, the broad mandates of the FCC and OfCom include regulating media content—including advertising—for traditional media such as radio and television (e.g., https://www.fcc.gov/media/program-content-regulations). In that sense, the German regulator's role is broadly analogous to the FCC, except that they extended their definition of "content provider" beyond television and radio to social media.

[17] In Appendix A1, we describe the Italian and French regulatory environment.

[18] CrowdTangle tracks over 2 million public Instagram accounts, including all public Instagram accounts with more than 75k followers and all verified accounts (CrowdTangle.com). It does not include paid ads unless those ads began as organic, non-paid posts that were subsequently "boosted" using Facebook's advertising tools. It also does not include activity on private accounts, or posts made visible only to specific groups of followers (CrowdTangle.com).

[19] In the main estimation sample, we restrict our sample period to go from January 2014 to December 2019.

[20] Using proprietary methods, HypeAuditor calculates the share of each influencer's followers who live in their country.

[21] Some influencers may live abroad while posting about local content. This does not seem to be the case. Influencers from Spain primarily post from Spain (although they also post from other locations). This makes sense given that even the most popular influencers are equivalent to local celebrities. Many advertisers who want to sponsor content with local influencers are also likely to be local.

[22] Local content preferences online have been persistently demonstrated in previous literature, such as Blum and Goldfarb (2006) and Ferreira and Waldfogel (2013).

[23] See Appendix A16 for more detail.

[24] For examples of disclosed-sponsored, non-sponsored, and ambiguously sponsored posts, see Appendix A2.

[25] While it would be possible to claim a non-sponsored post was sponsored via false disclosure, industry discussion never focuses on this category, so we do not try to measure it.

[26] This may be due to the underlying ambiguous nature of disclosure rules. As discussed in Section 2.1, there was disagreement among German courts about the extent and strictness of disclosure requirements under the new regulations. In Appendix A18, we discuss the results of a MTurk survey for a small random sample of *undisclosed* posts from Germany in the post-regulatory change period. For each post, we asked survey respondents whether the post was likely sponsored (i.e., whether the user posting it received compensation for that post). Survey results show that a large share of undisclosed posts are likely sponsored.

[27] Posts that have no captions are automatically labeled as non-sponsored under both approaches. We do not see substantial differences over time between Spain and Germany in the percentage of text-free posts, and omitting them from the analysis does not change our main results.

[28] We chose to use 2018 because our data suggest that disclosure rates in Germany stabilize around late 2017 (see Figure 1).

[29] See Appendix A3.2.1 for more details on our ML approach and classifier tuning.

[30] Since the embeddings are common across Germany and Spain, this also helps us with translating between German and Spanish.

[31] See Goanta and Wildhaber (2019) for more details on various contractual arrangements between influencers and brands, including examples of brands directing post text.

[32] Moreover, our theoretical model in Section 4, which captures influencer obfuscation, predicts that some mis-measurement of sponsored content will not affect the estimated *direction* of the effects of regulations on sponsorship. See additional discussion in Section 4.

[33] See Figures A2 and A3 and related discussion for more details. Results with other classifiers, such as Naive Bayes, Decision Tree, or Random Forest, are in Appendix A9 and are qualitatively similar to results in the main text. Additional comparisons between the classifiers are in Appendix A3.2.2.

[34] Disclosed-sponsored post embeddings do move over time in Germany, especially during the "post" period. This is likely because the "pre" disclosed-sponsored posts were a very small and unrepre-

sentative subset of all sponsored posts (see Appendix A12 for more discussion).

[35] Our main estimates are robust to including influencer observations with only one post per month.

[36] Additional details are in Appendix A6. As a robustness check, we use an alternative matching approach based on propensity score weighting. Results in Appendix A14 show this does not affect our main findings.

[37] More precisely, we are left with 618 influencers from Germany and 560 influencers from Spain.

[38] Additional summary statistics for the full non-matched sample are available in Table A12 in the Appendix. Estimates using the non-matched sample are qualitatively similar to the matched results (see Appendix A13).

[39] Since we map words to embeddings (numbers) in our empirical exercise, one can think of this as a vector of numbers.

[40] Similar to Fainmesser and Galeotti (2021) and for the sake of tractability, we model the choice of post type at a given independent occasion and abstract from multi-post sponsorship campaigns and other dynamic considerations.

[41] A sufficient condition is an Inada condition on c.

[42] In Appendix A8, we also estimate effects for *post*-level outcomes.

[43] We choose November 2016 since the "clarification" to German disclosure regulations came out in October 2016 (Section 2.1). This is likely understating the true effects of the changes, as the enforcement of regulations started in the middle of 2017. See Figures 1 and 2 for period-specific effect estimates.

[44] In Appendix A16, we also show that, as proxied by Google Trends search volumes, overall demand for Instagram content in Germany did not change relative to Spain after German regulations were introduced.

[45] We show similar effects on content without relying on any classification in Appendix A11. We find that the distribution of embeddings in Germany is changing between the pre- and post-regulatory change period more than the distribution of embeddings in Spain. In Figure A6, we calculate the average differences in cosine distance from 0 for each post in each country and find that German embeddings' average distance from 0 increases relative to Spanish embeddings' average distance from 0 after regulations come in. This is the case for both disclosed and undisclosed posts. We also show similar effects in an influencer–month-level regression with additional controls in Table A11.

[46] Results using alternative classifiers such as Naive Bayes, Decision Tree, or Random Forest are in Appendix A9. Results using the full sample of influencers, including those that are not matched by the CEM algorithm, are in Appendix A13. Key coefficient estimates are qualitatively similar throughout.

[47] Estimates in column (2), for the manual classifier, show a small and positive but not statistically significant coefficient. One possibility for why we find no effects is that our manual classifier is very "loose," including many words that only vaguely connote commercial intent. As such, we are biased toward finding a null effect. The same regression at the post-level in column (2) of the bottom panel of Table A6 shows a smaller discrepancy between SGD and manual point estimates.

[48] This is also consistent with evidence from Appendix A11 that uses embedding data to show that post text itself is changing in Germany relative to Spain after regulations.

[49] The average decrease in the mean number of comments (in Column 2) is bigger than the pre-treatment mean because of the skewed distribution of the variable and its growth over time.

[50] A more complete model that tried to directly connect attention and engagement and measure welfare effects would need additional elements, such as follower heterogeneity.

[51] This is an issue with the underlying data collection by CrowdTangle.com, which does not always collect the number of followers for each post. While it is possible to interpolate or extrapolate the number of followers for the missing observations, this would require strong assumptions. We find negative and statistically significant effects using the larger non-matched sample in Table A15.

[52] For example, CrowdTangle.com does not collect any information about the comments that posts receive (i.e., comment text) except for the aggregate number of comments.

References

Alatas, V., A. G. Chandrasekhar, M. Mobius, B. A. Olken, and C. Paladines. 2024. "Do Celebrity Endorsements Matter? A Twitter Experiment Promoting Vaccination in Indonesia." The Economic Journal 134, no. 659: 913-933.. When Celebrities Speak: A Nationwide Twitter Experiment Promoting Vaccination in Indonesia. Technical Report. National Bureau of Economic Research.

Anagol, S., S. Cole, and S. Sarkar. 2017. "Understanding the Advice of Commissions-Motivated Agents: Evidence from the Indian Life Insurance Market." *Review of Economics and Statistics* 99, no. 1: 1–15.

Arora, S., Y. Liang, and T. Ma. 2017. "A Simple but Tough-to-Beat Baseline for Sentence Embeddings." Paper presented at the 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26.

Ash, E., D. L. Chen, and S. Naidu. 2024. "Ideas Have Consequences: The Impact of Law and Economics on American Justice." Center for Law & Economics Working Paper Series 4.

Bhattacharya, V., G. Illanes, and M. Padi. 2019. Fiduciary duty and the market for financial advice. Technical Report. National Bureau of Economic Research.

Blum, B. S., and A. Goldfarb. 2006. "Does the Internet Defy the Law of Gravity?." *Journal of International Economics* 70, no. 2: 384–405.

Ducato, R. 2020. "One Hashtag to Rule Them All? Mandated Disclosures and Design Duties in Influencer Marketing Practices." In *The Regulation of Social Media Influencers*, edited by C. Goanta and S. Ranchordás. Edward Elgar Publishing.

Fainmesser, I. P., and A. Galeotti. 2021. "The Market for Online Influence." *American Economic Journal: Microeconomics* 13, no. 4: 332–372.

Ferreira, F., and J. Waldfogel. 2013. "Pop Internationalism: Has Half A Century of World Music Trade Displaced Local Culture?." *The Economic Journal* 123, no. 569: 634–664.

Gentzkow, M., J. M. Shapiro, and M. Taddy. 2019. "Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech." *Econometrica* 87, no. 4: 1307–1340.

Goanta, C., and I. Wildhaber. 2019. "In the Business of Influence: Contractual Practices and Social Media Content Monetisation." *Schweizerische Zeitschrift für Wirtschafts-und Finanzmarktrecht, SZW* 4.

Goanta, C., and S. Ranchordás. 2020. "The Regulation of Social Media Influencers: An Introduction." In *The Regulation of Social Media Influencers*, edited by C. Goanta and S. Ranchordás. Edward Elgar Publishing.

Godes, D., and D. Mayzlin. 2004. "Using Online Conversations To Study Word-of-Mouth Communication." *Marketing Science* 23, no. 4: 545–560.

Godes, D., and D. Mayzlin. 2009. "Firm-Created Word-of-Mouth Communication: Evidence from A Field Test." *Marketing Science* 28, no. 4: 721–739.

Hansen, S., M. McMahon, and A. Prat. 2018. "Transparency and Deliberation within the FOMC: A Computational Linguistics Approach." *The Quarterly Journal of Economics* 133, no. 2: 801–870.

He, S., B. Hollenbeck, and D. Proserpio. 2022. "The Market for Fake Reviews." *Marketing Science* 41, no. 5: 896–921.

Hinnosaar, M., and T. Hinnosaar. 2024. "Influencer Cartels." SSRN 3786617.

Horstmann, I., and G. MacDonald. 2003. "Is Advertising A Signal of Product Quality? Evidence from the Compact Disc Player Market, 1983–1992." *International Journal of Industrial Organization* 21, no. 3: 317–345.

Hughes, C., V. Swaminathan, and G. Brooks. 2019. "Driving Brand Engagement Through Online Social Influencers: An Empirical Investigation of Sponsored Blogging Campaigns." *Journal of Marketing* 83, no. 5: 78–96.

Inderst, R., and M. Ottaviani. 2012. "Competition through Commissions and Kickbacks." *American Economic Review* 102, no. 2: 780–809.

Joulin, A., P. Bojanowski, T. Mikolov, H. Jégou, and E. Grave. 2018. "Loss in Translation: Learning Bilingual Word Mapping with A Retrieval Criterion." *arXiv preprint arXiv:1804.07745* .

Mayzlin, D. 2006. "Promotional Chat on the Internet." *Marketing science* 25, no. 2: 155–163.

Mitchell, M. 2021. "Free Ad(vice): Internet Influencers and Disclosure Regulation." *RAND Journal of Economics* 52, no. 1: 3–21.

Müller, K., and C. Schwarz. 2023. "From Hashtag To Hate Crime: Twitter and Anti-Minority Sentiment." *American Economic Journal: Applied Economics* 15, no. 3: 270-312. 314910

Pei, A., and D. Mayzlin. 2021. "Influencing Social Media Influencers Through Affiliation." Marketing Science 41, no. 3: 593-615.

Sahni, N. S., and H. S. Nair. 2020. "Does Advertising Serve as a Signal? Evidence From a Field Experiment in Mobile Search." *The Review of Economic Studies* 87, no. 3: 1529–1564.

Yang, J., J. Zhang, and Y. Zhang. 2021. "First Law of Motion: Influencer Video Advertising on Tiktok." *SSRN 3815124* .

Supporting Information

Additional supporting information can be found online in the Supporting Information section.