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

Daniel Ershov[*]
Toulouse School of Economics
Université Toulouse 1 Capitole
daniel.ershov@tse-fr.eu

Matthew Mitchell
Graduate Department of Management
University of Toronto
matthew.mitchell@utoronto.ca

November 4, 2021

Abstract

We study the effects of advertising disclosure regulations in social media markets. Theory generates ambiguous predictions about the effects of regulations on the equilibrium content, user engagement and welfare. 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 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 a substantial adoption of disclosure after regulations, but also an increase in sponsored content including undisclosed sponsored content. We also find reductions in engagement, suggesting that followers were likely negatively affected.

[*] We would like to thank seminar participants at CEMFI-UC3M, Tilburg, Toronto, TSE, Mannheim, UCLA-Anderson, the QVMS, the 2019 Israeli IO Day and ACM-EC 2020 for their helpful comments on this project. 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 paper 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 support from ANR under grant ANR-17-EUR-0010 (Investissements d'Avenir program).

1 Introduction

This paper studies social media influencers in order 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 UK 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.[4]

Our research question is: what effects do advertising disclosure regulations have on content and engagement in user generated social media platforms? We use 2014-2020 post-level data from a large sample of Instagram influencers from Germany and Spain to answer this question. Our empirical strategy takes advantage of the strengthening of disclosure regulations in Germany starting at the end of 2016. We use a difference-in-differences approach, comparing the behaviour of similar Spanish and German influencers before and after regulations. Natural Language Processing (NLP) on post texts allows us to recover information about content: whether a given post is likely sponsored or not and whether it is disclosed as sponsored. We identify a substantial amount of sponsored content in Germany and in Spain, including *undisclosed-sponsored* content. We show that influencers in Germany post more sponsored content after regulations are introduced. Despite disclosure regulations, they also increase the amount of *undisclosed-sponsored* content. We also find reductions in consumer (follower) engagement, as measured by post likes and comments.

[1] Recent evidence highlights that online advice can have real effects (Alatas et al. 2019, Müller and Schwarz 2019).

[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" (Gov.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] For example, Google earns nearly 10% of its revenues from YouTube (TheVerge.com). A strict interpretation of disclosure regulations would require it to disclose Google search links directing users to YouTube as sponsored/compensated.

Economic theory is important to interpret our empirical findings. There are two competing views on the effects of disclosure regulations on content in existing literature. Drawing on theories of buyer-seller transactions, regulatory agencies (i.e., FCC in the US, or ASA in the UK) 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 papers such as Inderst and Ottaviani (2012), Fainmesser and Galeotti (2020), Pei and Mayzlin (2019) and Mitchell (2021) suggest it might be incomplete. These papers 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.

We introduce a theoretical model to capture these competing views in a way that reflects our empirical setting, especially highlighting the roles of organic, disclosed-sponsored, and undisclosed-sponsored posts. In the model, an influencer chooses where to post an organic or sponsored post, which determines the words used in the post.[5] The influencer earns higher payoffs from sponsored posts, but their revenues also depend on follower engagement and followers prefer to engage with organic posts. Absent regulations followers do not know a post's type and have to form beliefs about sponsorship probabilities. Engagement increases with followers' expected belief that a post is organic. Disclosure regulations label a portion of sponsored posts as sponsored, changing follower beliefs. Engagement falls for disclosed-sponsored posts, but could increase for undisclosed-sponsored posts that slip through the disclosure filter relative to organic posts. Depending on which engagement effect dominates, influencers respond to regulations by either increasing or decreasing the amount of sponsored content. Follower welfare may fall if the amount of sponsored content increases.[6]

The model results in several predictions to test in the data: does the amount of sponsored content, and especially undisclosed sponsored content increase after regulations? What happens to aggregate user engagement? And what happens to the engagement of undisclosed-sponsored relative to organic posts? To empirically evaluate the effects of disclosure regulations on content and engagement, we collect Instagram data on a random sample of twelve thousand *local* Instagram influencers in Germany and Spain from CrowdTangle.com. The German regulatory environment became substantially stricter towards the end of 2016: In October 2016, German state media authorities (analogous to the US FCC/ UK OfCom) 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.[7] We face two main empirical challenges using this data. First is a measurement challenge: while

[5] For example, sponsored posts may use the word "Sale" more frequently than organic posts. This is consistent with many existing contractual arrangements where influencers do not have direct control over the content of sponsored posts (Goanta and Wildhaber 2019).

[6] Previous literature proposed alternative mechanisms that also generate more sponsorship following disclosure regulation. In Fainmesser and Galeotti (2020) disclosure leaves influencers with followers who are less elastic to advertising. In Mitchell (2021) any cost of lower revenue for influencers might in turn hurt followers, since influencers are disciplined by the possibility of future ad revenues. Reduced per-post revenues may incentivize them to increase the number of sponsored posts. Neither paper, though, speaks directly to undisclosed-sponsored posts that we measure.

[7] We do not capture post images.

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 in Germany prior to the regulatory change. Also, our model suggests that changes in the amount of undisclosed-sponsored content is key to understanding the welfare effects of regulatory changes.

We address the first challenge 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).[8] (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 several popular ML classifiers: Gaussian Naive Bayes, Stochastic Gradient Descent, Decision Tree, and Random Forest. 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.[9] We also use a combination of the two approaches, labelling 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.

Our second challenge is an identification challenge: isolating the causal effect of regulations on sponsorship and engagement. We address the second challenge using 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.[10] 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.

Our difference-in-differences estimates show that disclosure regulations affect the type of content influencers post online. Results from all classification methods 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 Mean of 15-30 percentage points. At a minimum, sponsored shares increase by approximately 3 percentage points (10%). At a maximum, sponsored shares increase by 7 percentage points (over 50%). 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 and the sponsorship rate among

[8] See Appendix A.3 for a full list of keywords.

[9] Each post is represented by a 300-dimensional continuous vector (Arora et al. 2017). Posts that are 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). See Ash et al. (2019) for another recent application of embeddings in economics.

[10] Results from the non-matched sample are similar and are available in Appendix A.10.

undisclosed posts increases. We also show changes in engagement in response to the regulatory change. Both the mean number of likes and the mean number of comments that influencers in Germany receive falls after the regulatory change. This is consistent with followers in social media markets disliking sponsored content. Consistent with the model's predictions, we also show that the average number of likes per post increases for undisclosed-sponsored posts relative to non-sponsored posts.[11] Timing tests show that effects are not driven by pre-trends and the parallel trends assumption holds.

The contributions of this paper are two-fold. This is the first empirical paper 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.

Our findings are relevant for the broader question of regulation of online markets and platforms such as Google Search. Google Search also 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 are effectively ads. Google has been accused of biasing search results in favour of its own products and recently received a multi-billion Euro fine from the EU Commission. Google's acquisitions of other firms such as YouTube may also be related to its trade-offs between directing consumers to authentic vs sponsored content. Other popular online platforms such as Spotify face similar trade-offs (NYTimes.com). Our results help understand platform incentives in online markets. Our findings on increasing sponsorship, including increased undisclosed-sponsorship suggest that forcing platforms to disclose advertising may in fact increase the amount of advertising that consumers are exposed to. This is a key concern for policy-makers and regulators.

Our paper also contributes methodologically to the empirical economics literature that uses text as data (e.g. Gentzkow et al. 2019, Hansen et al. 2018, Ash et al. 2019). Our combined use of multi-lingual embeddings and supervised machine learning classification is novel to this literature. Such methods allow for straight-forward 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 paper proceeds as follows. Section 2 gives an overview of related literature. Section 3 describes industry background and the regulatory change we study. Section 4 presents the raw Instagram data we use in the paper and discusses the empirical methodology: the classification of sponsored content and the difference-in-differences estimation. Section 5 presents our theoretical model. Section 6 describes the classification outcomes and presents some summary statistics and descriptive evidence. Section 7 shows the main difference-in-differences regression estimates and discusses various robustness checks. Section 8 concludes.

[11] In Appendix A.6 we also show that the average number of likes per post falls in Germany after the regulatory change for non-sponsored posts. This does not happen for sponsored and undisclosed posts.

2 Related Literature

To our knowledge, there is no existing empirical research on the effects of advertising regulations on content production online/ on social media. There are several existing theoretical studies.[12] Fainmesser and Galeotti (2020) 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 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 over-supply 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 vs 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 behaviour. In equilibrium, the influencer rotates between periods of building up reputation by providing authentic content, and periods of cashing in via sponsored content. Key for the influencer's strategy is not to provide too 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 (2019) also study recommendations by influencers. In their paper, the influencer faces an explicit informational model in persuading a potential consumer. This generates an endogenous limit on the degree of endorsement that the influencer can give before recommendations are no longer followed. In that model, some form of credible commitment to what is and is not endorsed (like an FTC rule) is necessary for the market to function.

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

[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 (2019) and Goanta and Ranchordas (2019) among others.

sometimes have.

Prior empirical literature has studied the impact of disclosure regulations on paid intermediaries - for example, in the market for insurance advice (Bhattacharya et al. 2019), and for financial advice (Anagol et al. 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 by 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 online advertisement; a key difference in our setting is that the post comes from an intermediary. Information providing intermediaries have long worked to avoid 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 signalling, 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 signalling for intermediaries that provide information and advertisements side by side.

3 Background

3.1 Background: Advertising on Instagram

We consider the market for influence on Instagram, a social media platform with over 1 billion active users in 2019.[14] Instagram users post visuals accompanied by captions. Users can also "follow" each other and like or comment on other users' posts.

Instagram content is provided for free, but it is also a popular market for sponsored posts. Most sponsorship happens through independent online marketing agencies that connect advertisers and Instagram influencers.[15] Influencers and brands sign up with third party marketing agencies (i.e., Heepspy.com, HypeAuditor.com). Influencers give marketing agencies access to their analytics and are subdivided into types such as nano-influencers (less than 20k followers), micro-influencers (less than 100k followers) and so on (mention.com). Influencers are also divided based on their interests.

[13] For instance, see ConsumerReports.com.

[14] Instagram has been owned by Facebook since 2012.

[15] Instagram itself also connects advertisers and users for sponsorship purposes. However, the mechanism there is different. Rather than advertisements appearing in the feed of the influencer users follow, ads run by Instagram appear in user feeds regardless of whether they follow the influencer or not. They are also always clearly delineated as "Promoted." We abstract from this advertising channel.

Advertisers can sponsor posts in several ways. Advertisers can send influencers free products (or services) for a post. For example, influencers may receive a free trip to a city for a series of posts about that city. Advertisers can also commission posts from the influencers and pay them for each post. Payments per post broadly depend on the number of followers of the influencer and the "quality" of their audience (the engagement). These range from a 10 dollars per post for micro-influencers to over $1 million USD per post for Kylie Jenner (webfx.com). Brands who want to run a campaign can then choose how much money to allocate and what type of influencers they want to use. More rarely, advertisers can enter into long term agreements with the influencers that involve traditional advertising (i.e., billboards) as well as social media posts by the influencers. In all three cases, advertisers decide on some parameters for the sponsored posts - hash-tags or links, the text of the post and sometimes the image.[16]

3.2 Background: Advertising Disclosure Regulations in Germany

Social media advertising regulations are not standardized across EU countries.[17] 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 2019). 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.[18] 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: labelling any posts where the influencer has been remunerated by a brand, including free products, as an ad with a visible hashtag.

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

[17] 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).

[18] 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.

The role of Die Medienanstalten is comparable to the FCC in the US or OfCom in the UK. 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 which 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 labelled as advertising (Ducato 2019). 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 as a commercial post and labelled 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 similar guidelines in 2016 by the FTC in the US 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 but not to other popular influencers. "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.[19] 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.

4 Data

4.1 Data Description

We collect data from CrowdTangle, which describes itself as "a public insights tool owned and operated by Facebook" (CrowdTangle.com).[20] Our raw data is at the post level. We observe a full

[19] In Appendix A.1, we describe the Italian and French regulatory environment.

[20] 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

history of posts for each influencer. 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 6,000 German and 6,000 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.[21] 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.[22] This is important to make sure that influencers are only affected by laws of the country in question, rather than laws in other countries.[23] Spanish followers' conception of the world is also not being changed by regulation in Germany since most Spanish followers are not reading German posts.[24]

The number of likes and comments for each post are recorded at the time of data collection rather than at the time of the posting. This may introduce measurement error between earlier and later posts, but industry experts estimate that engagement on posts on Instagram peters out after less than 24 hours (SprocketWebsites.com). There are several reasons for this: Instagram user profiles are relatively difficult to scroll through, many users follow a large number of influencers and Instagram targets users with recent content in their feed. To the best of our understanding, the "clarifications" to German disclosure regulations in late 2016 (described in Section 3.2) do not apply retroactively - 2017 cases against German influencers were based on 2017 content they posted.

We collect additional country-year/month specific data. Germany and Spain are different in many respects that 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.[25]

4.2 Post Examples

Figure 1 shows examples of influencer posts from Kylie Jenner's Instagram account ("@kyliejenner") and highlight potential challenges in classifying posts as sponsored or non-sponsored and as disclosed or undisclosed.[26] All posts were made within a few days of one another in early December 2018. Two of the posts are easy to classify. The post in panel (a) is clearly sponsored and disclosed. It begins with disclosure (#ad) and offers a discount code for purchasing a product. The post in panel

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).

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

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

[23] 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.

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

[25] See Appendix A.11 for more detail.

[26] Kylie Jenner is one of the most prominent social media influencers with over 200 million Instagram followers as of January 2021 (SocialBlade.com).

(b) is clearly non-sponsored. There is no disclosure text, no links or any references to advertisers or products.

The posts in panel (c) and (d) are more ambiguous. The post in panel (c) is also sponsored. It provides a link for followers of Kylie Jenner to shop for Adidas shoes. Kylie Jenner had a contract as a model for Adidas at the time (Forbes.com). This contractual arrangement is not disclosed via a #ad but through another hashtag (#adidas_Ambassador). This form of disclosure is broadly consistent with German regulations, although some court decisions instructed influencers to include disclosures at the beginning and not at the end of post captions. We consider such a post as disclosed and sponsored.[27] We consider the post in panel (d) as an example of a possible undisclosed sponsored post. It is a post that thanks an interior decorator for decorating Kylie Jenner's house for Christmas. However, the post includes a link to the decorator's professional website. Through this website the decorator can be hired for additional jobs. In that sense, it is not different than posts in panels (a) and (c). It is possible to imagine a sponsored contractual arrangement where Kylie Jenner receives discounted or free interior decorating services in return for a post with a link.[28]

An interesting comparison is of the number of likes for each post in Figure 1. The clearly non-sponsored post received over 5 million likes. The clearly sponsored post selling a product received only 1.8 million likes. The post selling Adidas shoes that is more ambiguous (but still denoted as sponsored) received 2.6 million likes. The post in panel (d) which is an even more ambiguous example of a potentially sponsored post received 2.9 million likes. This suggests a hierarchy in terms of consumer preferences that we incorporated into our model in Section 5.

4.3 Detecting Disclosure

Our raw data does not provide us with a specific identifier for posts that are disclosed as sponsored. 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 A.2.

4.4 Detecting Sponsorship

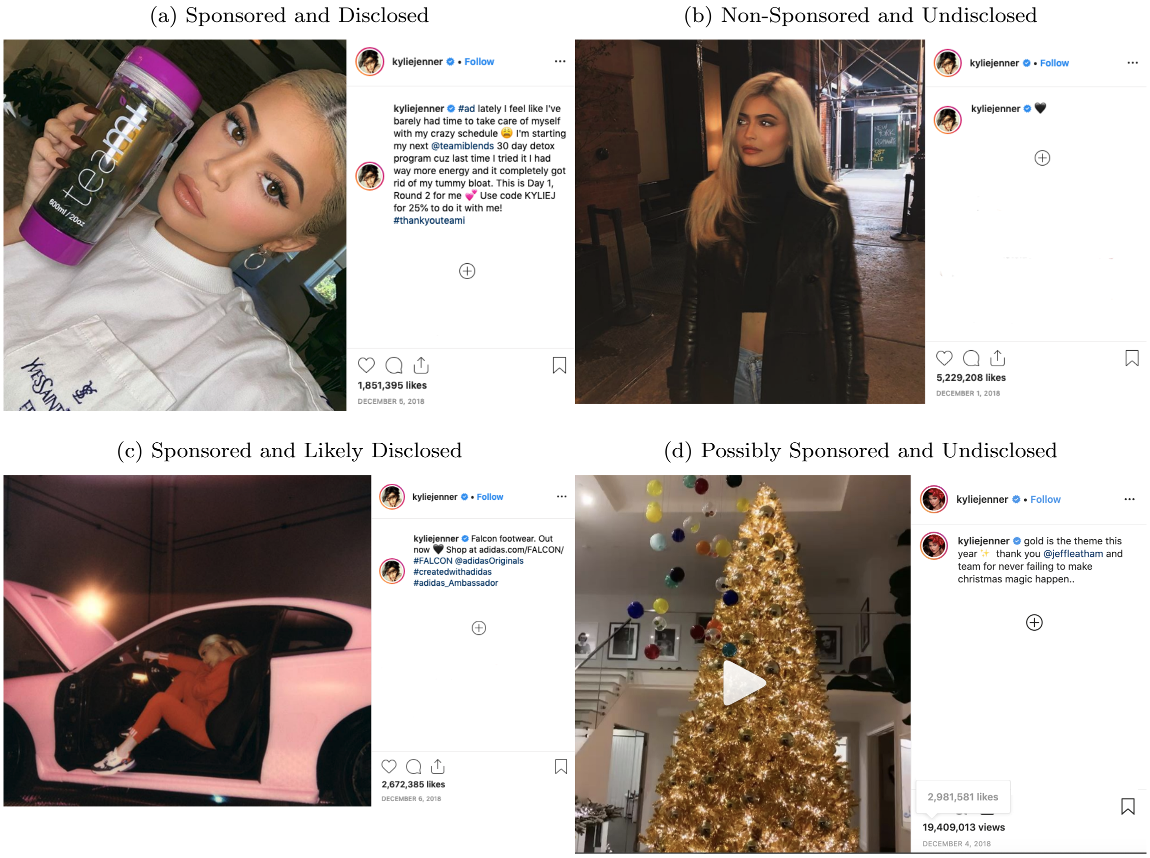
Our raw data does not provide us with a specific identifier for posts that are sponsored. While we can uncover disclosed-sponsored posts as per Section 4.3, undisclosed-sponsored posts are by definition hidden. Such posts are likely popular in Spain where there are no regulatory incentives for disclosure. There is also reason to suspect that some posts in Germany after the regulatory change are sponsored but undisclosed.[29]

[27] In our subsequent analysis, "ambassador" (in German or in Spanish) reflects disclosure, as well as any references to "collaboration," "partnership," "gift[ed]," etc. A full list of words defining disclosure is in Appendix A.2.

[28] Influencers often have similar arrangements where they receive services or items for free in return for posting about them (TaylorWessing.com).

[29] This may be due to the underlying ambiguous nature of disclosure rules. As discussed in Section 3.2, there was disagreement among German courts about the extent and strictness of disclosure requirements under the new regulations. In Appendix A.12 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.

Figure 1: Examples of Non-Sponsored, Sponsored and Disclosed Posts



(a) Sponsored and Disclosed

kyliejenner • Follow

kyliejenner #ad lately I feel like I've barely had time to take care of myself with my crazy schedule 😩 I'm starting my next @teamiblends 30 day detox program cuz last time I tried it I had way more energy and it completely got rid of my tummy bloat. This is Day 1, Round 2 for me 💕 Use code KYLIEJ for 25% to do it with me! #thankyouteami

1,851,395 likes

DECEMBER 5, 2018

(b) Non-Sponsored and Undisclosed

kyliejenner • Follow

kyliejenner 🖤

5,229,208 likes

DECEMBER 1, 2018

(c) Sponsored and Likely Disclosed

kyliejenner • Follow

kyliejenner Falcon footwear. Out now 🖤 Shop at adidas.com/FALCON/ #FALCON @adidasOriginals #createdwithadidas #adidas_Ambassador

2,672,385 likes

DECEMBER 6, 2018

(d) Possibly Sponsored and Undisclosed

kyliejenner • Follow

kyliejenner gold is the theme this year ✨ thank you @jeffleatham and team for never failing to make christmas magic happen..

2,981,581 likes

19,409,013 views

DECEMBER 4, 2018

The sample posts in Figure 1 showcase the difficulty in separating undisclosed-sponsored posts and non-sponsored posts.[30] We propose two approaches for identifying sponsored posts using text data: (1) A "manual" approach using a list of pre-determined keywords which generally denote sponsorship. (2) A supervised "automatic" approach using machine-learning classifiers.

Both methods rely on a conceptual framework that is consistent with the model in Section 5. Broadly speaking, there are two possible message types for influencers to send to their followers: an organic message, or a sponsored message. Based on the type of message they wish to send, they draw words from a message-type specific vocabulary.[31] 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. ML methods are generally probabilistic - a word can denote sponsored content with some probability but also authentic content with some probability. Additional words that are more "sponsored" push the post to be labelled as sponsored. The algorithm uncovers the probabilistic distribution of the words across message types.[32] We use posts from Germany after disclosure regulations as the training data. *Disclosed-sponsored* posts allow us to learn about the distributions of words across message types.

In the main analysis, we also use a combination of the two approaches, labelling a post as sponsored only if both a ML classifier and the manual approach classifies it as sponsored.

4.4.1 Manual Classification

In the first approach, we define a set of words that connotes sponsorship or commercial intent. 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"). We also consider words discussing events, launching products, references to shipping, references to dates (i.e., "out tonight"), references to new things, and influencers thanking someone (i.e., "thank you to L'Oreal"). Last, we include a large list of nearly 1,000 brands (fashion, retail and electronics), including German and Spanish specific brands (e.g., El Corte Ingles).[33] A full description of the keywords is in Appendix A.3.

The main benefit of this approach is that it is transparent and fast to implement by searching the text of each post for a pre-determined list of words. The main drawback of the approach is that it requires us to choose specific keywords. Selecting incorrect keywords or missing correct keywords introduces measurement bias. Some keywords are also ambiguous and are commonly used in both sponsored and non-sponsored contexts. Our keyword selection is purposefully *loose*. We include a number of words that can have ambiguous meaning, such as variants of "thank you."[34] Given this, our manual classifier may over-count the number of sponsored posts. We believe this is better

[30] 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.

[31] This reflects how sponsorship actually works - influencers are often contractually obligated to use certain vocabulary in their sponsored posts (Goanta and Wildhaber 2019).

[32] One of the algorithms, Stochastic Gradient Descent (SGD) is not probabilistic. Instead, it builds a series of hyper-planes to separate a "sponsored" space of words from a "non-sponsored" space of words and classifies posts belonging to the "sponsored" space as sponsored.

[33] We obtain the list of brands from Wikipedia.

[34] It is possible to be *even looser* in the choice of words. For example, we can assume that all "" tags denote sponsorship. This is not necessarily the case - tags are popular on Instagram as methods of communication across accounts or acknowledgement. However, they are often used to link to sponsoring pages.

than under-counting. Incorrectly selected words that do not belong to sponsored posts should not differentially change after regulation in Spain and Germany. Any changes that we find with this approach would under-state the true effects of regulation. There may be a concern that selected keywords are evolving differently in Germany and Spain prior to regulation. An analysis of the pre-trends suggests that this is not the case.[35]

4.4.2 Supervised Machine Learning (ML) Classification

In the second approach, we train supervised machine learning (ML) classification algorithms on labelled data and project the trained algorithms on non-labelled data. Posts from Germany in the period after disclosure regulations serve as training data. Since we know that disclosed posts are sponsored, we use ML algorithms to look for words/language associated with disclosure. In practice, since our raw data includes over 5 million German posts we randomly select a set of approximately 300,000 German posts from 2018 as our training dataset.[36]

A potential concern of applying a supervised classification approach with our training data is that it may not uncover words associated with sponsorship if disclosure is incomplete. This becomes especially problematic if different words are associated with disclosed and undisclosed posts. This could be the case if influencers are choosing to strategically disclose some types of sponsored posts but not others, or if influencers are unsure whether some types of ambiguous posts require disclosure.[37]

An additional novel challenge we face in applying text-based ML classifiers to our data are the languages used by influencers. Standard text-based ML classifiers such as Naive-Bayes use a "bag-of-words" approach, computing the probability of $P(word_i|disclosed)$ for each German word and recovering $P(disclosed|word_i)$. German and Spanish influencers in our data use different languages. Not only this, but influencers from both countries mix their local language with English words. Simple word-by-word translation is inaccurate, and with a very large dataset of 15 million posts and billions of characters, it is infeasible to perform more complicated translations using the context of the entire post.[38] Even if we could translate the posts, there would still be a concern that the translation would not pick up the "correct" words.[39] This is not a problem faced by previous "text as data" papers that use one language (Gentzkow et al. 2019, Hansen et al. 2018, Ash et al. 2019).

To deal with these challenges we transform all posts from "word-space" into "embedding/meaning-space." Word embeddings are a popular approach in natural language processing (NLP) and "rep-

[35] Robustness checks of our main results with a smaller set of manual keywords gives qualitatively and quantitatively similar results.

[36] We choose to use 2018 because our data suggests that disclosure rates in Germany stabilize around late 2017 (see Fig. 2).

[37] Our theory model in Section 5 also predicts that disclosed sponsored posts could look different than undisclosed sponsored posts in equilibrium. In Section 5 we include such an example where the dictionary consists of the words "Deal," "Sale" and "Love," and where each post consists of two randomly drawn words (e.g., "Deal Love"). In this example, "Deal" and "Sale" are associated with sponsored posts (are more likely to appear in sponsored posts) and "Love" is associated with non-sponsored posts. If the regulation is word-specific and "Sale" is regulated, every post with "Sale" is disclosed as sponsored but "Deal Deal" posts are sponsored and never disclosed.

[38] In a prior version of this paper with smaller data based on 50 influencers, we translated all Spanish posts into German using the full text of each post, and used a "bag-of-words" approach for classification. Results are consistent.

[39] For example: suppose all influencers in Germany use the word "amazing" to describe sponsored products. The translation from Spanish could translate words that are the same as "amazing" into another German word with a similar meaning such as "awesome." This would create a measurement error in projecting the trained German algorithm on the translated Spanish posts.

resent words as continuous vectors in a low dimensional space which capture lexical and semantic properties of words" (Arora et al. 2017). Put simply, each word is converted into a 300-dimensional continuous vector, and words with similar meanings are close to each other in this space. Computer science literature also shows that arithmetic done on word embeddings retains meaning.[40] Recently, embeddings have been used for translation. Joulin et al. (2018) present a method of "re-aligning" embeddings such that embeddings from two different languages are comparable.[41] They provide a set of pre-trained embeddings for research use (FastText.cc) and we apply those to our data. Our concerns about influencers mixing English and German or English and Spanish still hold. Transforming words into embedding space involves looking up specific words in a dictionary and recovering a corresponding numerical vector. The dictionaries are language specific, so German embedding mappings do not recognize English words and either return an empty vector or a nonsense vector. To deal with this issue we use an automated language detection algorithm (Joulin et al. 2016b,Joulin et al. 2016a) before converting the words into embedding space.[42]

Moving into embedding space gives us several advantages over "bag of words" approaches. If words used in undisclosed-sponsored posts are close in general language usage to words used in disclosed-sponsored posts, then our ML algorithms successfully label these posts as more likely to be sponsored. It should also be generally more difficult for influencers to be strategic about the *meaning* of their posts rather than specific words they use. For example, influencers may want to use somewhat more ambiguous words to encourage their followers to buy a product they were paid to promote, but the commercial meaning must still be there for the post to be effective. Last, if influencers in Spain and Germany use slightly different words that have similar meanings these words would be close in embedding space. Since we have data that stretches over several years across two countries, specific popular words can change while the underlying meanings of posts remains similar. Embeddings capture these.

Our text processing proceeds as follows (an illustrative example is in Appendix A.4). For each post in our data, we take the text of the post and remove punctuation, emojis and common stop-words in English and in German or Spanish.[43] We also take out words that are used to disclose advertising (see Appendix A.2 for a full list of these words). For each word, we then apply the language detection algorithm. If the language detection algorithms says the word is more likely to be German/Spanish than English, we apply the pre-trained multi-lingual German/Spanish embedding. If the algorithm says the word is more likely to be in English, we apply the multi-lingual English embedding. Once we converted each word, we compute the *average meaning* of a post by taking a simple average of all embeddings within the post.[44] At the end of the processing, each post is represented by a single 300-dimensional vector.

After text processing, we apply ML classifiers to the training data (300k German posts from

[40] i.e., subtracting the embedding of "paris" from "france" and adding the embedding of "tokyo" recovers the embedding for "japan" (Mikolov et al. 2013).

[41] i.e., the embedding for "cat" (in English) is approximately equal to the embedding for "gato" (in Spanish) and "katze" (in German).

[42] Applying this algorithm on a particular word gives a probability that the word is in a particular language. This determines whether a specific word is English, Spanish, or German and which embedding dictionary to apply.

[43] Stop-words are the most common words used in each language that do not add meaning to the text.

[44] There are other ways of aggregating word embeddings to extract the meaning of a sentence or paragraph. Some approaches using Recursive Neural Networks or weighted averaging seem to perform better than simple averaging (Arora et al. 2017). However, algorithms such as Recursive Neural Networks are too computationally expensive. For weighted averaging, it is not clear what the weights should be and whether they should be assumed to be the same between Germany and Spain. We consider simple averaging to be the most transparent and computationally feasible method.

2018). We use several algorithms that are commonly used in NLP and that perform well in text classification assignments (Silva et al. 2020): Gaussian Naive Bayes (NB), a linear Stochastic Gradient Descent with L1 loss (SGD), a Decision Tree (DT) and a Random Forest (RF).[45] The algorithms look for areas of the embedding space that are associated with the disclosure/sponsorship label.[46] The probabilistic classifiers (NB, DT and RF) compute the probabilities that different areas of the embedding space are associated with disclosure/sponsorship and then invert that mapping to compute a predicted probability that a post is disclosed/sponsored conditional on its average embedding distribution. The non-probabilistic classifier (SGD) constructs hyper-planes in the 300-dimensional embedding space to separate out disclosed/sponsored areas of that space from the undisclosed/non-sponsored. After classification is complete, we project the classifiers on our entire 15 million post dataset.

The ML algorithms allow for false-positives, which helps us address concerns about using disclosed-sponsored posts as a general stand-in for sponsored posts in the training data. Under imperfect disclosure, false-positives (Type I errors) are not prediction mistakes. Instead, they are likely to capture sponsored posts that are not disclosed even after regulation by finding posts that are near enough in meaning-space to disclosed sponsored posts but are not actually disclosed.[47] We tune ML classifiers' parameters so that they weight correctly predicting true-positive outcomes higher and are more forgiving on false-positive outcomes.[48] With this approach our classifiers uncover meanings associated with sponsorship rather than disclosure.[49]

4.4.3 Comparison of Classification Approaches

This section compares the predictions of our classifiers. We look at two different dimensions: (i) the quality of predictions as measured by Type I and Type II errors. (ii) how consistent are the predictions of different classifiers with one another. Both matter to whether we want to use one or several of the classifiers in the main analysis.

First, we use two metrics borrowed from the computer science literature to evaluate the performance of the classifiers. A *Recall Score* measures the number of correctly classified disclosed posts (true positives) as a share of the actual total number of disclosed posts. A classifier with higher Recall Score makes fewer Type II errors. We also look at the *Balanced Accuracy Score*, which is an average between the Recall Score (the true positive rate) and the true negative rate. Although the classifiers are tuned to be stricter on Type II errors, it is important to ensure that they do not go too far. A classifier labelling all posts as sponsored and disclosed will have no false negatives and no Type II errors, but will also be practically useless. The Balanced Accuracy Score accounts for

[45] Other algorithms such as XGBoost or Convolutional Neural Networks may produce better results, although recent computer science literature such as Silva et al. (2020) suggests this is mostly not the case.

[46] ML classifiers are often sensitive to imbalanced class data. As Figure 2 suggests, we have more undisclosed posts than disclosed posts in the training data. We re-balance the data by randomly oversampling the disclosed posts (with replacement) until the two classes are equal.

[47] In the three word dictionary/two-word post example from Section 5, if "Deal" is more frequently used together with "Sale" than with "Love," the algorithms will label "Deal Deal" as likely to be sponsored even if they are not disclosed.

[48] We perform the tuning for each algorithm using 5-fold stratified cross-validation. Since there is a concern that the algorithms will attempt to reach a perfect true positive rate by labelling every post as a "positive," we search the parameter state-space for parameters that maximize a combination of the Recall Score (the true-positive rate) and the Balanced Accuracy score which incorporates the false negative rate.

[49] In Section A.12 we verify that our ML predictions for undisclosed posts are consistent with a classification coming from a MTurk survey.

Table 1: Comparison of Classification Approaches

	(1) Gaussian Naive Bayes	(2) SGD w/ L1 Loss	(3) Decision Tree	(4) Random Forest
Mean Recall Score	0.816	0.715	0.761	0.737
Mean Balanced Accuracy Score	0.618	0.677	0.665	0.703

Notes: The Recall Score and Balanced Accuracy Scores are computed as averages from 5-fold stratified cross validation. Each column in the table looks at the performance of a different classifier, trained on embeddings generated from post text as discussed in Section 4.4.2. Each ML classifier was trained on a sample of 300,000 German post-disclosure regulation posts as described in the main text.

this.[50]

Table 1 shows the performance of the two metrics for the different classifiers. For each classifier we show the average of the two scores based on 5-fold cross-validation applied to the training sample of 300,000 post-regulation German posts.[51] We find that all classifiers perform well in terms of Recall Scores. They largely correctly uncover true positives while minimizing false negatives. The NB approach catches nearly 82% of the actually disclosed posts (on average), but the other approaches also manage to catch over 70% of actually disclosed posts.[52] However, Table 1 also shows a trade-off between Type I and Type II errors. Classifiers with higher Recall Scores have lower Balanced Accuracy Scores and vice versa.

Next we test whether the classifiers' predictions agree with one another. We start by comparing the classifiers' predictions for the sample Kylie Jenner posts from Figure 1. Table 2 shows the predicted labels for each post. As expected, the classifiers generally agree that the posts in Panels (a) and (c) should be labelled as sponsored. Both of these posts are also disclosed as sponsored, although the disclosure in the Panel (c) post is weaker. Similarly, no classifier predicts the post from Panel (b) as sponsored because there is no text with which it could sponsor something. There is more disagreement on the post from Panel (d), which is most ambiguous. Both the manual and the Decision Tree/ Random Forest approaches label it as sponsored but the Naive Bayes and SGD do not.

Table 2 only looks at four posts. We also compare measures of "agreement" between the classifiers on the full sample of German and Spanish posts. In Table A1 we show pairwise agreement percentages and measures of Cohen's κ for the different classifiers. We do this separately for disclosed and undisclosed posts. We find that the classifiers have relatively high measures of agreement and correlation in predictions but are far from unanimous. The high degree of correlation is consistent with the good Recall score performance of the classifiers. However, while on average the different classification approaches are consistent, they may also be picking up different facets of

[50] Another popular measure that balances out Type I and Type II errors is the Mean ROC-AUC Score. Results from this measure are similar to those from the Balanced Accuracy Score.

[51] We also tested the performance of the classifiers, including the manual classifier on the full data. Recall Scores from the full data are similar to those from the training sample and the manual classifier has a recall score similar to the Naive Bayes classifier (close to 80%). As part of this exercise we checked whether the performance of the classifiers changes over time and before and after the regulatory change in Germany. We find that most classifiers perform worse in the pre-regulation period in Germany. However, the manual classifier, which is not trained on any particular time period, also performs worse in the pre-regulation period. This reflects the likely selection of disclosed sponsored posts in the pre-regulation period rather than any changes in language over time. The exception is the SGD classifier, which performs equally well before and after regulations.

[52] Crucially, the classifiers are applied *after* removing any words that denote disclosure, as discussed in the previous section.

Table 2: Classification of Sample Posts from Figure 1

	(1) Manual	(2) Gaussian Naive Bayes	(3) SGD w/ L1 Loss	(4) Decision Tree	(5) Random Forest
Panel (a)	Spon.	Spon.	Spon.	Non-Spon.	Non-Spon.
Panel (b)	Non-Spon.	Non-Spon.	Non-Spon.	Non-Spon.	Non-Spon.
Panel (c)	Spon.	Non-Spon.	Spon.	Spon.	Non-Spon.
Panel (d)	Spon.	Non-Spon.	Non-Spon.	Spon.	Spon.

Notes: Data for each row in this table comes from the text of the posts in Figure 1. Column (1) applies the manual approach to classifying sponsored and non-sponsored posts as outlined in Section 4.4.1. Columns (2)-(5) show the predicted labels for each post from projecting a trained ML classifier on embeddings, which were generated from post text as discussed in Section 4.4.2. Each ML classifier was trained on a sample of 300,000 German post-disclosure regulation posts as described in the main text.

sponsorship/advertising.

We also compare the predictions made by the ML classifiers to results from an MTurk survey for a sample of post-regulation German posts in Appendix A.12. In the survey, we asked respondents to classify captioned as sponsored or non sponsored. We find strong correlation between both the manual classification and the ML approaches and the classification of MTurk survey respondents.

Overall, there are trade-offs between the different classifiers. More "complex" classifiers like Random Forest perform better according to some measures (Balanced Accuracy), but not according to others (Recall Score). Although, as explained in Section 4.4.2, we tune our classifiers to be more lenient towards false-positive predictions, we do not have priors about which trained classifier we should use. Some classifiers may be too lenient and allow for too many false-positives. It is also possible that the different classifiers pick up different advertising posts, or even different types of advertising. We choose to use all four ML classifiers in the rest of the analysis.

4.5 Matching and Influencer-Month Summary Statistics

After the classification procedure described in Sections 4.3 and 4.4, we merge the post-level data with monthly country level data. The post-level data includes 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.[53]

An important concern with our data is the comparability of German and Spanish influencers. Although the influencers were selected randomly, it does not mean that the average German influencer in our sample is necessarily similar to the average Spanish influencer. We formally test this in Table 3 by comparing the average monthly characteristics of influencers in Spain and in Germany in 2015 (over one year before any regulatory changes in Germany). We do this by regressing the average monthly number of likes per post, number of comments per post, and the average number of posts for an influencer on a dummy that is equal to 1 when the influencer is German. In Columns (1)-(3) we do this for the full sample of influencers in 2015. These regressions suggest

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

Table 3: Balancing Test for Main Observables

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Sample:		Full Sample			CEM Matched Sample			Random 10% Sample	
Pre-Treatment Variable:	Mean Likes	Mean Comments	N Posts	Mean Likes	Mean Comments	N Posts	Mean Likes	Mean Comments	N Posts
Germany	309.244***	5.353***	-6.271***	-32.480	-2.208	-1.705	246.608**	1.091	-7.075***
	(39.046)	(0.957)	(0.567)	(77.339)	(1.781)	(1.085)	(117.621)	(4.059)	(1.398)
Observations	248,981	248,981	248,981	30,913	30,913	30,913	28,959	28,959	28,959

Notes: Sample for Columns (1)-(3) includes all influencer-month observations from 2014 and 2015. Sample for Columns (4)-(6) includes all influencer-month observations from 2014 and 2015 for influencers who were "matched" by the Coarsened Exact Matching procedure explained in Section 4.5. Sample for Columns (7)-(9) includes influencer-month observations from 2014 and 2015 for 600 randomly selected German influencers and 600 randomly selected Spanish influencers. Standard errors are clustered at the influencer level. *** p<0.01, ** p<0.05, * p<0.1.

that the average characteristics of German influencers in our full sample are statistically different than for Spanish influencers before German regulations. In 2014 and 2015, German influencers have 300 more likes per post, 5 more comments per post, and they post 6 fewer posts per month than Spanish influencers.[54]

To address this concern we restrict our sample by matching German influencers to observationally similar Spanish influencers. We use Coarsened Exact Matching (CEM), a matching method commonly used in economics literature to balance covariates (Azoulay et al. 2019, Sarsons 2019, Aneja and Xu 2020).[55] To apply this method, we use the following average characteristics of influencers in 2015 (one year before regulations were introduced in Germany): the mean number of monthly posts, the mean number of comments per post and the mean number of likes per post. In addition to these characteristics, we match influencers based on their market/industry. HypeAuditor provided us with broad "Topics" associated with each influencer, such as "Acting," "Modelling," "Dogs," "Mom," "Cars," etc. We consider a topic to be an influencer's industry. We exactly match the German and Spanish influencers on their market/industry and on their starting year on Instagram, and we coarsely match the other 2015 characteristics.

After matching we are left with only 10% of our original sample, but this still accounts for approximately 600 influencers from Germany and a similar number from Spain.[56] The share of matched to non-matched observations is similar to other applications of this approach (Azoulay et al. 2019, Sarsons 2019). In Columns (4)-(6) of Table 3 we repeat the exercise from Columns (1)-(3) using this sample. We find that there are no average statistically significant differences in observable characteristics between the matched German and Spanish influencers. The magnitude of the coefficients is also much smaller than for the full sample.

One concern with results from Columns (4)-(6) may be that we do not find statistically significant differences simply because we drop 90% of observations. The smaller sample size increases the

[54] These differences are not just statistically significant but meaningful in magnitude. Table A5 shows the summary statistics for the full sample. An influencer on average has 1,900 likes per post, so the German-Spanish average difference is 15%. The difference in the number of posts per month is nearly 30%.

[55] CEM was first described in Iacus et al. (2008). This approach "coarsens" some covariates into bins and discards treated and control observations that do not fall into the same bins. Other covariates are used "exactly" - i.e., treatment and control observations have to have the same value of the covariates to not be discarded. CEM is non-parametric and does not involve functional form assumptions embedded in other matching methods such as propensity score matching. This is not a 1:1 matching model, so there may be multiple control observations that are matched to a given treatment observation based on covariates or vice-versa.

[56] More precisely, we are left with 618 influencers from Germany and 560 influencers from Spain. The matching procedure is not 1:1.

Table 4: Influencer/Month Summary Statistics for Germany and Spain (CEM Matched Sample)

Variable	Obs	Mean	Std. Dev.
Mean Likes per Post	67,235	1,651	3,870
Mean Comments per Post	67,235	46	229
N Followers	14,567	92,492	89,085
N Posts per Month	67,235	19	21
Account Age (months)	67,235	40	21
First Account Year	67,235	2013	1
Mean Disclosed Post Share	67,235	.109	.217
Mean Manual Predicted Sponsored Post Share	67,235	.534	.298
Mean NB Predicted Sponsored Post Share	67,235	.404	.355
Mean SGD Predicted Sponsored Post Share	67,235	.429	.266
Mean RF Predicted Sponsored Post Share	67,235	.342	.25
Mean DT Predicted Sponsored Post Share	67,235	.235	.295

standard errors and reduces statistical significance, but there could still be underlying differences. To address this issue we perform the following exercise: we randomly select the same number of influencers as are in the matched sample (i.e., 600 from Germany and 600 from Spain) and repeat the comparisons of observable characteristics. Results from the fully random sample of influencers are in Columns (7)-(9) of Table 3. We find that in the non-matched random sample of 1,200 influencers there are still very large differences in observables between Germany and Spain. Coefficients from Columns (7)-(9) are statistically significant and similar in magnitude to Columns (1)-(3).

Overall, balancing tests in Table 3 show that CEM trims the data such that the matched German and Spanish influencers are comparable to one another. We use the CEM matched sample in the main analysis below.[57]

Summary statistics for the CEM-matched sample are in Table 4.[58] Influencers in our sample have an average 1,650 likes per post, and an average 46 comments per post. The average influencer has approximately 90,000 followers.[59] The influencers in our sample generate a substantial amount of content. An average influencer account is 40 months old and generates 19 posts per month. Approximately 11 percent of monthly posts are disclosed as sponsored.[60] On average, between 24 and 53 percent of monthly posts are predicted to be sponsored by our various classification methods. Additional descriptive analysis of our predicted sponsorship measures is in the next section.

[57] As a robustness check, we repeat the main analysis using the non-matched sample. The results are broadly similar, although the difference-in-differences coefficients in the non-matched regressions appear to be biased upwards relative to the matched sample. Regression estimates are in Tables A6-A8.

[58] Additional summary statistics for the full non-matched sample are available in Table A5 in the Appendix.

[59] Follower counts are likely biased upwards since we do not observe followers for all influencers and for all time periods, especially for earlier years. This is an issue with the raw data and the reason for why the number of observations is smaller than for other characteristics.

[60] These summary statistics capture both Spain, where there are no disclosure regulations, and Germany prior to disclosure regulations.

5 Model

This section introduces a conceptual framework for our empirical work. The model has a continuum of two types of agents, influencers (who choose what type of post to make) and followers (who decide how much engagement they spend on a particular post). Influencers differ ex post in terms of their value of posting sponsored content, but otherwise each type of agent is identical to other agents on their side of the market. [61] As in Fainmesser and Galeotti (2020), we model a static interaction between an influencer and follower to highlight the incentives of influencers and their followers. The model shows the ambiguous effect that disclosure regulations might have, and therefore the need for measurement.

5.1 Influencers

To model the contrast between sponsored and organic posts, we assume that posts are one of two types, $i \in \{A, B\}$. Type A posts are organic and type B posts are sponsored. Posts are made out of *words*. A post of type i generates a draw of a collection of words $\omega \subset \Omega$ drawn from a distribution $f_i(\omega)$. The set Ω is often referred to in natural language processing as the *dictionary*. We assume that the distributions have common support S. Below we specialize to the case where Ω has just two or three words and posts contain a single word to illustrate some specific features of the model.

Suppose an influencer has an occasion to make a new post. We assume that the influencer privately chooses between post types and does not choose individual words. The choice of post type i leads to a draw of words for that individual post.[62] This mapping from choice of i to words matches the notion that when an influencer chooses an ad campaign most words are determined by the advertiser.

The choice of post type is driven by a trade-off between revenue and follower engagement. For each posting occasion, the influencer realizes a random draw γ from some distribution with CDF $\Gamma(\gamma)$. Fixing engagement, revenues are lower for type A posts by a factor of $\gamma < 1$ relative to B posts. We assume that type B posts generate higher payoffs for the influencer but lower payoffs for a follower (see Section 5.2 below), to highlight the tradeoff between quality and sponsorship.[63]

The influencer's payoff, fixing the type of post, is proportional to engagement $x(\omega) \geq 0$ by followers given words ω. The determination of $x(\omega)$ will come from the followers and is described in Section 5.2 below. The expected engagement for a type i post is:[64]

$$\int x(\omega) f_i(\omega) d\omega$$

In choosing which type of post to make, the influencer compares $\int x(\omega) f_B(\omega) d\omega$ to $\gamma \int x(\omega) f_A(\omega) d\omega$. This allows the following immediate result:

Proposition 1. *If $\int x(\omega) d\omega > 0$ then there exists a cutoff γ^* such that an influencer posts a type*

[61] The model can be modified to allow for further heterogeneity.

[62] 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.

[63] This is both standard in the literature (e.g., Mitchell 2021,Fainmesser and Galeotti 2020) and consistent with our data. Average engagement for sponsored posts is lower than engagement for non-sponsored posts. See Table 10 for more detail.

[64] All integrals are over the support S of f_i

B post if it draws $\gamma < \gamma^*$ where

$$\gamma^* = \frac{\int x(\omega) f_B(\omega) d\omega}{\int x(\omega) f_A(\omega) d\omega}$$

If $\int x(\omega) d\omega = 0$ then any choice of post is optimal.

The observed distribution of words is therefore a mixture of the two distributions, with the probability of a B post being $\Gamma(\gamma^*)$. Assuming multiple independent posting occasions, this means the share of B posts the influencer posts is also $\Gamma(\gamma^*)$. The equilibrium determination of γ^* depends on follower engagement, which we model next.

5.2 Consumers/Followers

The follower chooses costly engagement with a post based on their beliefs about its type given its content ω. The return to each unit of engagement is r_A for posts of type A and r_B for type B, where $r_A > r_B$. Without loss let $r_B = 0$ and $r_A = 1$. The cost of engagement $x(\omega)$ on a post of words ω is $c(x)$; since engagement is costly, the follower prefers engagement with likely A posts and not likely B posts. Let the follower's posterior probability of a post being type A be $g(\omega)$. Given a set of words, they choose engagement $x(\omega)$ to solve

$$x(\omega) = arg \max_x g(\omega)x - c(x)$$

For simplicity of notation we assume that $c(x) = ln(x)$, but it is direct to generalize, for instance, to $c(x) = \frac{\epsilon}{\epsilon+1} x^{(\epsilon+1)/\epsilon}$ for $\epsilon > 0$ so that $x(\omega) = g(\omega)^\epsilon$.

5.3 Equilibrium

With some abuse of notation let the belief about a post of words ω given the equilibrium posting cutoff γ for influencers be $g(\omega|\gamma)$, so that

$$g(\omega|\gamma) = \frac{(1-\Gamma(\gamma)) f_A(\omega)}{\Gamma(\gamma) f_B(\omega) + (1-\Gamma(\gamma)) f_A(\omega)}$$

Suppose $\gamma < 1$. Then $g(\omega) > 0$ for all ω, and therefore $\int g(\omega|\gamma) d\omega > 0$. As a result, for $\gamma < 1$ we can define the optimal influencer cutoff by

$$R(\gamma) = \frac{\int_\Omega g(\omega|\gamma) f_B(\omega) d\omega}{\int_\Omega g(\omega|\gamma) f_A(\omega) d\omega}$$

For $\gamma = 1$, $g(\omega) = 0$ for all ω, and so let $R(1) = [0, 1]$ since any cutoff can be optimal. The correspondence $R(\gamma)$ denotes the optimal response of influencers if followers expect them to use cutoff γ; we therefore define

Definition. A cutoff γ^* is an equilibrium if $\gamma^* \in R(\gamma^*)$

Notice that $\gamma^* = 1$ can always be an equilibrium: every post is type B, and no attention is paid. This equilibrium, however, could be broken by a small amount of attention paid to every post independent of $g(\omega)$. We therefore focus on equilibria where attention is paid, i.e. equilibria where $\gamma < 1$ and where $R(\gamma)$ is single valued. These equilibria can be written as $\gamma^* = R(\gamma^*)$.[65] In these equilibria the probability of a given B post and the share of B posts the influencer makes is $\Gamma(R(\gamma^*))$.

5.4 Regulation

5.4.1 Word-Independent Regulation and a Two-Word Example

Suppose that regulation makes a fraction r of all chosen $B-$posts (chosen independently of ω, given B) add a disclosure word that is never present in type A posts.[66] This allows us to consider both disclosed-sponsored and undisclosed-sponsored posts. We assume regulation does not change the number of posting occasions and total number of posts by the influencer.[67] Compared to the equilibrium without regulation, the relevant beliefs are for undisclosed posts, since disclosed posts are all known to followers to be type B. Let $g(\omega)$ denote beliefs for undisclosed posts. The return to a type-B post is therefore

$$(1-r)\int x(\omega)f_B(\omega)d\omega$$

so that the cutoff rule for choosing B is

$$\gamma^* < (1-r)\frac{\int g(\omega)f_b(\omega)d\omega}{\int g(\omega)f_a(\omega)d\omega} \tag{1}$$

The first term on the left hand side is the taxation effect: $r > 0$ makes some messages disclosed and have minimal engagement, reducing the productivity of type B posts. At the same time, disclosure also changes the second term (the ratio) on the left hand side, which could make B posts more trusted relative to A posts as compared to a world without disclosure regulations. The net effect of regulation on the cutoff γ^* is ambiguous.

A simple two-word dictionary and single-word post example clearly illustrates how regulation impacts the incentive to send sponsored posts. Suppose that each post is one of two words ("Love" and "Sale"), $\Omega = \{l, s\}$. Posts of type i are a fraction p_i "Love", where $p_A > p_B$ ("Love" is more likely to appear in Type A posts). Then define

$$\hat{R}(\gamma, r) = \frac{\int g(\omega|\gamma, r)f_b(\omega)d\omega}{\int g(\omega|\gamma, r)f_a(\omega)d\omega} = \frac{p_B V(\gamma, r) + (1-p_B)}{p_A V(\gamma, r) + (1-p_A)}$$

where $V(\gamma, r) > 1$ is the ratio of $g(l|\gamma)$ to $g(s|\gamma)$. Direct calculations verify that $V(\gamma, r)$ is increasing in γ and decreasing in r. Since $p_B < p_A$ it is immediate that:

Proposition 2. *$\hat{R}(\gamma, r)$ is decreasing in γ and increasing in r*

[65] We verify below in the two word example that R is decreasing in γ for $\gamma < 1:$ more type B posts makes followers have less engagement with type B-looking posts. Therefore the equilibrium where attention is paid is unique in that case. However, in general there is no necessity to that monotonictity, as we discuss below.

[66] Such imperfect porous regulation is consistent with the types of disclosure regulations instituted in many jurisdictions. Due to ambiguities in legislative language or in the enforcement of legislation it is often unclear whether some sponsored posts require disclosure. Section 3.2 and Online Appendix A.1 provide more details about German, French, Italian and US regulations.

[67] In the empirical results we show that policy changes do not significantly alter post volume.

The result highlights the opposing impacts of regulation. The cutoff is determined by $(1 - r)\hat{R}(\gamma, r)$; on the one hand there is the direct effect, that disclosed sponsored posts receive less engagement, lowering the return to a B post. On the other hand, the relative trustworthiness of "Love" compared to "Sale" declines with regulation, increasing the incentive to make B posts.[68] As in the general case above, the net effect of regulation on the cutoff is ambiguous. When regulation decreases B posts it unambiguously increases welfare: followers like A posts better and moreover, some of the remaining B posts are filtered out by disclosure. However, when regulation increases B posts the conclusion for welfare is ambiguous: some of the additional B posts are filtered out but followers also end up (rationally) engaging more with undisclosed B posts as compared to A posts. The key takeaway is that the effects of a disclosure regulation on undisclosed sponsored content is key to understanding its welfare effects.

5.4.2 Three Words and Word-Specific Regulation

To see how regulation and the impact on posts could be even more complicated, suppose that $\Omega = Love, Sale, Deal$, and that *Sale* and *Deal* are relatively more likely in B posts. Suppose also that only the word *Sale* gets regulatory scrutiny, so that a fraction r of all sponsored posts with the word *Sale* are labelled with *#ad* but the other two words are never marked.[69]

For fixed λ, the posterior, and therefore the engagement, with *Love* or *Deal* are unchanged with regulation since those words are not regulated. As before, trust in undisclosed *Sale* posts rises with regulation, therefore making the return to type B posts potentially higher, and increasing the equilibrium proportion of those posts. Since *Sale* and *Deal* are both more likely with B type posts, this can increase both *Sale* and *Deal* posts. As *Deal* posts are not disclosed but more likely to be sponsored, this can drive an increase in specifically undisclosed, sponsored ads.

Importantly, with word specific regulation and many words there is not necessarily a monotonic relation between the proportion of undisclosed posts that come from the B distribution and the ratio of expected attention paid to B versus A posts.[70] In other words, it is an open question whether regulation makes the share of sponsored content among undisclosed posts go up or down, further motivating our empirical exercise. As with word-independent regulations, the relative engagement of undisclosed-sponsored content relative to non sponsored content is key to understanding the potential welfare effects of word-specific regulations.

Our model assumes that regulation may depend on words, but content determines words. If words can be chosen strategically, a regulated influencer may try to choose words to avoid regulation, making post-regulation A and B type posts more similar. Such strategic word choice would make it more difficult to find any differences in sponsorship across regulated and unregulated regimes.[71]

[68] In this two word example, an increase in sponsored content with regulation can only come if the sponsored-share of undisclosed posts falls. This need not be true in more complicated multi-word post settings.

[69] The example also works if the regulator regulates *Sale* posts independent of which category they come from.

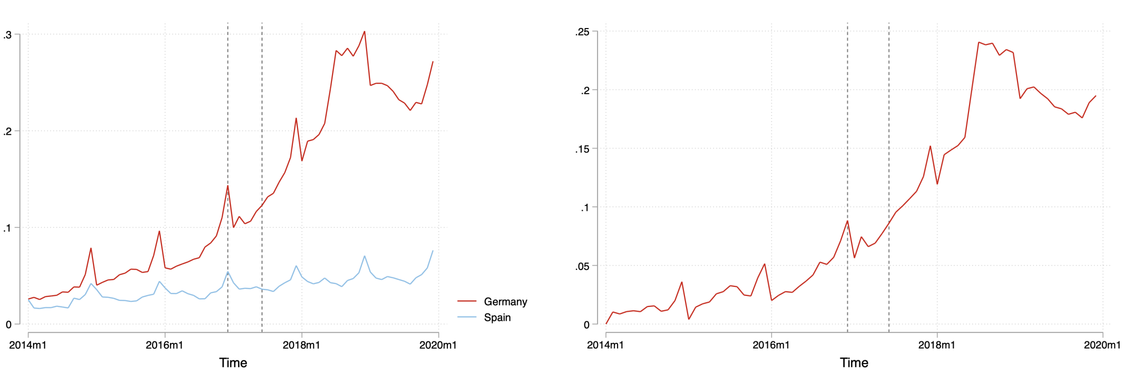
[70] Formally, the ratio of B to A posts is an expectation with respect to a constant, and the ratio R is an integral with respect to a function that shifts with regulation. The two word example adds structure that guarantees monotonicity.

[71] We leave modelling of strategic word choice, and the nature of regulation across different words, to future work.

Figure 2: Share of Posts Disclosed as Advertising in Germany and Spain

(a)

(b)



Notes: Each line in panel (a) shows the total number of posts labelled as "disclosed" advertising over the total number of posts in month t in Germany or Spain ($\frac{N Disclosed Posts}{N Posts}$). A post is labelled as disclosed if it includes one of the disclosure words from Appendix A.2. Full (non-CEM matched) sample of influencers used. The line in panel (b) shows the difference between the disclosure share in Germany and the disclosure share in Spain in month t. This difference is normalized relative to the difference in January 2014 (the first month of the sample). The first dashed vertical line represents the initial changes to German disclosure regulations in November 2016 (see Section 3.2). The second dashed vertical line represents the first fines handed out to German influencers in mid 2017.

6 Descriptive Evidence

6.1 Disclosure in Germany and Spain

Figure 2 shows the percentage of all posts disclosed as advertising in Germany and Spain during our sample period in panel (a) and the difference between German and Spanish percentages in panel (b). 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 3.2 for more details). Disclosure in Germany increases dramatically around changes in the regulatory environment.[72] There are no similar changes in disclosure in Spain over the same period.

6.2 Sponsorship in Germany and Spain

We plot differences in monthly percentages of predicted sponsored posts between Germany and Spain in Figure 3.[73] 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. We then subtract Spain's ratio from Germany's ratio and normalize it relative to the difference in January 2014. On the left panel of the figure, sponsored posts are labelled exclusively by the ML classifiers. On the right panel, a post is only labelled as sponsored if both a ML classifier and our manual approach classify it as

[72] We test for changes in disclosure more formally in Table 7.

[73] Time series of non-differenced country-level mean sponsored post shares are in Figure A1 in the Appendix.

sponsored. In each panel, we also plot the mean difference in ratios between Germany and Spain across all predictions.

Figure 3 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.[74] These preliminary findings are consistent with the theoretical model and suggest that content in Germany is changing in response to changes in the regulatory environment.[75]

Notably, changes in the shares of sponsored posts in Germany and Spain in Figure 3 do not perfectly track changes in disclosure from Figure 2. For example, there are several classifiers which suggest that Spain had more sponsorship than Germany in 2015. The differences between Germany and Spain in terms of sponsorship rates are much smaller on average than differences in disclosure rates. This suggests the existence of a large number of undisclosed but sponsored posts in Spain (also see Figure A1 in the Appendix). Altogether, we are uncovering something novel about the underlying text data through our classification procedure, rather than simply re-stating the changes in disclosure.

7 Estimation Methodology and Results

7.1 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.[76] We model outcome Y_{it} (i.e., share of sponsored posts) for influencer i at month t as:

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

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.[77] X_{it} are a set of influencer/time varying controls, such as account age and country characteristics (i.e., popularity of Instagram,

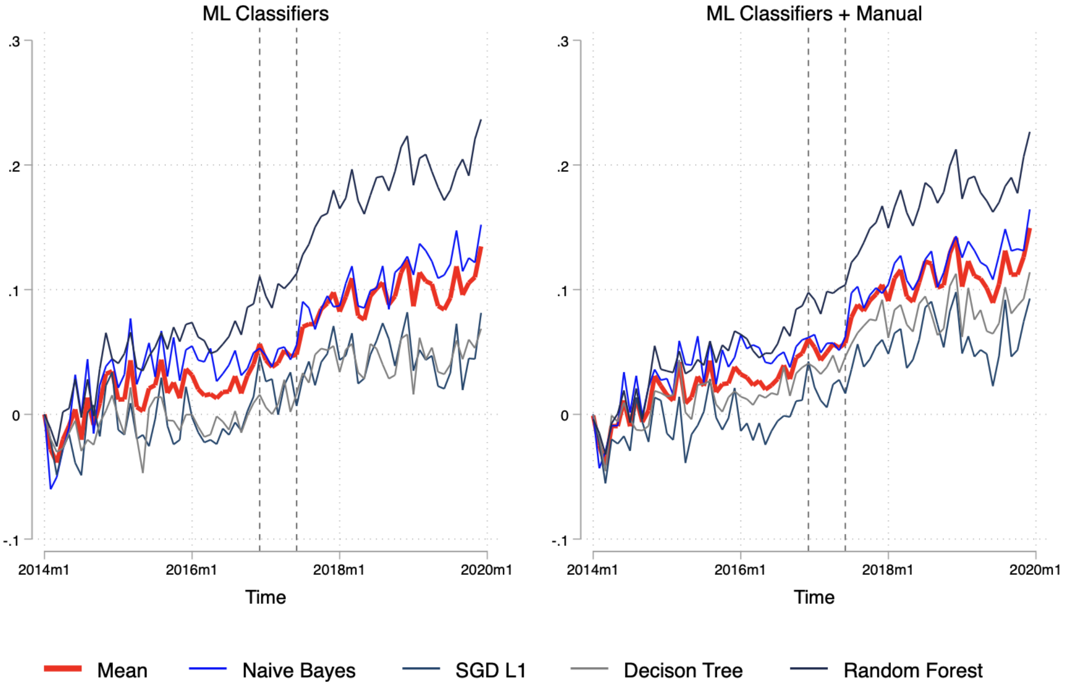
[74] Additional results show that sponsorship in Germany is also increasing in the absolute, rather than just the relative sense. Time series of non-diffferenced country-level mean sponsored post shares are in Figure A1 in the Appendix.

[75] We show similar effects on content without relying on any classification in Appendix A.8. 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 A4, 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 A3.

[76] In Appendix A.6 we also estimate effects for *post*-level outcomes.

[77] We choose November 2016 since the "clarification" to German disclosure regulations came out in October 2016 (Section 3.2). This is likely understating the true effects of the changes, as the enforcement of regulations started in the middle of 2017. See Appendix A2 for period-specific effect estimates.

Figure 3: Differences in Predicted Sponsored Post Shares Between Germany and Spain for Different Classifiers



Notes: Each line represents the difference in the average share of predicted sponsored posts of the total number of posts ($\frac{\text{N Predicted Sponsored Posts}}{\text{N Posts}}$) between Germany and Spain in month t according to one of four ML classifiers (Gaussian Naive Bayes, SGD L1, Decision Tree, Random Forest). All differences are normalized relative to the difference in January 2014 (the first month of the sample). The solid red line represents the mean of the four other lines. CEM matched sample of influencers used. In data used to generate left-hand panel figure, posts are labelled as sponsored by ML classifiers only. In data used to generate right-hand panel figure, posts are only labelled as sponsored if both a ML classifier and the manual approach classifies them as sponsored. The first dashed vertical line represents the initial changes to German disclosure regulations in November 2016 (see Section 3.2). The second dashed vertical line represents the first fines handed out to German influencers in mid 2017. Time series of non-diffferenced country-level mean sponsored post shares are in Figure A1 in the Appendix.

GDP per capita). δ_i and δ_t are influencer and year/month fixed effects which absorb country and Treated Period$_t$ fixed effects.

There are several concerns with the ability of this estimation strategy to capture the average treatment effect of the policy. One concern might be anticipation effects, or other country-time specific shocks that are correlated with the timing of the policy. Regulatory events in other countries (such as FTC regulations in the US) may also have spillovers on Germany and Spain. The policy itself might also be endogenously driven by pre-regulation behaviour of influencers in Germany. For example, increasing advertising on Instagram in Germany may have incentivized strict regulation.

All of these concerns would be reflected by diverging pre-regulation period patterns between Germany and Spain. We test for these directly with a timing test in Appendix A.7. As well, we include a number of country/time varying controls to try to capture demand shocks, such as the popularity of Instagram and GDP per capita (which may influence consumption or advertising behaviour). With respect to FTC regulations, there is no reason why regulations in another country would affect influencers in Germany and Spain differentially.

Another concern relates to the comparability of influencers in Germany and Spain. Even though our data contains a random sample of influencers, they are not necessarily similar. The influencer names we received were not randomly sampled *within* a particular market (i.e., fashion influencers) or based on any characteristics except having more than 10,000 followers in May 2019. We address this concern by restricting our sample to Coarsened-Exact-Matched (CEM) influencers as described in Section 4.5.[78]

7.2 Main Results

This section shows our main difference-in-differences regression estimates in five tables. Table 5 shows the effects of intensified disclosure regulations on advertising/sponsorship rates. This is measured by the number of sponsored posts over the number of total posts at the influencer/month level. Table 6 shows the effects of intensified disclosure regulations on the rates of sponsorship among posts that are not disclosed as sponsored. We do this separately to test whether disclosure regulations affect the content of *all* posts rather than just the disclosed posts. Table 7 shows the effects of changes in disclosure regulations on other outcomes at the influencer/month level, such as the mean number of posts. Table 8 shows the effects of disclosure regulation on follower engagement measures: the average number of likes received by an influencer, average number of comments and the average number of followers. Table 9 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 5 shows the first set of results. Outcome variables in the table are all shares - the number of predicted sponsored posts over the number of total posts for an influencer/month observation. Each column uses a different ML algorithm to label posts as sponsored. Columns (1) and (5) use NB, Columns (2) and (6) SGD, Columns (3) and (7) DT and Columns (4) and (8) RF. In data used to generate the top panel (Columns 1-4), posts are labelled as sponsored by ML classifiers only. In data used to generate the bottom panel (Columns 5-8), posts are only labelled as sponsored if both a ML classifier and the manual approach classifies them as sponsored. 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

[78] We also replicate our main analysis with the non-matched sample. The results are qualitatively similar to the main estimates and are available in Appendix A.10.

Table 5: Influencer/Month DiD Estimates - Sponsored Share

	(1)	(2)	(3)	(4)
Outcome:		Predicted Sponsored Shares		
Classifier:	Naive Bayes	SGD L1	Decision Tree	Random Forest
Germany × Treated Period	0.034***	0.046***	0.038***	0.076***
	(0.011)	(0.008)	(0.007)	(0.009)
Pre-Treatment Mean	0.277	0.382	0.264	0.142
Country Controls	YES	YES	YES	YES
Influencer FE	YES	YES	YES	YES
Year-Month FE	YES	YES	YES	YES
Account Age FE	YES	YES	YES	YES
Account Age × First-Account-Year FE	YES	YES	YES	YES
Observations	67,235	67,235	67,235	67,235
R-squared	0.670	0.522	0.525	0.674
	(5)	(6)	(7)	(8)
Outcome:		Predicted Sponsored Shares		
Classifier:	Naive Bayes +Manual	SGD L1 +Manual	Decision Tree +Manual	Random Forest +Manual
Germany × Treated Period	0.038***	0.045***	0.038***	0.071***
	(0.010)	(0.008)	(0.008)	(0.009)
Pre-Treatment Mean	0.201	0.216	0.168	0.115
Country Controls	YES	YES	YES	YES
Influencer FE	YES	YES	YES	YES
Year-Month FE	YES	YES	YES	YES
Account Age FE	YES	YES	YES	YES
Account Age × First-Account-Year FE	YES	YES	YES	YES
Observations	67,235	67,235	67,235	67,235
R-squared	0.655	0.571	0.565	0.656

Notes: 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 4.5. The dependent variable in each regression is the number of posts that were labelled as sponsored for influencer i in month t as a share of the total number of posts made by influencer i in month t. Each column uses a different ML classifier to label posts as sponsored. The classifiers use embeddings generated from post text and are trained on a sample of 300,000 German posts as discussed in Section 4.4.2. In data used for the top panel, posts are labelled as sponsored by ML classifiers only. In data used for the bottom panel, posts are only labelled as sponsored if both a ML classifier and the manual approach classifies them as sponsored. "Germany × Treated Period" is a dummy equal to one for all German influencer observations after November 2016 and zero 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.

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 suggest that there is a statistically significant increase in the share of sponsored posts in Germany after the strengthening of disclosure regulations. This result holds for all ML classifiers on their own and when combined with the manual classifier. The changes in sponsored post shares are large in magnitude. Relative to an average Pre-Treatment Mean advertising rate of between 10 and 30 percentage points (i.e., one in 10 to one in three posts), the increase after regulations is between 3.4 and 7.6 percentage points. At the minimum (in Columns 1 and 5) sponsored shares increased by 3.4/3.8 percentage points or approximately 12%. At the maximum (in Columns 4 and 8) sponsored shares increased by 7.1/7.6 percentage points or by over 50%. This is consistent with the prediction of Fainmesser and Galeotti (2020) that sponsorship unambiguously rises with disclosure regulation, and a risk that can occur in Mitchell (2021) and the model presented in Section 5.

Our model suggests that an important question about the results in Table 5 is whether the changes in sponsorship rates occur only for disclosed posts. In our model as well as in Fainmesser and Galeotti (2020), a sponsored post that is disclosed can be ignored, which improves welfare for a given level of sponsorship. Table 6 shows that the sponsorship increase is not driven just by disclosed sponsored posts. Outcome variables in this table are the share 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. As in Table 5, the columns reflect different classification methods to predict sponsorship. Once again, all estimates show that there is a statistically significant increase in sponsored content in Germany after changes to regulations. Point estimates in Table 6 are smaller than in Table 5, suggesting that some of the increase in sponsored content reflects *disclosed* sponsored content. However, there is still a substantial increase in *undisclosed* sponsored/advertising content. On average, point estimates show that undisclosed sponsored content increased by approximately 2 percentage points (although Columns 4 and 8 suggest it could be as high as 5 percentage points). This corresponds to an increase of around 10% on average. Interpreting these results through the lens of our model suggests that consumer welfare could decrease in response to the strengthening in regulations.

Table 7 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 2 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 percent. The increase in disclosure reflects two different factors: the disclosure of existing sponsored content, and the increase in disclosed and sponsored content. However, as Tables 5 and 6 show, there was also an increase in undisclosed sponsored content.

Column (2) in Table 7 looks at the effects of disclosure regulations on sponsored post shares using the manual approach to label posts as sponsored (see Section 4.4.1). Results are consistent with those in Tables 5 and 6. Column (3) shows that the number of posts per month for influencers in Germany relative to influencers in Spain do not change. This suggests that the increases in sponsored post shares in Tables 5 and 6 reflect changes in the number of sponsored posts rather than the number of posts that influencers are posting or other strategies.[79]

[79] This is also consistent with evidence from Appendix A.8 which uses embedding data to show that post text itself

Table 6: Influencer/Month DiD Estimates - (Sponsored Share | Non-Disclosure)

	(1)	(2)	(3)	(4)
Outcome:	(Predicted Sponsored Shares | Non-Disclosure)			
Classifier:	Naive Bayes	SGD L1	Decision Tree	Random Forest
Germany × Treated Period	0.022*	0.025***	0.018**	0.055***
	(0.012)	(0.008)	(0.007)	(0.009)
Pre-Treatment Mean	0.269	0.373	0.257	0.133
Country Controls	YES	YES	YES	YES
Influencer FE	YES	YES	YES	YES
Year-Month FE	YES	YES	YES	YES
Account Age FE	YES	YES	YES	YES
Account Age × First-Account-Year FE	YES	YES	YES	YES
Observations	65,984	65,984	65,984	65,984
R-squared	0.644	0.474	0.466	0.625

	(5)	(6)	(7)	(8)
Outcome:	(Predicted Sponsored Shares | Non-Disclosure)			
Classifier:	Naive Bayes +Manual	SGD L1 +Manual	Decision Tree +Manual	Random Forest +Manual
Germany × Treated Period	0.022**	0.019**	0.016**	0.050***
	(0.010)	(0.008)	(0.008)	(0.009)
Pre-Treatment Mean	0.193	0.207	0.161	0.107
Country Controls	YES	YES	YES	YES
Influencer FE	YES	YES	YES	YES
Year-Month FE	YES	YES	YES	YES
Account Age FE	YES	YES	YES	YES
Account Age × First-Account-Year FE	YES	YES	YES	YES
Observations	65,984	65,984	65,984	65,984
R-squared	0.617	0.515	0.501	0.598

Notes: 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 4.5. The dependent variable in each regression is the number of posts that were labelled as sponsored for influencer i in month t but were not disclosed as sponsored/advertising as a share of the total number of posts made by influencer i in month t that were not disclosed as sponsored/advertising. A full list of words used to detect disclosure is in Appendix A.2. Each column uses a different ML classifier to label posts as sponsored. The classifiers use embeddings generated from post text and are trained on a sample of 300,000 German posts as discussed in Section 4.4.2. In data used for the top panel, posts are labelled as sponsored by ML classifiers only. In data used for the bottom panel, posts are only labelled as sponsored if both a ML classifier and the manual approach classifies them as sponsored. "Germany × Treated Period" is a dummy equal to one for all German influencer observations after November 2016 and zero 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 7: Influencer/Month DiD Estimates - Additional Post Content Outcomes

	(1)	(2)	(3)
Outcome:	Disclosed Share	Manual Sponsored Share	N Posts
Germany × Treated Period	0.091***	0.019**	0.974
	(0.007)	(0.009)	(0.777)
Pre-Treatment Mean	0.0509	0.452	19.29
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	67,235
R-squared	0.576	0.556	0.579

Notes: 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 4.5. "Germany × Treated Period" is a dummy equal to one for all German influencer observations after November 2016 and zero 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 8: Influencer/Month DiD Estimates - Engagement

	(1)	(2)	(3)
Outcome:	Mean N Likes	Mean N Comments	Mean N Followers
Germany × Treated Period	-483.217***	-22.663***	-4,693
	(157.693)	(7.232)	(8,275)
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

Notes: 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 4.5. "Germany × Treated Period" is a dummy equal to one for all German influencer observations after November 2016 and zero 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 8 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 fall 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%).[80] This is consistent with the intuition that an increase in sponsorship will reduce engagement, on average. Fainmesser and Galeotti (2020) and Mitchell (2021) generate theoretical predictions consistent with this result. It may indicate a decrease in average consumer welfare in Germany after disclosure regulations are introduced.[81] 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.[82]

The model in Section 5 generates predictions about the effects of regulations on follower beliefs and engagement across different post types. A key prediction is 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 does not have information about the beliefs of followers.[83] Nonetheless, we test the engagement predictions in the 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. This ratio proxies the expected engagement ratio in Section 5. If the model's predictions are correct, we should observe an increase in the ratio after the strengthening of disclosure regulations in Germany.

Estimates of these regressions are in Table 9. As in previous tables, each column represents a different classifier or combination of classifiers 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. For example, the Naive Bayes classifier in Column (1) suggests that relative to a Pre-Treatment Mean engagement ratio of 0.7, the ratio in the post regulation period is approximately 1. This is also the case for the SGD classifier in Column (2), although it starts from a higher baseline ratio. 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 receive minimal attention and engagement, at least relative to undisclosed-sponsored and non-sponsored posts. This is the case for the Kylie Jenner posts in Figure 1, 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 Naive Bayes predicted mean like ratio for sponsored and disclosed posts over non-

is changing in Germany relative to Spain after regulations.

[80] 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.

[81] As discussed in Section 5, this need not be the case.

[82] 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 A8.

[83] 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.

sponsored posts is approximately 0.5 (compared to a ratio of nearly 1 for sponsored undisclosed posts). The SGD predicted mean like ratio for sponsored and disclosed posts is 0.4 (also compared to a ratio of nearly 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 10.

We further confirm these influencer-level estimates by estimating a series of post-level difference-in-differences regressions in Appendix A.6. 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 A2 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.

7.3 Robustness

We address several concerns with our empirical approach and estimates in the Appendix and summarize them below:

- *Timing Tests (Appendix A.7)*: an important concern with difference-in-differences estimation is whether the control and treated groups are similar in their outcomes prior to treatment. We show timing tests for diverging pre-trends in Figure A2. There are no systematic statistically significant differences in the main outcomes from Table 5 between Germany and Spain prior to November 2016 in Germany.[84]

- *Classifier-free Embedding Results (Appendix A.8)*: rather than use our classifiers, we use post embeddings to directly look at the effects of stronger disclosure regulation in Germany on the distribution of language that influencers use. Similar to Ash et al. (2019), we calculate the cosine-distance of each post from a vector and test for changes in distances.[85] We show how the distributions of post-level distances change in Figures A3 and A4 and estimate difference-in-differences regressions with influencer/month mean cosine-distances as outcome variables in Table A3. We show that embeddings in Germany move on average relative to embeddings in Spain, even for undisclosed posts.

- *Predicted post-level sponsorship probabilities (Appendix A.9)*: the theoretical model in Section 5 has a probabilistic description of sponsorship/advertising. In the main results, we classify posts as sponsored/non-sponsored, but some of our ML classifiers (NB, DT and RF) generate post-level probabilistic predictions about sponsorship. We show how the distributions of these probabilities change between the pre- and post- strong regulation period in Figure A5 and estimate difference-in-differences regressions with them as outcome variables in Table A4. Our results using these outcomes are qualitatively similar to those in the main text.

- *Non-Matched Sample (Appendix A.10)*: we replicate the analysis in Tables 5-7 using the full sample of approximately 12,000 influencers rather than the matched sample of approximately

[84] Other outcomes similarly show no statistically significant differences in the pre-regulation period.

[85] Unlike Ash et al. (2019) we use an arbitrary vector (of zeroes) to measure the distance. We do this to be as agnostic as possible about possible changes in content.

Table 9: Influencer/Month DiD Estimates - Relative Engagement (Likes)

	(1)	(2)	(3)	(4)
Outcome:		$\frac{\text{Mean N Likes for Sponsored-Undisclosed Posts}}{\text{Mean N Likes for Non-Sponsored Posts}}$		
Classifier:	Naive Bayes	SGD L1	Decision Tree	Random Forest
Germany × Treated Period	0.312**	0.056*	0.011	0.109*
	(0.146)	(0.028)	(0.016)	(0.065)
Pre-Treatment Mean	0.725	0.906	0.844	0.499
Country Controls	YES	YES	YES	YES
Influencer FE	YES	YES	YES	YES
Year-Month FE	YES	YES	YES	YES
Account Age FE	YES	YES	YES	YES
Account Age × First-Account-Year FE	YES	YES	YES	YES
Observations	59,209	63,298	64,476	63,807
R-squared	0.101	0.054	0.098	0.052

	(5)	(6)	(7)	(8)
Outcome:		$\frac{\text{Mean N Likes for Sponsored-Undisclosed Posts}}{\text{Mean N Likes for Non-Sponsored Posts}}$		
Classifier:	Naive Bayes +Manual	SGD L1 +Manual	Decision Tree +Manual	Random Forest +Manual
Germany × Treated Period	0.124*	0.012	0.009	0.102*
	(0.073)	(0.016)	(0.018)	(0.058)
Pre-Treatment Mean	0.605	0.729	0.674	0.440
Country Controls	YES	YES	YES	YES
Influencer FE	YES	YES	YES	YES
Year-Month FE	YES	YES	YES	YES
Account Age FE	YES	YES	YES	YES
Account Age × First-Account-Year FE	YES	YES	YES	YES
Observations	62,671	64,722	65,042	64,513
R-squared	0.054	0.114	0.118	0.047

Notes: 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 4.5. The dependent variable in each regression is a ratio of the mean number of likes of posts that were labelled as sponsored and undisclosed for influencer i in month t over the mean number of likes of posts that were labelled as non-sponsored. A full list of words used to detect disclosure is in Appendix A.2. Each column uses a different ML classifier to label posts as sponsored. The classifiers use embeddings generated from post text and are trained on a sample of 300,000 German posts as discussed in Section 4.4.2. In data used for the top panel, posts are labelled as sponsored by ML classifiers only. In data used for the bottom panel, posts are only labelled as sponsored if both a ML classifier and the manual approach classifies them as sponsored. "Germany × Treated Period" is a dummy equal to one for all German influencer observations after November 2016 and zero 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 10: Relative Post Engagement in Germany in Treated Period

	(1)	(2)	(3)	(4)
Classifier:	Naive Bayes	SGD L1	Decision Tree	Random Forest
Average $\frac{\text{Mean N Likes Spon-Undisc.}}{\text{Mean N Likes Non-Spon.}}$	1.00	0.96	0.91	0.73
Average $\frac{\text{Mean N Likes Spon-Disc.}}{\text{Mean N Likes Non-Spon.}}$	0.54	0.42	0.40	0.39

	(5)	(6)	(7)	(8)
Classifier:	Naive Bayes	SGD L1	Decision Tree	Random Forest
	+Manual	+Manual	+Manual	+Manual
Average $\frac{\text{Mean N Likes Spon-Undisc.}}{\text{Mean N Likes Non-Spon.}}$	0.87	0.81	0.80	0.67
Average $\frac{\text{Mean N Likes Spon-Disc.}}{\text{Mean N Likes Non-Spon.}}$	0.47	0.39	0.38	0.37

Notes: 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 4.5. Each cell shows a ratio of either (i) the mean monthly number of likes of posts that were labelled as sponsored and undisclosed over the mean monthly number of likes of posts that were labelled as non-sponsored, or (ii) the mean monthly number of likes of posts that were labelled as sponsored and disclosed over the mean monthly number of likes of posts that were labelled as non-sponsored. A full list of words used to detect disclosure is in Appendix A.2. Each column uses a different ML classifier to label posts as sponsored. The classifiers use embeddings generated from post text and are trained on a sample of 300,000 German posts as discussed in Section 4.4.2. In data used for the top panel, posts are labelled as sponsored by ML classifiers only. In data used for the bottom panel, posts are only labelled as sponsored if both a ML classifier and the manual approach classifies them as sponsored.

1,200 influencers. Estimates from regressions using this sample are in Tables A6-A8 and are similar to the main results.

- *Google Trends Results (Appendix A.11)*: we 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. This helps address concerns that the prominence of debates around deceptive advertising on social media affected consumer demand and content production through other channels than what we outline above.

8 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 labelled 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 2020,Mitchell 2021), we also show that there are potentially adverse effects to such regulations. The number and percentage of sponsored posts increases at the influencer level, even for undisclosed posts. Overall, our findings suggest that in markets with no direct compensation mechanisms, regulations that distort indirect compensation mechanisms can have large and unanticipated effects in supply of sponsored content.

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 advertising may in fact increase the amount of advertising that consumers are exposed to. This is a key concern for policy-makers 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 (follower) utility in this market, our findings suggest that consumer welfare falls after regulation. We also find changes in relative engagement for undisclosed-sponsored and non-sponsored 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 the content of posts 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 paper. Nonetheless, these are open avenues for future research. A more fully specified model, together with the data used in this paper, could also allow for direct estimation of parameters governing influencer and follower behaviour. 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 (2020) and Mitchell (2021).

References

Alatas, V., Chandrasekhar, A. G., Mobius, M., Olken, B. A., and Paladines, C. (2019). When celebrities speak: A nationwide twitter experiment promoting vaccination in indonesia. Technical report, National Bureau of Economic Research.

Anagol, S., Cole, S., and Sarkar, S. (2017). Understanding the advice of commissions-motivated agents: Evidence from the indian life insurance market. *Review of Economics and Statistics*, 99(1):1–15.

Aneja, A. and Xu, G. (2020). The costs of employment segregation: Evidence from the federal government under wilson. Technical report, National Bureau of Economic Research.

Arora, S., Liang, Y., and Ma, T. (2017). A simple but tough-to-beat baseline for sentence embeddings. In *5th International Conference on Learning Representations, ICLR 2017*.

Ash, E., Chen, D. L., and Naidu, S. (2019). Ideas have consequences: The impact of law and economics on american justice. *Center for Law & Economics Working Paper Series*, 4.

Azoulay, P., Fons-Rosen, C., and Graff Zivin, J. S. (2019). Does science advance one funeral at a time? *American Economic Review*, 109(8):2889–2920.

Bennett, P. N. (2000). Assessing the calibration of naive bayes posterior estimates. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA SCHOOL OF COMPUTER SCIENCE.

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

Blum, B. S. and Goldfarb, A. (2006). Does the internet defy the law of gravity? *Journal of international economics*, 70(2):384–405.

Ducato, R. (2019). One hashtag to rule them all? mandated disclosures and design duties in influencer marketing practices. *Mandated Disclosures and Design Duties in Influencer Marketing Practices (May 21, 2019). Ranchordás, S.(ed.) & Goanta, C.(Eds), The Regulation of Social Media Influencers.*

Fainmesser, I. P. and Galeotti, A. (2020). The market for online influence. *American Economic Journal: Microeconomics*. forthcoming.

Ferreira, F. and Waldfogel, J. (2013). Pop internationalism: has half a century of world music trade displaced local culture? *The economic journal*, 123(569):634–664.

Gentzkow, M., Shapiro, J. M., and Taddy, M. (2019). Measuring group differences in high-dimensional choices: method and application to congressional speech. *Econometrica*, 87(4):1307–1340.

Goanta, C. and Ranchordas, S. (2019). The regulation of social media influencers: An introduction. *The Regulation of Social Media Influencers: An Introduction'in C. Goanta and S. Ranchordas (eds), The Regulation of Social Media Influencers (Edward Elgar, 2020, Forthcoming).*

Goanta, C. and Wildhaber, I. (2019). In the business of influence: Contractual practices and social media content monetisation. *Schweizerische Zeitschrift für Wirtschafts-und Finanzmarktrecht, SZW*, 4.

Goodman-Bacon, A. (2018). Difference-in-differences with variation in treatment timing. Technical report, National Bureau of Economic Research.

Hansen, S., McMahon, M., and Prat, A. (2018). Transparency and deliberation within the fomc: a computational linguistics approach. *The Quarterly Journal of Economics*, 133(2):801–870.

Horstmann, I. and MacDonald, G. (2003). Is advertising a signal of product quality? evidence from the compact disc player market, 1983–1992. *International Journal of Industrial Organization*, 21(3):317–345.

Iacus, S. M., King, G., and Porro, G. (2008). Matching for causal inference without balance checking. *Available at SSRN 1152391*.

Inderst, R. and Ottaviani, M. (2012). Competition through commissions and kickbacks. *American Economic Review*, 102(2):780–809.

Joulin, A., Bojanowski, P., Mikolov, T., Jégou, H., and Grave, E. (2018). Loss in translation: Learning bilingual word mapping with a retrieval criterion. *arXiv preprint arXiv:1804.07745*.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., and Mikolov, T. (2016a). Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016b). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*.

Mitchell, M. (2021). Free ad(vice): Internet influencers and disclosure regulation. *RAND Journal of Economics*, 52(1):3–21.

Monti, S. and Cooper, G. F. (2013). A bayesian network classifier that combines a finite mixture model and a naive bayes model. *arXiv preprint arXiv:1301.6723*.

Müller, K. and Schwarz, C. (2019). From hashtag to hate crime: Twitter and anti-minority sentiment. *Available here: https://ssrn. com/abstract*, 3149103.

Pei, A. and Mayzlin, D. (2019). Influencing the influencers. *Available at SSRN 3376904*.

Sahni, N. S. and Nair, H. S. (2020). Does advertising serve as a signal? evidence from a field experiment in mobile search. *The Review of Economic Studies*, 87(3):1529–1564.

Sarsons, H. (2019). Interpreting signals in the labor market: Evidence from medical referrals.

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

Sun, L. and Abraham, S. (2020). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*.

A Appendix

A.1 Other European Advertising Regulations

In Italy, regulation of influencer behaviour followed two parallel tracks. The first is self-regulation by the Italian Advertising Self Regulatory Institute (IAP). In 2016 the IAP produced a set of non-binding recommendations about social media influencer conduct. Those broadly reflected disclosure guidelines promoted by EASA. In 2018, a popular Italian Instagram influencer and a car manufacturer were forced by the IAP to remove content for improper labelling of sponsorship (lexology.com). There was also a second more formal track by the government and competition authorities. This was mainly conducted in 2017 and 2018 and appears to be related to government pressure to reduce tax evasion (Osborne-Clarke.com). In 2017 the Italian competition authorities, the ACGM, began investigating influencer conduct as a form of deceptive advertising. In late 2017, they issued a series of public letters to prominent influencers inviting them to clearly disclose any sponsored content. Similar letters were sent to brands which advertised with the influencers. In June 2017, the Italian parliament passed a bill to regulate the behaviour of online influencers requiring them to clearly identified sponsored tags to any posts that relate to products influencers were paid to promote or received for free (Osborne-Clarke.com). By 2019, recommendations of the IAP were folded into official government regulations.

French regulations were similar to Italy. France's self regulatory advertising association (ARPP) instituted mandatory disclosure recommendations for influencers at the end of 2017 (ARPP.org, InternationalLawOffice.com). Misleading advertising also falls under French consumer protection law and is punishable by up to 300k EUR fines for the influencer and up to 1.5 million EUR fine for the brand (Osborne-Clarke.com). Content is considered to be sponsored even in cases when there is no direct financial compensation: for example, whenever a brand (or someone other than the influencer) has editorial control over the post, or whenever some compensation is provided (i.e., free product). However, no influencers have been fined yet and there is no evidence that fines were handed out.[86]

A.2 Manual Keywords Used For "Disclosed"

- #*ad*, #*paid*, #werbung, #anzeige, #anuncio
- sponsor[...], promo[...]
- collab[...], partner[...]
- ambassad[...], publici[...], adver[...]
- patrocin[...]
- #gift[...], and its translations

[86] In March 2019, the French government began a case against two influencers who advertised "dropshipping" websites which sell fake products (i.e., fake watches) and re-sell products from Amazon for inflated prices. The case is due for a hearing in 2020 (ladn.eu).

A.3 Manual Keywords Used For "Sponsored"

We assume that the following sets of words denote sponsorship (each set has been translated):

- Links to websites (".com", ".de", etc)
- References to availability
- References to discount codes, contests, and "%" off
- References to campaigns, mentions of "official," "new," "tonight"
- References to shopping
- References to "launching"
- Various versions of "Thank you," or "Thanks"
- Instructions for users to "follow"
- Brand names: we use a list of around 1,000 brand names

A.4 Illustration of Text Processing

In this section we illustrate how our text processing and multi-lingual text embeddings work. We use two sample sentences, one in Spanish and one in German. To illustrate the difficulty we face with the real data, we include English words in both German and Spanish sentences:

1. "Ich liebe meine kinder sehr! #Loving life!" (I love my kids very much! #Loving life!)
2. "Qué gran oferta! Available now - link in my bio!" (What a great deal! Available now - link in my bio!)

As described in Section 4.4, we first process the sentences and transform them into a list of words/tokens. We remove all "stopwords" in German, English and Spanish. We also remove all hashtag signs, punctuation and drop emojis. Last, we take out all words that are less than 3 characters long or more than 20 characters long. Words from the processed sentences are below:

1. kinder, liebe, life, loving
2. available, gran, link, oferta

Next we check to see the language of each word. We apply FastText's language detection algorithm. For the words in our text, the detected languages are:

1. kinder = German, liebe = German, life = English, loving = English
2. available= English, gran= Spanish, link= English, oferta= Spanish

With these detected languages, we convert the words into embeddings using the relevant embedding dictionary for each language. We then calculate a simple average of the embeddings of each word to obtain a post-level embeddings.

Table A1: Post-Level Agreement % and Cohen's κ

Sample: Disclosed Posts

	Gaussian Naive Bayes	SGD w/ L1 Loss	Decision Tree	Random Forest
Manual	Agree =75.54% κ=0.3291	Agree= 64.29% κ=0.1356	Agree=62.34% κ= 0.1690	Agree= 62.86% κ=0.2215
Gaussian Naive Bayes		Agree=63.79% κ=0.1758	Agree=68.26% κ=0.3208	Agree=75.13% κ=0.4861
SGD w/L1 Loss			Agree=67.29% κ=0.3173	Agree=71.65% κ=0.4210
Decision Tree				Agree=76.73% κ=0.5294

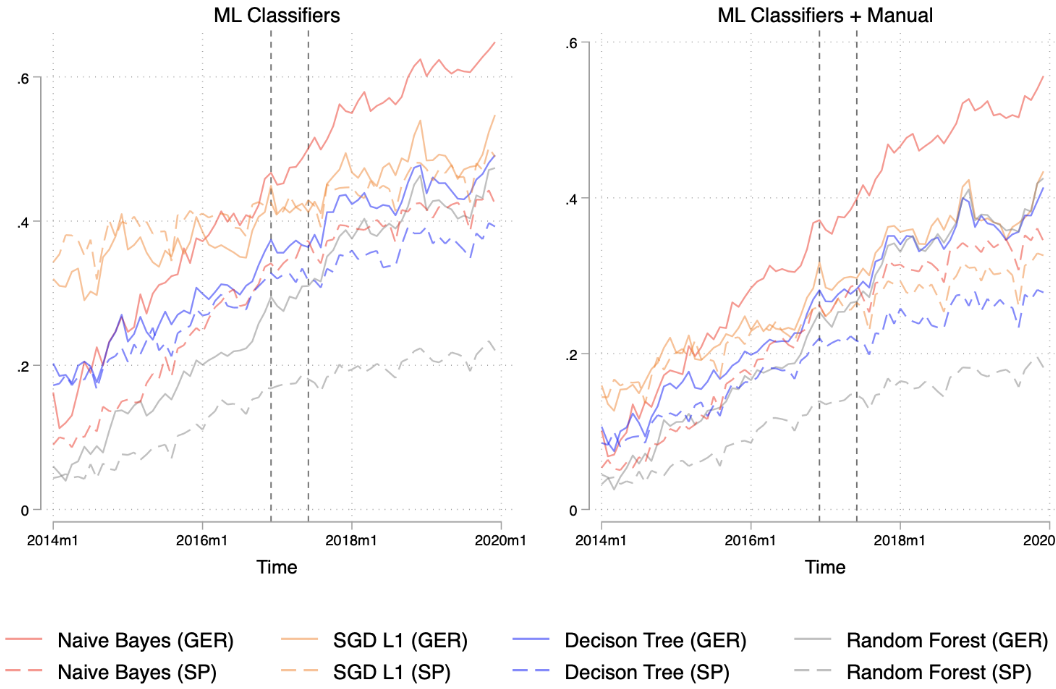
Sample: Undisclosed Posts

	Gaussian Naive Bayes	SGD w/ L1 Loss	Decision Tree	Random Forest
Manual	Agree=67.73% κ= 0.3588	Agree=58.60% κ=0.1760	Agree=60.53% κ=0.2179	Agree=60.53% κ=0.2222
Gaussian Naive Bayes		Agree=60.64% κ=0.1592	Agree=70.15% κ=0.3205	Agree=77.66% κ=0.4462
SGD w/L1 Loss			Agree=66.81% κ=0.2700	Agree=69.92% κ=0.2972
Decision Tree				Agree=79.68% κ=0.4397

Notes: This table shows averages of pairwise classifier agreement % and Cohen's κ for all posts made by CEM-matched influencers. Matching is described in Section 4.5. ML classifiers use embeddings generated from post text and are trained on a sample of 300,000 German posts as discussed in Section 4.4.2. Top panel looks only at posts that are disclosed as sponsored. Bottom panel looks only at posts that are not disclosed as sponsored. A full list of words used to detect disclosure is in Appendix A.2.

A.5 Additional Data Description

Figure A1: Predicted Sponsored Post Shares for Germany and Spain



Notes: Each dashed line represents the average share of predicted sponsored posts of the total number of posts ($\frac{\text{N Predicted Sponsored Posts}}{\text{N Posts}}$) in Germany or Spain in month t according to one of four ML classifiers (Gaussian Naive Bayes, SGD L1, Decision Tree, Random Forest). CEM matched sample of influencers used. In data used to generate left-hand panel figure, posts are labelled as sponsored by ML classifiers only. In data used to generate right-hand panel figure, posts are only labelled as sponsored if both a ML classifier and the manual approach classifies them as sponsored. The first dashed vertical line represents the initial changes to German disclosure regulations (see Section 3.2). The second dashed vertical line represents the first fines handed out to German influencers in mid 2017.

A.6 Post Level Results

In addition to influencer level regressions we also estimate post-level regressions. We model outcome Y_{pit} for post p created by influencer i at time t as:

$$Y_{pit} = \alpha \left(\text{Germany}_i \times \text{Treated Period}_t\right) + \beta X_{pit} + \delta_i + \delta_t + \epsilon_{pit} \tag{3}$$

where X_{pit} country/influencer/post specific characteristics. Post level characteristics include whether the post is an image, a video, or a photo album, the length of the post's text and the day of week on which the post was made. Using this regression, we can compare similar posts made by the same influencers over time and see how their outcomes change after regulations are introduced in Germany (relative to Spain).

Our model predicts that after regulations, followers are going to "trust" or engage with sponsored but not disclosed posts more **relative** to non-sponsored posts.[87] We want to test for this prediction by using the number of post likes as an outcome measure, imagining likes as a measure of engagement with the post, which is a common interpretation in the industry.

In order to compare posts by similar influencers before and after regulations, we introduce influencer popularity as a control. We measure influencer popularity in month t as the mean number of likes the influencer's posts receive (excluding post p).[88] This allows us to account for overall changes in influencer popularity, which are in and of themselves affected by regulations (Table 7). We capture the effects of regulations on the popularity of different types of posts conditional on how popular the influencers are.

Table A2 shows estimates from post-level regressions. We segment the sample of posts by disclosure and sponsorship, separately estimating Equation (3) for non-sponsored posts and undisclosed-sponsored posts. We also separately use each one of our ML classifiers (and the combination of ML and manual classifiers) to denote sponsorship. Estimates from these regressions show that the average number of likes for non-sponsored posts falls in Germany after regulations relative to Spain. However, there is no similar change in likes for undisclosed sponsored posts.[89] These results are consistent with the predictions of the model about followers are less engaged with non-sponsored posts relative to sponsored posts after disclosure regulations.

[87] In absolute terms, non-sponsored posts are still likely to generate more engagement and be trusted more than sponsored and undisclosed posts.

[88] The most natural measure of popularity is the number of followers at the time the post was made. Unfortunately, this data is fragmentary. We do not wish to extrapolate it since that would require making substantial assumptions. The average number of likes an influencer's posts receive and the number of followers is highly correlated in the data where the number of followers is available.

[89] We also estimate a similar regression for disclosed sponsored posts and also find no changes in Germany relative to Spain after regulatory changes. However, there is a very small number of disclosed sponsored posts in Germany prior to any regulatory changes and in Spain throughout our sample period. The disclosed sponsored posts that exist in these settings are likely substantial outliers and it is not clear how to think about disclosure there.

Table A2: Post Level DiD Estimates - Post Likes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Outcome:					N Post Likes			
Classifier:	Naive Bayes	SGD L1	Decision Tree	Random Forest	Naive Bayes +Manual	SGD L1 +Manual	Decision Tree +Manual	Random Forest +Manual
					Sample:Non-Sponsored Posts			
Germany × Treated Period	-353.220	-367.952*	-294.034	-316.333*	-323.048	-334.989*	-297.014	-309.726*
	(229.062)	(205.191)	(190.275)	(190.632)	(204.220)	(195.070)	(186.331)	(186.599)
Observations	701,776	671,899	780,329	918,696	811,943	847,031	894,883	960,266
R-squared	0.629	0.643	0.624	0.633	0.631	0.638	0.627	0.633
					Sample: Sponsored & Undisclosed Posts			
Germany × Treated Period	-149.441	-188.643	-233.825	-53.672	-121.555	-146.018	-177.992	-52.247
	(146.381)	(148.517)	(155.136)	(173.475)	(159.710)	(150.880)	(160.886)	(192.944)
Observations	465,272	495,159	386,725	248,331	355,104	320,022	272,152	206,754
R-squared	0.687	0.630	0.671	0.689	0.690	0.645	0.689	0.691
Country Controls	YES	YES	YES	YES	YES	YES	YES	YES
Post Type Controls	YES	YES	YES	YES	YES	YES	YES	YES
Influencer Popularity Controls	YES	YES	YES	YES	YES	YES	YES	YES
Influencer FE	YES	YES	YES	YES	YES	YES	YES	YES
Year-Month FE	YES	YES	YES	YES	YES	YES	YES	YES
Account Age FE	YES	YES	YES	YES	YES	YES	YES	YES
Account Age × First-Account-Year FE	YES	YES	YES	YES	YES	YES	YES	YES

Notes: Sample includes post level observations from January 2014 to December 2019. Posts in the sample belong to CEM-matched influencers. Matching is described in Section 4.5. The dependent variable in each regression is the number of likes for the post. The top panel includes posts that were not disclosed as sponsored or labelled as sponsored by the different classifiers. The bottom panel includes posts that were not disclosed as sponsored but were labelled as sponsored by a classifier. A full list of words used to detect disclosure is in Appendix A.2. Each column uses a different ML classifier to label posts as sponsored. The classifiers use embeddings generated from post text and are trained on a sample of 300,000 German posts as discussed in Section 4.4.2. In Columns (1)-(4), posts are labelled as sponsored by ML classifiers only. In Columns (5)-(8), posts are only labelled as sponsored if both a ML classifier and the manual approach classifies them as sponsored. "Germany × Treated Period" is a dummy equal to one for all German posts after November 2016 and zero 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. Regressions also include post type fixed effects, day of week fixed effects and influencer popularity controls. Standard errors are clustered at the influencer level. *** p<0.01, ** p<0.05, * p<0.1.

A.7 Timing Tests

Formal timing tests help us ensure that average outcomes between the control and treated groups are not statistically significantly different prior to treatment taking place. In our context, we test whether sponsorship rates between German and Spanish influencers are similar prior to the enactment of regulations in Germany. Formal timing tests also help us understand the timing of treatment effects. In our baseline regressions we "activate" treatment in November 2016, but we do not necessarily expect the effects to be immediate. Even the "clarified" regulations could be somewhat ambiguous, and it may not be immediately clear to influencers the extent to which they will be enforced. It may take time for influencers to understand that the regulations are binding. As mentioned in Section 3.2, the first high profile fines for German influencers for not following regulations did not come until the middle of 2017. We may expect to see stronger effects after that period.

We estimate the following regression for influencer i at month t:

$$Y_{it} = \sum_{t=\tau-10}^{t=T} \alpha_t \left(\text{Germany}_i \times D_t\right) + \beta X_{it} + \delta_i + \delta_t + \epsilon_{it} \tag{4}$$

where D_t is a dummy equal to 1 in period t and where the initial regulatory change occurs at time $t = \tau$ for Germany. We allow for heterogeneous treatment effects (αs) to occur from that time period until the last period in our sample ($t = T$). We also allow for placebo heterogeneous treatment effects for 10 periods before regulation takes place.[90] In many applications estimating this dynamic difference-in-differences regression is challenging. If different observations are treated at different points in time, the lead and lag coefficients do not correctly recover treatment effects (Goodman-Bacon 2018, Sun and Abraham 2020). This is not a concern in our setting, where all influencers in Germany are subject to the same regulations that come in at the same time and all influencers in Spain are not subject to regulations.

We show estimates from these regressions for the main outcomes from Table 5 in Figure A2.[91] They confirm that there are no statistically significant differences in sponsorship shares between Germany and Spain in the 10 months before the initial changes in the German regulatory regime. This is true for all classifiers. We also find that treatment effects intensify after the first fines are levied on German influencers in mid-2017.

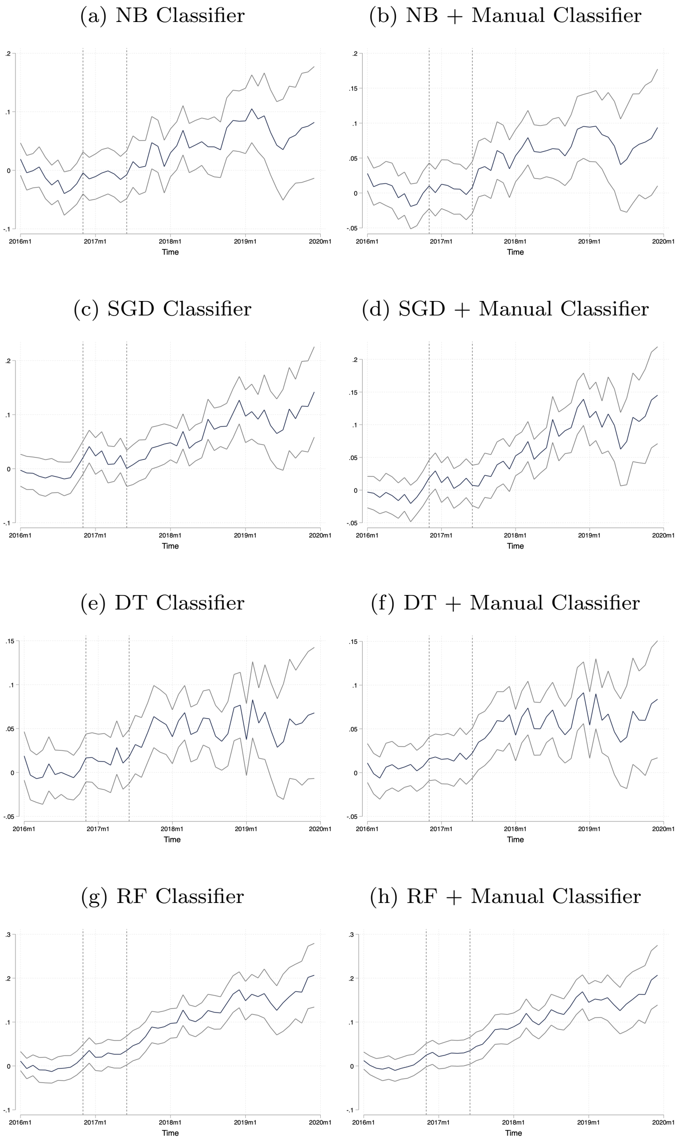
A.8 Embeddings Estimates

Estimates in Tables 5 and 6 show that the amount of sponsored content increases in Germany after regulations. A possible concern with these results is that they depend on the various classifiers (ML or manual) to tell us what is a sponsored post. Differences in classifier quality over time or across countries could affect estimates. To test whether our estimates are driven by underlying changes in post content, we look directly at post embeddings. These are the 300 dimensional vectors representing the average location of a post in "word meaning" space. For each post, we calculate the cosine-distance of the embedding from an arbitrary fixed location (a 300 dimensional vector of

[90] Our estimates are robust to different numbers of lead periods.

[91] Figures for other outcomes show similar patterns.

Figure A2: Timing Tests for Predicted Sponsored Share Outcomes in Table 5



(a) NB Classifier

(b) NB + Manual Classifier

(c) SGD Classifier

(d) SGD + Manual Classifier

(e) DT Classifier

(f) DT + Manual Classifier

(g) RF Classifier

(h) RF + Manual Classifier

Notes: Each panel shows monthly treatment effects estimates for Germany relative to Spain for a separate regression. In each case, influencer/month share of sponsored posts is the outcome variable. 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 4.5. All regressions include influencer and time fixed effects, as well as account age fixed effects and account age × first-account year fixed effects. Additional 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. 95% confidence interval around estimates pictured. The first dashed vertical line represents the initial changes to German disclosure regulations (see Section 3.2). The second dashed vertical line represents the first fines handed out to German influencers in mid 2017. In data used to generate left-hand panel figure, posts are labelled as sponsored by ML classifiers only. In data used to generate right-hand panel figure, posts are only labelled as sponsored if both a ML classifier and the manual approach classifies them as sponsored. NB refers to a Gaussian Naive Bayes classifier. SGD refers to a Stochastic Gradient Descent classifier with L1 loss. DT refers to a Decision Tree classifier. RF refers to a Random Forest classifier.

zeroes).[92] This is similar to the approach of Ash et al. (2019).[93] We show the distribution of post-level cosine distances before German regulations and after German regulations for both Germany and Spain in Figure A3. We do this for all posts, and also only for posts that were not disclosed as advertising. Visually, cosine-distance changes more for the German posts than for the Spanish posts over time.[94]

Analogous to Figure 3 in the main text, we look at the difference between average country-level cosine distance in Germany and average country-level cosine distance in Spain around the initial introduction of stronger regulatory scrutiny in Germany. Figure A4 shows this time series. It suggests that mean cosine distance in Germany increases relative to Spain after the regulations. Finally, we replicate the regressions in Tables 5 and 6 with mean influencer-month cosine distance from 0 as the dependent variable. We look at both the average distance for *all* influencer posts in a month (Column 1), and the average distance for *undisclosed* influencer posts in a month (Column 2). These regressions all include influencer and time fixed effects, and additional country/influencer time-varying controls. We find that there is a statistically significant change in post content in Germany relative to Spain after changes to German disclosure regulations. This is the case even for posts that are not disclosed as advertising (at the 90% percentile confidence level).[95]

As mentioned previously, the main advantage of looking directly at embeddings is that they are as close as possible to looking at the "raw data." They do not require any additional steps such as choosing words that indicate sponsorship or training ML classifiers. The main disadvantage is that both the magnitude and the direction of the changes do not have obvious interpretation. The measure of cosine-distance may also hide other changes in the high-dimensional vectors that could also reflect changes in content (i.e., one dimension moving closer to zero, while another dimension moving further away). Nonetheless, the fact that we find average statistically significant changes in these high-dimensional objects after regulation in Germany is meaningful and confirms the more easily interpretable results in the main text.

A.9 Predicted Sponsorship Probabilities

The model in Section 5 allows influencers to choose between two types of posts (sponsored/non-sponsored) but words used in the two types of posts may overlap. This means that a post may have some probability of being sponsored conditional on the words used. This is reflected in the probabilistic beliefs of consumers/followers in the model. A consumer has a posterior probability that she is looking at a sponsored post conditional on the words in that post.

Estimates in Tables 5 and 6 rely on a discrete classification of posts into sponsored/non-sponsored labels. Three of the ML classifiers we use (NB, DT and RF) are probabilistic and use conditional probabilities to classify the posts: if a post has a higher than 50% probability of being sponsored conditional on the words used, it is labelled as sponsored.[96] Discrete classification

[92] The cosine distance between two vectors A and B is equal to $1 - \frac{A \cdot B}{||A|| \times ||B||}$.

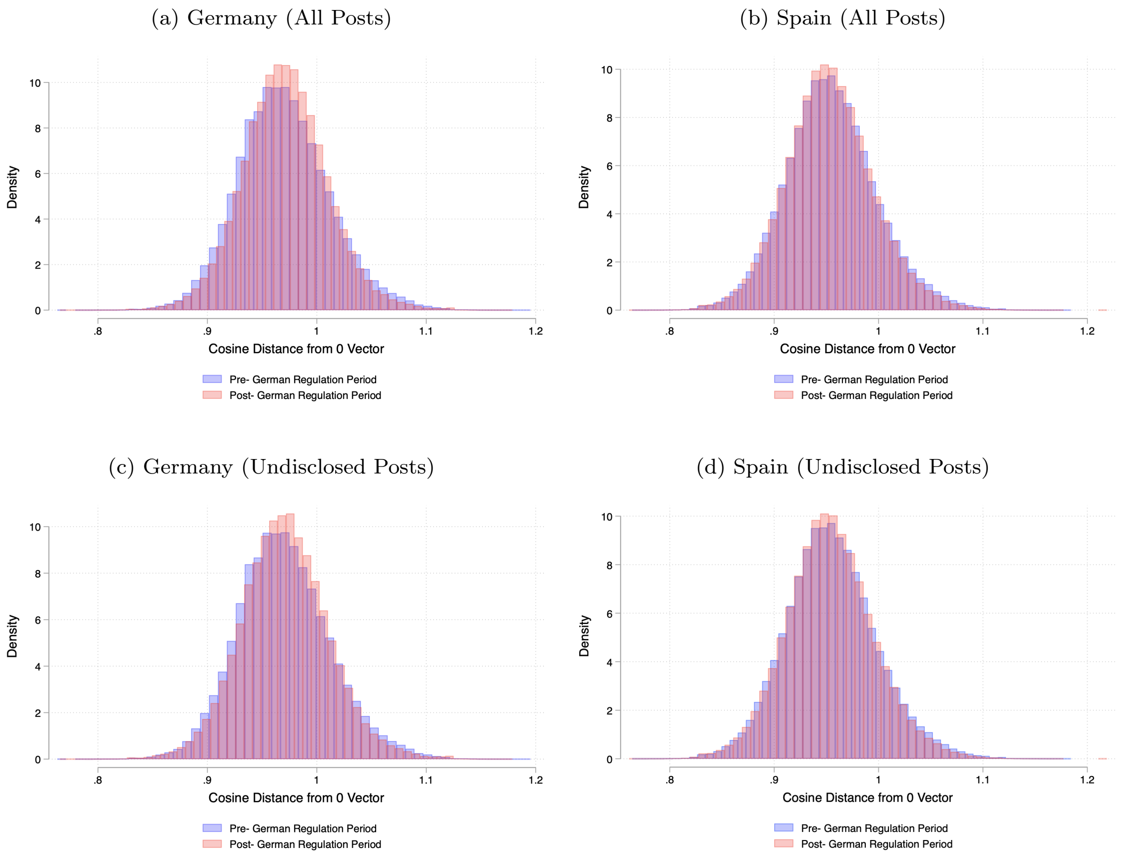
[93] In that paper, they measure the cosine distance from a particular vector with some pre-defined meaning. We use an arbitrary vector to avoid any subjective choices in assigning specific meaning to different parts of the embedding space.

[94] Formal tests of distribution equivalences (KS and Mann-Whitney) show that both the German and Spanish distributions change between the two periods. This is likely due to the sensitivity of both tests to large sample sizes. Nonetheless, both tests suggest that German posts change more - i.e., have larger test statistics.

[95] Formal timing tests confirm that there are no statistically significant average differences in cosine distance prior to changes in regulations in Germany.

[96] The fourth classifier, Stochastic Gradient Descent (SGD), is not probabilistic. Instead it builds a series of hyper-planes to separate a "sponsored" space of words from a "non-sponsored" space of words and classifies posts belonging

Figure A3: PDFs of Post-Level Embedding Cosine Distances from 0


(a) Germany (All Posts)

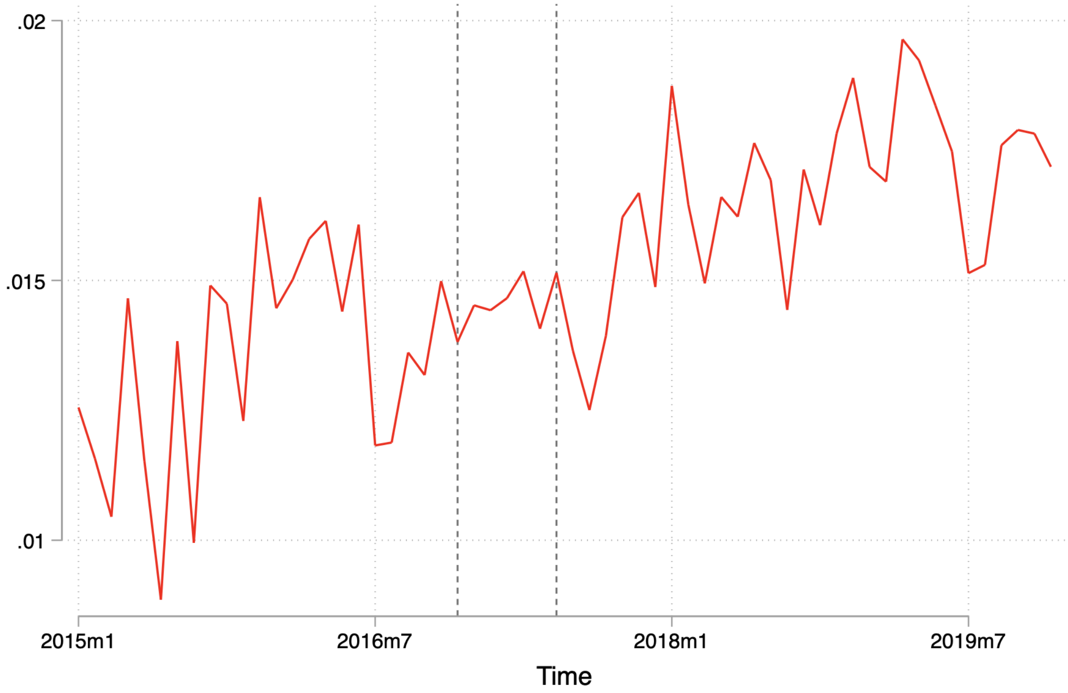
(b) Spain (All Posts)

(c) Germany (Undisclosed Posts)

(d) Spain (Undisclosed Posts)

Notes: Each panel shows the distribution of post-level embedding cosine distances from a vector of 0. Blue distributions show posts made before the period of intensified regulatory scrutiny in Germany (pre Nov 2016). Red distributions show posts made during the period of intense regulatory scrutiny in Germany (post Nov 2016). We only include posts from the sample of CEM-matched influencers. The matching is described in Section 4.5. Embeddings are generated from post text as described in Section 4.4.2. Panels (a) and (b) include all posts made by the sample of influencers. Panels (c) and (d) only include posts that are not disclosed as sponsored. A full list of words used to detect disclosure is in Appendix A.2.

Figure A4: Differences in Cosine Distance from 0 Between Germany and Spain



Notes: Figure shows a time series of differences in country-level cosine-distances from a vector of zeroes. In each month and country, we calculate the cosine distance for each post from a vector of 0 and take a simple average. We then subtract the Spanish average from the German average for each month. We only include posts from the sample of CEM-matched influencers. The matching is described in Section 4.5. Embeddings are generated from post text as described in Section 4.4.2. The first dashed vertical line represents the initial changes to German disclosure regulations (see Section 3.2). The second dashed vertical line represents the first fines handed out to German influencers in mid 2017.

Table A3: Influencer/Month DiD Estimates - Cosine-Distance from 0

	(1)	(2)
Sample:	All Posts	Undisclosed Posts
Outcome:	Mean Post	Cosine-Distance from 0
Germany × Treated Period	0.00155**	0.00119*
	(0.000670)	(0.000674)
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	67,084	65,815
R-squared	0.441	0.417

Notes: 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 4.4.2. Outcome in Column (1) is the mean cosine distance from 0 for post embeddings for all posts by influencer i in month t. Outcome in Column (2) is the mean cosine distance from 0 for post embeddings for all posts by influencer i in month t that are not disclosed as advertising/sponsored. A full list of words used to detect disclosure is in Appendix A.2. variable "Germany × Treated Period" is a dummy equal to one for all German influencer observations after November 2016 and zero 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.

has some benefits - for example, allowing us to combine ML and manual classifiers. But the probabilistic classification is closer to the theoretical model and could potentially allow us to uncover more subtle effects. Influencers may change their content such that all of their posts are more likely to be sponsored without necessarily "flipping" many posts from a non-sponsored to sponsored label.

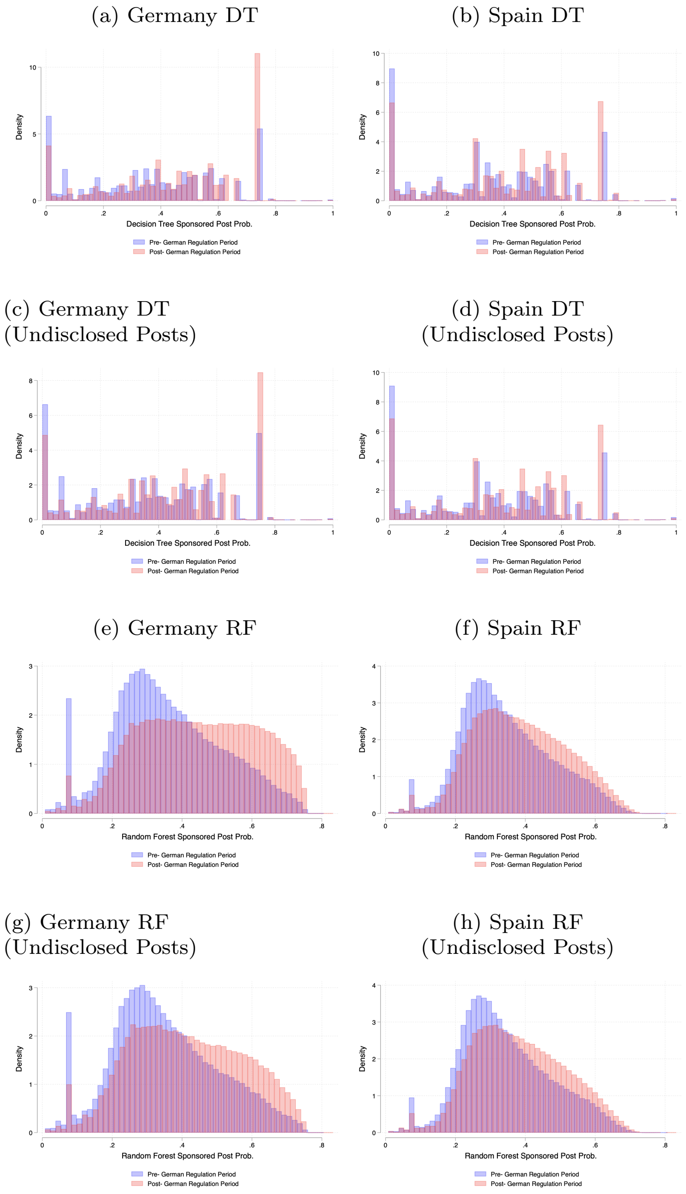
We show the probability distributions of post-level predicted sponsorship probabilities for two of our three probabilistic classifiers, DT and RF, in Figure A5.[97] We separately show the distributions for Germany and for Spain, before and after disclosure regulations in Germany. We also separately show the distributions for all posts and for only posts that were not disclosed as advertising. Figure A5 suggests that after disclosure regulations were introduced in Germany, German post-level probabilities moved to the right much more so than Spanish post-level probabilities. This consistent across both DT and RF classifiers and also holds when we look at only undisclosed posts.

Table A4 shows results from influencer/month level regressions with average predicted sponsorship probabilities as the outcome variables. Estimates from these regressions are similar qualitatively and quantitatively to the results in Tables 5 and 6. They show that there is a statistically significant increase in sponsorship probability in Germany relative to Spain after advertising disclosure regulations intensify in Germany. Even posts that are not disclosed as sponsored become more likely to be sponsored.

to the "sponsored" space as sponsored.

[97] The NB classifier produces post-level predicted posterior probabilities that are either very close to zero or very close to one. This is an issue commonly referenced in ML literature, which suggests that although NB is an accurate classifier it does not produce useful posterior probabilities (Bennett 2000,Monti and Cooper 2013).

Figure A5: PDFs of Post-Level Predicted Probabilities of Sponsorship



(a) Germany DT

(b) Spain DT

(c) Germany DT (Undisclosed Posts)

(d) Spain DT (Undisclosed Posts)

(e) Germany RF

(f) Spain RF

(g) Germany RF (Undisclosed Posts)

(h) Spain RF (Undisclosed Posts)

Notes: Each panel shows the distribution of post-level predicted sponsorship probabilities. Blue distributions show posts made before the period of regulatory scrutiny in Germany (pre Nov 2016). Red distributions show posts made after the period of regulatory scrutiny in Germany (post Nov 2016). We only include posts from the sample of CEM-matched influencers. The matching is described in Section 4.5. Panels (a)-(d) show predicted probabilities from a Decision Tree (DT) classifier. Panels (e)-(h) show predicted probabilities from a Random Forest (RF) classifier. The classifiers use embeddings generated from post text and are trained on a sample of 300,000 German posts as discussed in Section 4.4.2. Panels (a), (b), (e) and (f) include all posts made by the sample of influencers. Panels (c), (d), (g) and (h) only include posts that are not disclosed as sponsored. A full list of words used to detect disclosure is in Appendix A.2.

Table A4: Influencer-Month DiD Estimates - Sponsorship Probabilities

	(1)	(2)	(3)
Sample:		All Posts	
Outcome:	Mean Naive Bayes Prob.	Mean Decision Tree Prob.	Mean Random Forest Prob.
Germany × Treated Period	0.035***	0.022***	0.028***
	(0.011)	(0.004)	(0.003)
Pre-Treatment Mean	0.282	0.343	0.340
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	67,127	67,127	67,127
R-squared	0.668	0.586	0.724
	(4)	(5)	(6)
Sample:		Undisclosed Posts	
Outcome:	Mean Naive Bayes Prob.	Mean Decision Tree Prob.	Mean Random Forest Prob.
Germany × Treated Period	0.022*	0.013***	0.019***
	(0.012)	(0.004)	(0.003)
Pre-Treatment Mean	0.273	0.339	0.337
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	65,855	65,855	65,855
R-squared	0.642	0.538	0.687

Notes: 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 4.5. The dependent variable in each regression is the average probability that an influencer i's posts in month t are sponsored. Each column uses a different ML classifier to generate the predicted probabilities. The classifiers use embeddings generated from post text and are trained on a sample of 300,000 German posts as discussed in Section 4.4.2. In the top panel, we calculate the average sponsorship probability for all posts by the influencer. In the bottom panel we calculate the average sponsorship probability only for posts that are not disclosed as sponsored. A full list of words used to detect disclosure is in Appendix A.2. "Germany × Treated Period" is a dummy equal to one for all German influencer observations after November 2016 and zero 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.

A.10 Non-Matched Sample

Table A5: Summary Statistics for Non-Matched Sample

Variable	Obs	Mean	Std. Dev.
Mean Likes per Post	607,219	1,881	45,599
Mean Comments per Post	607,219	46	206
N Followers	137,183	92,087	92,750
N Posts per Month	607,219	22.634	28
Account Age (months)	607,219	38	23
First Account Year	607,219	2014	2

Table A6: Non-Matched Influencer/Month DiD Estimates - Sponsored Share

	(1)	(2)	(3)	(4)
Outcome:		Predicted Sponsored Shares		
Classifier:	Naive Bayes	SGD L1	Decision Tree	Random Forest
Germany × Treated Period	0.020***	0.034***	0.026***	0.060***
	(0.004)	(0.003)	(0.003)	(0.003)
Pre-Treatment Mean	0.258	0.365	0.244	0.118
Country Controls	YES	YES	YES	YES
Influencer FE	YES	YES	YES	YES
Year-Month FE	YES	YES	YES	YES
Account Age FE	YES	YES	YES	YES
Account Age × First-Account-Year FE	YES	YES	YES	YES
Observations	607,219	607,219	607,219	607,219
R-squared	0.678	0.539	0.548	0.704

	(5)	(6)	(7)	(8)
Outcome:		Predicted Sponsored Shares		
Classifier:	Naive Bayes +Manual	SGD L1 +Manual	Decision Tree +Manual	Random Forest +Manual
Germany × Treated Period	0.027***	0.034***	0.031***	0.058***
	(0.004)	(0.003)	(0.003)	(0.003)
Pre-Treatment Mean	0.187	0.203	0.155	0.0961
Country Controls	YES	YES	YES	YES
Influencer FE	YES	YES	YES	YES
Year-Month FE	YES	YES	YES	YES
Account Age FE	YES	YES	YES	YES
Account Age × First-Account-Year FE	YES	YES	YES	YES
Observations	607,219	607,219	607,219	607,219
R-squared	0.669	0.594	0.589	0.690

Notes: Sample includes influencer/month level observations from January 2014 to December 2019 with at least two posts in a month. The dependent variable in each regression is the number of posts that were labelled as sponsored for influencer i in month t as a share of the total number of posts made by influencer i in month t. Each column uses a different ML classifier to label posts as sponsored. The classifiers use embeddings generated from post text and are trained on a sample of 300,000 German posts as discussed in Section 4.4.2. In data used for the top panel, posts are labelled as sponsored by ML classifiers only. In data used for the bottom panel, posts are only labelled as sponsored if both a ML classifier and the manual approach classifies them as sponsored. "Germany × Treated Period" is a dummy equal to one for all German influencer observations after November 2016 and zero 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 A7: Non-Matched Influencer/Month DiD Estimates - (Sponsored Share | Non-Disclosure)

	(1)	(2)	(3)	(4)
Outcome:	(Predicted Sponsored Shares | Non-Disclosure)			
Classifier:	Naive Bayes	SGD L1	Decision Tree	Random Forest
Germany × Treated Period	0.010**	0.019***	0.011***	0.042***
	(0.004)	(0.003)	(0.003)	(0.003)
Pre-Treatment Mean	0.252	0.358	0.239	0.112
Country Controls	YES	YES	YES	YES
Influencer FE	YES	YES	YES	YES
Year-Month FE	YES	YES	YES	YES
Account Age FE	YES	YES	YES	YES
Account Age × First-Account-Year FE	YES	YES	YES	YES
Observations	596,869	596,869	596,869	596,869
R-squared	0.654	0.494	0.489	0.660

	(5)	(6)	(7)	(8)
Outcome:	(Predicted Sponsored Shares | Non-Disclosure)			
Classifier:	Naive Bayes	SGD L1	Decision Tree	Random Forest
	+Manual	+Manual	+Manual	+Manual
Germany × Treated Period	0.014***	0.016***	0.014***	0.040***
	(0.004)	(0.003)	(0.003)	(0.003)
Pre-Treatment Mean	0.181	0.197	0.149	0.0910
Country Controls	YES	YES	YES	YES
Influencer FE	YES	YES	YES	YES
Year-Month FE	YES	YES	YES	YES
Account Age FE	YES	YES	YES	YES
Account Age × First-Account-Year FE	YES	YES	YES	YES
Observations	596,869	596,869	596,869	596,869
R-squared	0.636	0.544	0.530	0.640

Notes: Sample includes influencer/month level observations from January 2014 to December 2019 with at least two posts in a month. The dependent variable in each regression is the number of posts that were labelled as sponsored for influencer i in month t but were not disclosed as sponsored/advertising as a share of the total number of posts made by influencer i in month t that were not disclosed as sponsored/advertising. A full list of words used to detect disclosure is in Appendix A.2. Each column uses a different ML classifier to label posts as sponsored. The classifiers use embeddings generated from post text and are trained on a sample of 300,000 German posts as discussed in Section 4.4.2. In data used for the top panel, posts are labelled as sponsored by ML classifiers only. In data used for the bottom panel, posts are only labelled as sponsored if both a ML classifier and the manual approach classifies them as sponsored. "Germany × Treated Period" is a dummy equal to one for all German influencer observations after November 2016 and zero 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 A8: Non-Matched Influencer/Month DiD Estimates - Additional Outcomes

	(1)	(2)	(3)	(4)	(5)	(6)
Outcome:	Disclosed Share	Manual Sponsored Share	Mean N Likes	Mean N Comments	Mean N Followers	N Posts
Germany × Treated Period	0.069***	0.017***	58.251	-3.182*	-8,491.299***	2.029***
	(0.003)	(0.004)	(55.657)	(1.899)	(2,933.124)	(0.457)
Pre-Treatment Mean	0.0398	0.448	831	18.41	90,218	24.87
Country Controls	YES	YES	YES	YES	YES	YES
Influencer FE	YES	YES	YES	YES	YES	YES
Year-Month FE	YES	YES	YES	YES	YES	YES
Account Age FE	YES	YES	YES	YES	YES	YES
Account Age × First-Account-Year FE	YES	YES	YES	YES	YES	YES
Observations	607,219	607,219	607,219	607,219	137,091	607,219
R-squared	0.578	0.572	0.691	0.260	0.907	0.616

Notes: 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 4.5. "Germany × Treated Period" is a dummy equal to one for all German influencer observations after November 2016 and zero 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.

A.11 Google Trends

A possible concern with our research design is that debates surrounding disclosure regulations can affect consumer preferences for Instagram and change consumer demand and content production. For example, prominent critical discussions of hidden and deceptive advertising could reduce the willingness of consumers to spend time on social media, independently of the actual amount of advertising in the market. This would mean German regulations could affect the number or composition of German Instagram followers relative to Spain. The changing composition of German followers could in turn affect the content production of German influencers: an additional channel which is different than the channel we outline in the main text.

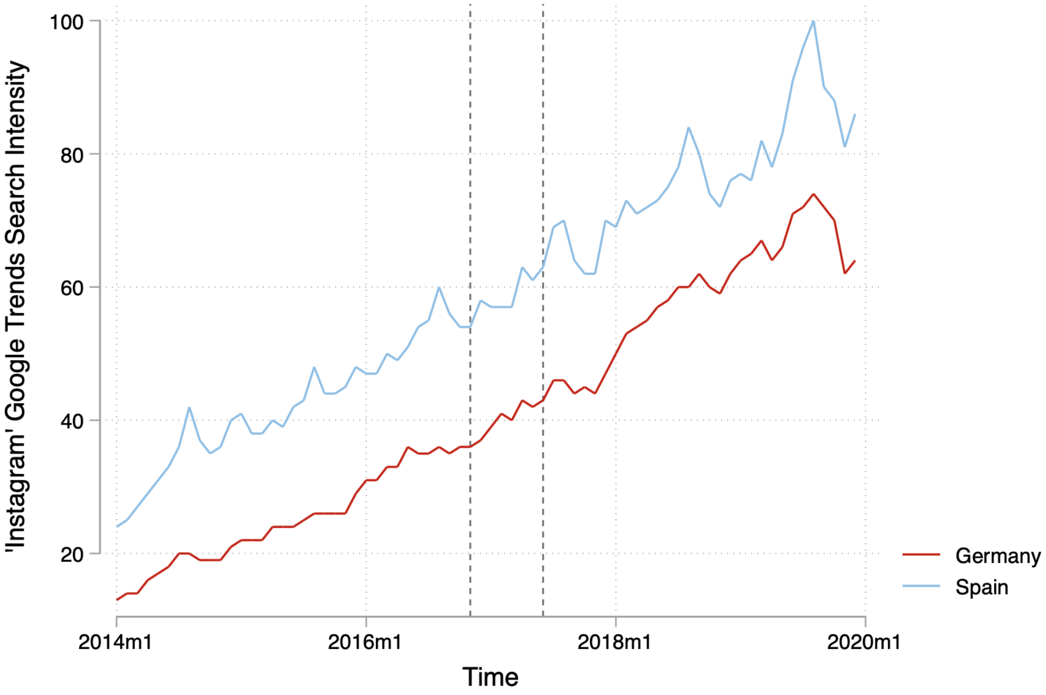
We address this concern partially by controlling for the time-varying popularity of Instagram in Germany and Spain in our main regressions. However, we can also evaluate whether regulations affected Germans' general demand for Instagram more directly. In Figure A6 we show the evolution of a Google Trends search index for the term "Instagram" in Germany and Spain throughout our sample period. Google Trends data is always indexed relative to the peak popularity of the search term, which in this case is Spain towards the end of the sample. The figure shows Instagram's popularity was increasing in both Germany and Spain throughout the sample period. Although the popularity of Instagram in Germany appears to be lower than Spain's, there are no apparent changes in relative popularity after regulations are introduced in Germany.

We also test for changes more formally by estimating a regression using country/month data. For country c in month t:

$$\text{Google Trends}_{ct} = \alpha\left(\text{Germany}_c \times \text{Treated Period}_t\right) + \beta X_{ct} + \delta_c + \delta_t + \epsilon_{ct} \tag{5}$$

where the outcome variable is the level of the Google Trends search index. $(\text{Germany}_c \times \text{Treated Period}_t)$ is a dummy variable equal to one for Germany after the initial changes in the regulatory environment (November 2016). X_{ct} include the log of GDP per capita and the log of population in each country. δ_c and δ_t are time and country fixed effects.

Figure A6: Google Trends Search Index for "Instagram" Term



Notes: Data comes from Google Trends. The vertical axis shows the popularity of "Instagram" as a search term in each country in a given month (relative to the peak popularity of the search term across the two countries). The first dashed vertical line represents the initial changes to German disclosure regulations (see Section 3.2). The second vertical line represents the first fines handed out to German influencers in mid 2017.

Table A9: Google Trends DiD Estimates

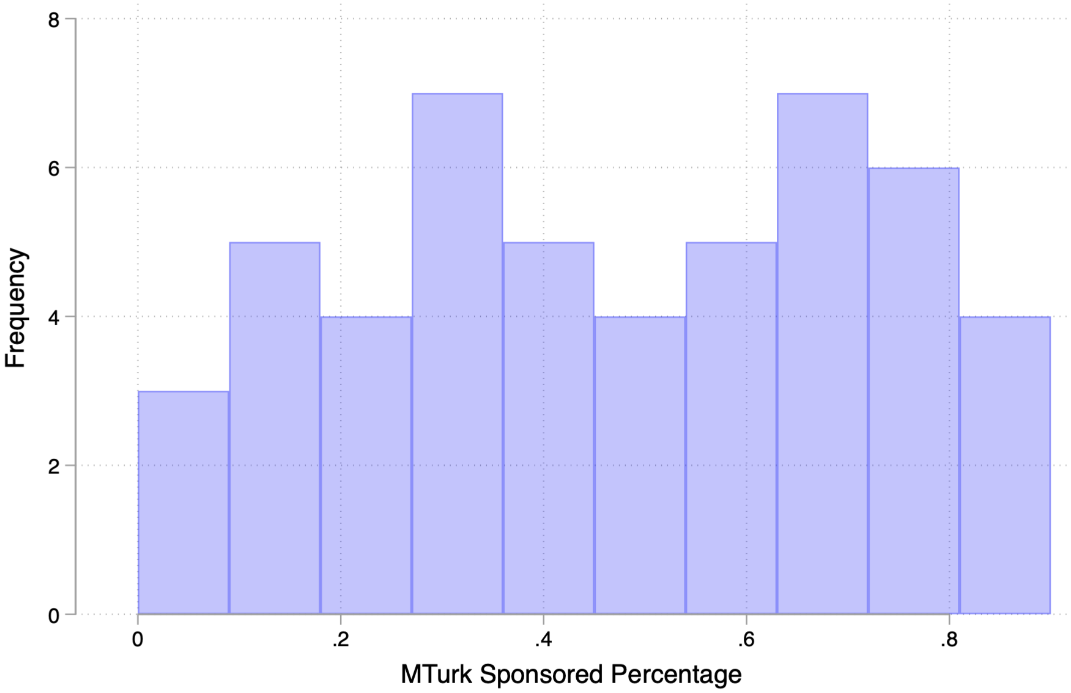
	(1)	(2)
Outcome:	Google Trends ”Instagram” Search Intensity	
Germany × Treated Period	-1.264	0.613
	(3.238)	(0.878)
Germany	-17.114***	27.365
	(1.974)	(68.690)
Treated Period	31.657***	
	(2.396)	
Year-Month FE	NO	YES
Country Controls	NO	YES
Observations	144	144
R-squared	0.774	0.995

Notes: Sample includes country/month level observations from January 2014 to December 2019. “Germany × Treated Period” is a dummy equal to one for all German observations after November 2016 and zero otherwise. Country controls include quarterly GDP per capita and quarterly population. Standard errors are robust to heteroskedasticity. *** p<0.01, ** p<0.05, * p<0.1.

We show estimates of this regression in Table A9. The first column includes only a “Germany” dummy, a “Treated Period” (post- November 2016) dummy and the interaction of the two. The second column includes the full set of year/month fixed effects as well as the additional time-varying country-level controls. Estimates from both regressions show that there were no statistically significant changes in Instagram search intensity in Germany relative to Spain after changes to the German regulatory environment.

Altogether, this evidence suggests that changes in German advertising regulations did not shift the overall perception or demand of Instagram content in Germany.

Figure A7: MTurk Percentage Sponsored Histogram



Notes: Data comes from 50 posts evaluated by 20 MTurk users each. Horizontal axis shows the share of MTurk respondents (out of 20) who label a post as "sponsored (advertising)." The vertical axis shows the number of posts that are at that sponsored percentage.

A.12 MTurk Survey

In this section we describe the results of an Amazon MTurk survey asking a sample of MTurk workers to classify whether specific Instagram post captions belong to a sponsored or non-sponsored post. This survey had two goals: test whether there was a large number of *undisclosed sponsored* posts in our sample, and benchmark the performance of our automated text-based classifiers.

We randomly chose a sample of 50 posts made by German influencers in 2018, after changes in the regulatory environment settled down. We specifically chose posts that were *not* disclosed as advertising according to our disclosure words. Each post was evaluated by 20 MTurk users.[98] The users were shown captions one at a time and directed with the prompt "does this Instagram caption describe (or belong to) a sponsored post/advertisement? Did the user receive a reward / payment for posting it?" and were asked to choose between two options: "Sponsored (Advertising)" and "Non-Sponsored."

[98] Because of various formatting restrictions on the MTurk platform we were required to translate each post into English. We did this using Google Translate. This means that the MTurk users who evaluated the posts were not German and were not required to speak German.

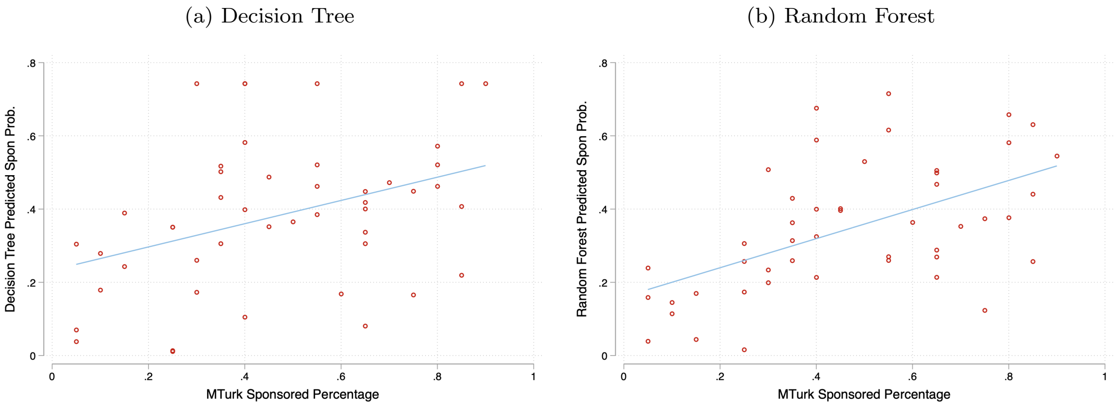
We find that MTurk users believe there are many sponsored and undisclosed posts in Germany after disclosure regulations are introduced. If we label a post as "MTurk-Sponsored" using a majority rule when 10 out of 20 respondents label it as sponsored, then 24 out of 50 posts are sponsored. Even if we use a stricter labelling requirement and only label a post as sponsored when 75% (15 out of 20) label it as sponsored, 10 out of 50 posts (20%) are sponsored. We also find some disagreement among respondents about whether posts are sponsored. Figure A7 shows a frequency histogram of posts by the percentage of respondents who think they are sponsored. More than 10 of the posts that would not be labelled sponsored by simple majority-rule (at least 10 out of 20 users) are labelled as sponsored by over 30% of the respondents.

There is substantial agreement between sponsorship labelling based on the MTurk survey and sponsorship labelling based on the other classifiers we use in the paper. Of the 50 posts, the Naive Bayes (NB) classifier labels 25 as sponsored, the SGD and Random Forest (RF) label 11 as sponsored, and the Decision Tree (DT) labels 12 as sponsored. Our manual classification approach labels 22 as sponsored. Of the 25 NB classified sponsored posts, 17 are also labelled as sponsored by MTurk user majority-rule. Similarly, of the 11 SGD-labelled sponsored posts, 7 are also labelled as sponsored by MTurk user majority-rule.

The Decision Tree (DT) and Random Forest (RF) classifiers also produce a continuous $[0, 1]$ post-level sponsorship probability rather than a discrete label (see Section A.9 for more discussion). We plot DT and RF predicted sponsorship probabilities against the share of MTurk users in Figure A8. This figure shows that the predicted sponsorship probabilities are strongly positively correlated with the share of MTurk users who believe that the posts are sponsored. This is particularly the case for RF probabilities. Pairwise correlations between MTurk shares and DT and RF probabilities are 0.38 and 0.55, respectively.

Overall, the MTurk survey helps us observe that there is likely a substantial number of sponsored and undisclosed posts even after disclosure regulations are introduced in Germany. The high agreement and substantial correlation in responses between MTurk users and our automated approaches also helps confirm that our classifiers correctly identifying the areas in embedding space associated with sponsorship.

Figure A8: MTurk Sponsorship Comparison with ML Sponsorship


(a) Decision Tree

(b) Random Forest

Notes: Data comes from 50 posts evaluated by 20 MTurk users each. Each dot on a scatter plot is a post. The horizontal axis shows the share of MTurk respondents (out of 20) who label a post as "sponsored (advertising)." The vertical axis shows a ML classifier predicted probability that the post is sponsored. Panel (a) uses a Decision Tree predicted probability and panel (b) uses a Random Forest predicted probability. The classifiers use embeddings generated from post text and are trained on a sample of 300,000 German posts as discussed in Section 4.4.2.