The Effects of Influencer Advertising Disclosure Regulations: Evidence From Instagram

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

Matthew Mitchell
Rotman School of Management
University of Toronto
matthew.mitchell@rotman.utoronto.ca

May 28, 2020

PRELIMINARY AND INCOMPLETE
PLEASE DO NOT QUOTE OR CITE

Abstract

We collect data from fifty top Instagram influencers in Germany and Spain from 2014 to 2019. Germany experienced changes in disclosure regulation for social media sponsorship during the sample period. Using a difference-in-difference approach, we study the impact of the the rules on the content of posts and the nature of interaction of followers with the posts. On the content side, we measure whether posts include suggested disclosure terms and show variable but substantial adoption of disclosure. We use an approach based on a fixed list of words associated with sponsorship (i.e. links, mentions of brands, use of words like "sale") as well as natural language processing to assess the likelihood that a post is sponsored. We show that sponsored content use may have increased after changes in disclosure and that followers may have been negatively affected. On the other hand, there is evidence that consumers' reaction to sponsored posts, measured by likes, may be quite different under stricter disclosure rules, suggesting that the rules could have a substantial impact on information transmission.

*We would like to thank workshop participants at the University of Toronto and the 2019 Israeli IO Day for their helpful comments on this project.

1 Introduction

Proliferation of options online makes product discovery difficult for consumers. This is particularly true in many online markets where prices are either uniform or zero (i.e., iTunes, Netflix, internet webpages as a whole). Consumers often rely on advice from intermediaries such as blogs and popular social media users ("influencers") that review and recommend products, or Google search results. Influencers can have potentially important effects on markets. A randomized experiment on Twitter suggests that social-media influencers can change consumer beliefs (Alatas et al. 2019). Müller and Schwarz (2019) similarly show that conversations on social media influence real world behaviour.

Online advice is often sponsored, as influencers are compensated to post sponsored content about specific products. 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). Google search also mixes "organic" search results and sponsored links. Google earned more than $130 billion USD from advertising in 2018 (AndroidAuthority.com). How to regulate influencers both large and small is an important question in the online world. Influence can be actively impacted by sponsorship. In recent years, concerns about hidden sponsored content and misleading online advertising led to regulatory scrutiny. A growing number of countries (i.e., Germany, the UK and the United States) instituted "disclosure regulations" on social media posts (ASA.org.uk). Under a disclosure regime, users have to clearly identify content with a "#ad" (or "#werbung") if they were somehow compensated for it. Both the legislation and enforcement of disclosure regulations varies substantially across countries, but in some countries such as Germany failure to comply has resulted in fines for influencers and advertisers (ISLA.com).

Drawing on theories of buyer-seller transactions, regulatory agencies (i.e., FCC in the US, or ASA in the UK) view disclosure regulations are strictly welfare increasing. More information is better, as consumers are less likely to unknowingly click (and purchase) low-quality sponsored products. As influencers (and brands) observe that sponsored content is less effective, they will produce less of it. Inderst and Ottaviani (2012), Mitchell (2020), Pei and Mayzlin (2019) and Fainmesser and Galeotti (2018) suggest that such a view might be incomplete. 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, leading to lower market welfare. In Mitchell (2020) 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. In Fainmesser and Galeotti (2018) disclosure generates selection of followers, leaving influencers with followers who are less elastic to advertising. In both cases, there may be an increase in sponsored content and a reduction in authentic advice.

This paper empirically tests the two theories, asking: What effects do advertising disclosure regulations have on content in user generated social media platforms? We collect new Instagram data on 50 of the most popular *local* Instagram influencers in two EU countries, Germany and Spain. Germany introduced disclosure regulations in November 2016 and Spain had no similar regulations. Top German and Spanish influencers in our data are broadly similar to one another, and these are the influencers who would likely be affected by a disclosure policy (i.e., most likely to face enforcement).

Our data spans 2014 to 2019. 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. Using a difference-in-differences approach, we exploit the introduction (and timing) of regulation by comparing the behaviour of similar users in Germany and Spain before

and after regulation is introduced. Relative changes in influencer behaviour after the regulation is introduced identify the effects. We look at a number of influencer/month level outcomes: how much do influencers disclose (does the regulation actually work?), as well as how much sponsored content (either disclosed or non-disclosed) is posted by influencers. We also look at whether the type of sponsored content changed, and whether the number of followers of influencers was affected by disclosure regulations. In addition, we consider the effects on follower engagement that helps underline the mechanism through which disclosure affects the market. We test whether the average number of likes and comments changes after disclosure regulations for disclosed-sponsored, non-disclosed-sponsored, and non-sponsored posts.

For this empirical exercise we need to detect which posts are sponsored. Detecting a sponsored post that is disclosed is trivial, but separating sponsored content that is not disclosed from non-sponsored content is more challenging. We use two methods to define sponsored posts: (1) a manual-rule based approach that defines sponsored posts as posts that include certain keywords associated with sponsorship (i.e., "contest", "promotion," "promo code," or a brand name). (2) a supervised machine learning based approach that labels non-disclosed posts with similar language to disclosed sponsored posts as "sponsored." We use posts from Germany after regulation takes place as training data. Germany is a country with relatively strict disclosure regulations. We use a Naive Bayes algorithm for classification.[1] The manual and supervised machine learning based classifiers produce similar results.

We find that the introduction of disclosure affects the type of content that influencers post online. The share of NB-labelled sponsored content influencers post increases by 7% (relative to a pre-treatment baseline of 19%). In addition to the shares, the total number of sponsored posts increases. Disclosure shares naturally increase after regulation, although there are still a substantial number of posts that are not disclosed. Consistent with the other effects, the number of followers falls likely because consumers in these markets do not like sponsored content. Timing tests suggest that these effects are not driven by pre-trends and the parallel trends assumption holds.

At the post-type level, we find that the average number of likes per post falls in Germany after regulation (controlling for the number of followers). This is primarily driven by non-disclosed posts. We also find that the average number of comments per post increase after regulation in Germany (controlling for both the number of followers and the number of likes per post). A decomposition suggests that this is primarily driven by non-disclosed sponsored posts. Instagram does not have a "dislike" button, but followers who dislike a post can mention this by commenting on it. The increase in comments for non-disclosed sponsored posts then reflects increasing consumer dislike of hidden advertising.

The contributions of this paper are two-fold. This is the first paper looking at the effects of online advertising disclosure regulations in a market where there is no direct compensation between the advisor and the advisee. While there is some recent theoretical literature on the subject, their predictions have not been tested empirically.[2] There is also widespread skepticism in the popular press about the effectiveness of such regulations. We show that disclosure regulations have an effect on actual disclosure and also influence content production otherwise. This paper is part of a growing body of literature investigating the effects of disclosure regulations on intermediaries - for example, in the market for insurance advice (Bhattacharya et al. 2019). In these markets, however,

[1] We also experimented with using decision trees and random forests. Both give qualitatively similar results to NB, but we are more concerned with overfitting using these methods. Computer science literature testing various ML approaches for supervised classification on social media text suggests that NB outperforms most other methods (Silva et al. 2020).

[2] There is also a legal literature on advertising disclosure regulations. Recent works include Ducato (2019) and Goanta and Ranchordas (2019) among others.

there is direct compensation between the intermediaries and consumers. Changing disclosure rules may then have different effects. Unlike other markets, we also have clear indicators of disclosure and can see when disclosure does not happen. We also have very detailed information about the content of actual "advice" influencers give since we observe post texts, allowing us to capture multiple dimensions of biased advice.

This project is also related to the literature on the effects of transparency regulations on communication, especially Hansen et al. (2018). We exploit a natural experiment of changes in regulation to identify the effects of enforced transparency. We also measure changes in communication using machine learning base methods. However, our focus is on intermediary behaviour rather than regulator behaviour in a very different market. Our setting also allows using supervised rather than unsupervised clustering algorithms which reduces potential measurement errors due to subjectivity.

Our findings are relevant for regulators 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 also 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 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.

The paper proceeds as follows. Section 2 describes the industry background, Section 3 the conceptual theoretical frameworks. Section 4 presents the data we use and discusses how we define sponsored content. Section 5 presents our empirical identification strategy. Section 6 shows the results and Section 7 concludes.

2 Institutional Background

2.1 Advertising on Instagram

We consider the market for influence on Instagram, a social media platform with over 1 billion active users in 2019.[3] Instagram users post photos accompanied by captions. Users can also "follow" each other and respond to other users' posts by "liking" or commenting.

As in other social media networks, content is produced and provided by users for free. Instagram is also a popular market for sponsored posts. The majority of sponsorship happens through independent online marketing agencies which connect advertisers and Instagram influencers.[4] Users and brands sign up with intermediaries (i.e., Heepspy.com, HypeAuditor.com). The users (influencers) give access to intermediaries of their analytics and are then subdivided into types such as nano-influencers (less than 20k followers), micro-influencers (less than 100k followers) and so on (mention.com). Users are also divided based on the "quality" of their audience and their interests.

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

[4]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 clearly delineated as "Promoted." We abstract from this advertising channel.

Advertisers can sponsor posts in three ways. First, they can send influencers free products (or services) in returns to a post. For example, influencers may receive a free trip to a city in returns for a series of posts about that city. Second, advertisers can 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. Third, 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.[5]

2.2 Advertising Disclosure Regulations

Social media advertising regulations are not standardized across countries within the EU. Influencer advertising is not directly subject to the GDPR. Consumers choose to follow influencers and the advertising does not involve the collection of personal data outside of agreements that influencers sign (sideqik.com). There are EU-wide existing advertising disclosure regulations which apply to traditional media (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). Most influencers cannot be simply defined as "traders" - a travel influencer posting pictures of herself on trips does not obviously have commercial interests.[6] 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). However, these were 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 by competition authorities. Jurisdiction is based influencer residence - influencers who live in Italy are subject to Italian regulations. Below we describe German regulations.[7] To the best of our knowledge, there are no disclosure regulations in Spain beyond the baseline non-binding EU regulations.

In Germany, "die medienanstalten," a group of 14 Media Authorities responsible for licensing and supervision of media released new guidelines on social media advertising on October 18, 2016 (Osborne-Clarke.com). These guidelines require labelling of any posts where the influencer has been renumerated by a brand (including free products) as an ad. The guidelines referred to relevant laws such as the German Marketing Law (UWG - also known as the Unfair Competition Law). In 2017 and 2018 there was a series of high-profile cases and fines against German social media influencers. A German YouTube fitness influencer with 1.3 million followers was fined over 10k EUR for failing to disclose a post 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. In 2018, a court in Berlin ruled that if the purpose of an influencer is merely

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

[6] 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.

[7] In Appendix A.1, we also describe Italian and French regulations which are looser than German ones and similar to FCC regulations in the US.

to keep followers updated about trends, a link to the brand is unnecessary and clearly denotes commercial intent that should be labelled as advertising (Ducato 2019).[8]

To the best of our understanding, German regulations on influencer advertising and their enforcement are the world's strictest.

Some observers of this market suggest that Italian regulations and enforcement are stricter than the German ones and the French ones (Ducato 2019). However, the number of high profile cases and fines related to influencers is larger in Germany, and they are notably not limited to top influencers but also those with fewer followers. German laws related to misleading advertising are also competition laws rather than consumer protection laws.

3 Conceptual Framework

We broadly rely on Mitchell (2020) and Fainmesser and Galeotti (2018) to frame our thinking about the behaviour of agents in this market and what we would expect to find empirically after the introduction of advertising regulations. The two papers set up very different models but reach similar conclusions regarding the effects of advertising.

Mitchell (2020) 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 sponsoship. 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 (2020) mimics disclosure regulations through a counterfactual that lowers the influencer's returns for posting sponsored content. The model predicts that this may negatively impact followers in two ways: it reduces the ability for followers to reward authentic content by also engaging with sponsored content, and may in turn lead to worse content.

Fainmesser and Galeotti (2018) 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 form beliefs about the degree of authenticity of the influencers. 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 remove the asymmetric information between influencers and 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 total welfare falls in the market after transparency.

[8] An even stricter interpretation was provided by a lower court in another case - any post by an influencer who has previously used their account for commercial gain can be considered to be a commercial post and should be labelled as an ad, even if the post does not mention an advertiser. This interpretation was overturned by an upper court of appeals.

In summary, the main predictions from both models suggest that transparency can have unintended consequences - an increase in the share of sponsored content at the expense of authentic content. Empirically, this means that the share of sponsored content could increase as a fraction of total influencer posts. Follower welfare could also fall. This is reflected in the loss of followers in equilibrium. Again, this is a clear empirical prediction. Engagement between the remaining followers and influencers should also change. As influencers increase the amount of sponsored content they provide, there should be a decrease in the beliefs of followers about the quality of their authentic advice. This means that in the data we should expect to see decreased engagement in non-disclosed posts - a smaller number of likes. We should also observe increased "dislike" of posts.

4 Data

4.1 Data Description

We collected data from Instagram in a semi-automatic way. Raw data is at the post-level and we observe a full 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 one image, multiple images or a video).

We have data for a sample of 50 influencers from Germany and 50 influencers from Spain. The influencers are popular - the most popular Instagram users residing in each country (according to HypeAuditor.com). These are the users who are most likely to face enforcement of a disclosure policy. Because of the power-law nature of popularity online, there is still substantial heterogeneity in influencer popularity within the sample. The top 5 influencers in our sample in each country have over a million followers by the end of the sample period (2019), whereas the bottom 5 influencers have less than 100k followers. We select our influencers to be local by looking at the % of their followers from the country (available on HypeAuditor.com) and comparing Google Trend search query volumes in different countries. We exclude influencers that are popular across many countries.[9] We do this to make sure that influencers are only affected by laws of the country in question, rather than laws in other countries.[10] The Spanish followers' conception of the world is also not being changed by regulation in Germany since most Spanish followers are not reading German posts.[11]

Because of our selection procedure, German and Spanish influencers in our data are similar to one another. In both countries the sampled influencers are primarily female and are either "lifestyle" influencers, locally popular musicians, locally popular actresses, or fashion models.

We supplement post-level data with partial historical data on the number of followers for influencers scraped from socialblade.com. This website tracked daily Instagram follower counts from 2014 until March 2018. We merge follower counts with post-level data using influencer user-names and post dates. We also have contemporaneous follower counts from the date of data collection from Instagram (i.e., if we collected post data from influencer X on January 5 2019, we observe the number of followers for that day). We linearly interpolate the number of followers between March 2018 and 2019.

[9] This means we exclude soccer players from our sample.

[10] 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.

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

Table 1: Influencer/Month Summary Statistics for Germany and Spain

Variable	Obs	Mean	Std. Dev.
Mean Likes per Post	5,067	51,429	101,366
Mean Comments per Post	5,067	1,365	5,690
N Followers	3,954	905,766	1,467,323
N Posts per Month	5,067	27	28
Account Age (months)	5,067	37	20
First Account Year	5,067	2013	1

The number of likes and comments are recorded at the time of data collection rather than at the time of the posting. This may introduce measurement error as posts made earlier would have more time to accumulate likes. However, industry experts estimate that engagement on posts on Instagram dies 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.[12]

We also include country-year/month specific data. Germany and Spain are different in many respects - such as income - which could affect the number and frequency of advertisements posted by influencers. The same is true for the popularity of Instagram as a whole. We control for cross-country differences using data on per-capita income and population from the OECD. We control for the popularity of Instagram by including Google Trends search query volumes for the keyword "Instagram" from each country between 2014 and 2019.

The merged data is aggregated to the influencer/month level from 2014 to 2019. In our regression analysis below we consider influencers who have more than 2 posts in a given month. Summary statistics are in Table 1. These show that influencers in our sample are very popular by Instagram standards - with an average of almost 900,000 followers. Despite having a relatively small sample of influencers, they generate a substantial amount of content. An average influencer account is 38 months old and generates 27 posts per month. The average influencer's post is liked around 50,000 times and is commented on 1,300 times.

4.2 Definition of Sponsored Posts

Our goal is to separate three types of posts using text data: non-sponsored posts, disclosed-sponsored posts and non-disclosed sponsored posts. Representative examples of the four types of posts from Kylie Jenner's Instagram account ("@kyliejenner") from early December 2018 appear in Figure A1. 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 (b) is clearly non-sponsored. There are no links or any text that refers to an advertiser.

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 disclosure is not at the beginning of the post but at the end of the post and it does not clearly denote sponsorship. Nonetheless,

[12] On other social networks, the life of posts is even shorter. Engagement for posts were estimated to have died out on average after 18 minutes on Twitter (moz.com).

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

(a) Sponsored and Disclosed

(b) Non-Sponsored and Non-Disclosed

(c) Sponsored and Likely Disclosed

(d) Possibly Sponsored and Non-Disclosed

to be as conservative as possible, we consider such a post disclosed and sponsored.[13] We consider the post in panel (d) as an example of a possible non-disclosed sponsored post. On the surface 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 webpage. Through this webpage the decorator can be hired for additional jobs. In that sense, it is not different than the Adidas or diet tea posts in the other panels. It is possible to imagine a sponsored contractual arrangement whereby Kylie Jenner receives discounted interior decorating services in return for a post with a link.[14]

An interesting comparison is of the number of likes for each post in Figure A1. 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 clearly suggests a hierarchy in terms of consumer preferences that we will continue to explore in subsequent sections.

[13] In our subsequent analysis, "ambassador" (in German and Spanish) reflects disclosure, as well as any references to "collaboration," "partnership," etc. A full list of words we use to connote disclosure is in Appendix A2.

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

The definition or detection of sponsored posts which are disclosed via a "#ad" (or its German or Spanish equivalents) is trivial. We use additional words to detect disclosure that come from national and international advertising guidelines (a full list of words used to detect disclosure is in Appendix A.2). The definition or detection of sponsored posts which are not disclosed and separating those from non-sponsored posts which are not disclosed is more difficult. We propose two approaches for this: (1) A "manual" approach: separating posts into sponsored and non-sponsored using a list of manually pre-determined keywords which seem to generally denote sponsorship. (2) A supervised "automatic" approach: separate posts into sponsored and non-sponsored using a machine-learning based classifier. More details about the two method are below.

Both methods rely on a conceptual framework which is consistent with theoretical work outlined in Section 3. Broadly speaking, there are two possible message types for influencers to send to their followers: an authentic message, or a sponsored message. Influencers pick words out of some vocabulary to send the appropriate message type.[15] The manual method assumes that if certain words are present, the message is definitely sponsored. It also assumes that the set of words that denote sponsorship is known to the researchers. The automatic method is probabilistic - a word can denote sponsored content with some probability but also authentic content with some probability. A higher share of words that are more "sponsored" push the message (post) to be labelled as sponsored. The algorithm uncovers the probabilistic distribution of the words across message types.

4.2.1 Manual Classification

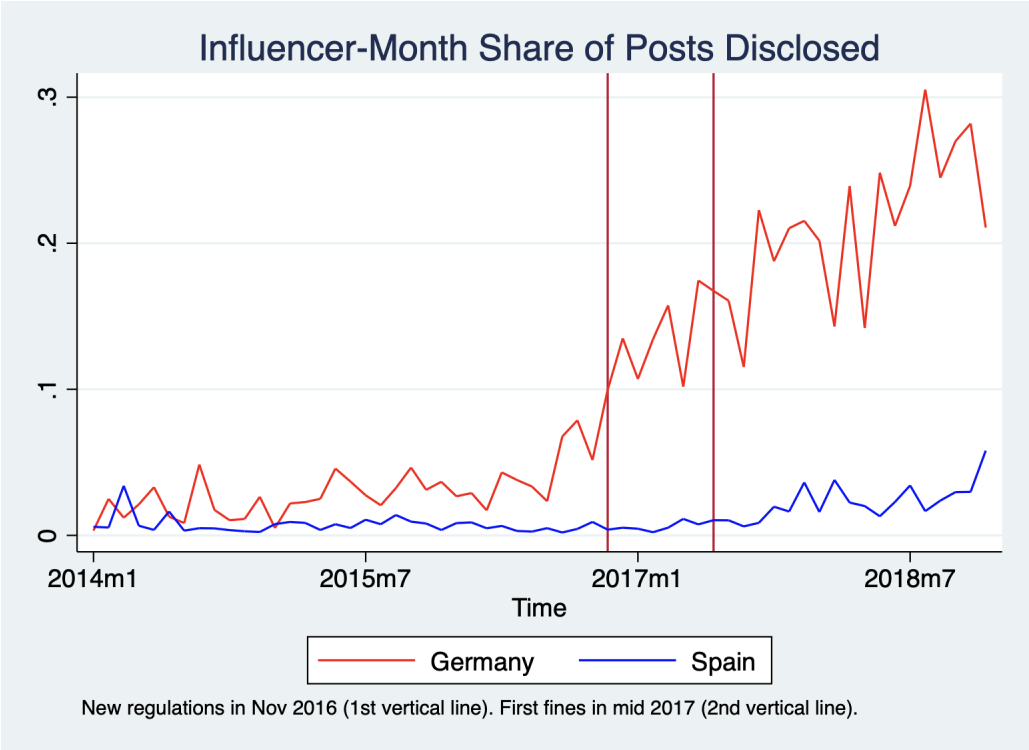
We define a set of words that we believe connotes sponsorship. We use translations of English, Spanish and German words. These include references to coupons, contests or discount codes - many sponsorships allow influencers to offer discounts for products. Relevant keywords also include any links to outside websites (anything that ends with ".com," ".de," etc), 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"). We also include a large list of nearly 1,000 brands, including German and Spanish specific brands (e.g., El Corte Ingles). This manual definition would capture the three sponsored and possibly-sponsored posts from Figure A1. A full description of the keywords is in Appendix A.3.

The main benefit of this approach is that it is fully transparent and fast to implement by searching the text of each post for a pre-determined list of keywords. The drawbacks of this approach is that we need to pick the keywords that denote sponsorship. This means that we can either undershoot or overshoot the true sponsorship rates if we miss some keywords, select keywords that do not actually denote sponsorship, or select ambiguous keywords that are commonly used in both sponsored and non-sponsored posts. We chose a loose manual definition of sponsorship with a number of words that can be ambiguous.[16] Given this, our manual definition likely overstates the amount of sponsored content in our sample. With this approach we believe that it is better to overstate rather than under-state sponsored content. If we select words that do not actually belong

[15] This is conceptually similar to a multinomial discrete choice model. It is also consistent with the way that sponsorship actually works - influencers are often required to use certain vocabulary in their sponsored posts by advertisers.

[16] It is possible to be *even looser* in the choice of words. For example, we can assume that all tags denote sponsorship. Broadly speaking, this is clearly not true - tags are popular on Instagram as methods of communication across accounts or acknowledgement. They are also often used to link to sponsoring pages.

Figure 2: Post Disclosure in Germany and Spain



to sponsored posts we should not see a change after regulation. This means our estimates are a lower-bound on 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. Robustness checks of our main results with a smaller set of manual keywords gives qualitatively and quantitatively similar results.

4.2.2 Supervised Machine Learning Classification

A supervised machine learning approach takes a labelled dataset, trains a classification algorithm on this data, and then projects the trained algorithm on non-labelled data. We take German data after disclosure regulations went into effect as our training data. This seems to be a reasonable approach. Figure 2 presents a graph of the percentage of total posts disclosed as advertising in Germany and Spain from 2014 to 2019. There are two vertical lines on this graph. The first represents the initial introduction of disclosure regulations in Germany in 2016. The second vertical line represents the beginning of enforcement of disclosure in Germany via fines to influencers (see Section 2.2 for more details). This figure shows that disclosure in Germany increased drastically almost immediately after the introduction of disclosure regulations. Anecdotal evidence obtained browsing the feeds of several German influencers also suggests that influencers are disclosing all or nearly all of their sponsored content. There is no similar change in disclosure for Spain.

We have around 30,000 posts in training data (post-regulation Germany). In the non-labelled (testing) data, we have 30,000 posts from Germany before regulation and 100,000 posts from Spain. Prior to running the ML algorithm we translate all Spanish posts into German using Google Translate. We then take out common stopwords in both German and English as well as

words that are actually used for disclosure of advertising (see Appendix A.2 for full list of words reflecting disclosure). We also throw away text that includes numbers/ are non-words, such as emojis.[17] We then tokenize our texts, leaving us with a dictionary of words and their incidences in each post.

In the post-regulation period for Germany, we compare the posts that are labelled as disclosed and posts that are labelled as non-disclosed. The algorithm then looks for words that are associated with the disclosure label. We effectively compute a probability of $P(disclosed|words)$ for a given post, or equivalently the ratio of this to $P(non-disclosed|words)$. Various algorithms compute this probability differently. For example, a Multinomial Naive Bayes algorithm computes $P(disclosed|words)$ as a function of $P(words|disclosed)$ which can be calculated as a product of $P(word_i|disclosed)$ for each $word_i$ in our text. When we project the algorithm on the testing data, we obtain a $P(disclosed|words)$ and $P(non-disclosed|words)$ for each post i. We then label a post that has a higher probability of being disclosed conditional on the words as a sponsored post, and a post which has a higher probability of being non-disclosed as a non-sponsored post.

For our main results, we use a Multinomial Naive Bayes classification algorithm.[18] Naive Bayes is a fast algorithm to train and to apply. It is also parsimonious and generates intuitive conditional probabilities. It admits an interpretation as coming from a simple strategy of the influencer across words for different types of posts. Suppose that an influencer (similar to a sender in much of the literature on strategic communication) is sending a message of type $\theta \in \{S, N\}$ with a list of words $\{w_i\}$. One can think of the type as whether or not the post is sponsored. Each word has a probability p_i^m of being sent in a message of type m. Under this strategy, Naive Bayes is equivalent to the true probability of the type of a message. Choosing many words allows the sender to have "degrees of separation" between messages, in terms of their probability of being sponsored.

The main disadvantage of the Naive Bayes algorithm is its naivety, or the conditional independence of words. There are alternative classification algorithms - Random Forests, Gradient Boosting, SVM, Neural Networks - that loosen this restriction. These algorithms are more complex to apply and are potentially more subject to issues such as overfitting and false positives. Despite its potential drawbacks Naive Bayes often performs as well as more complex algorithms. Silva et al. (2020) test multiple algorithms in the classification of Facebook posts as political advertising or non-political advertising. They find that Naive Bayes generates similar Accuracy, AUC and Macro-F1 scores as SVM, Random Forests, Gradient Boosting and Convlutional Neural Network classifiers. They also find that Naive Bayes substantially outperforms all classifiers in its True Positive Rate for a given False Positive Rate, suggesting that it is less suspect to overfitting.[19]

More generally, the main strength of a supervised classification approach is that it does not require us to pre-specify what we think are words that denote sponsorship or non-sponsorship. Rather, we let the algorithm decide. Our results suggest that this works well. Appendix A.4 shows the top 20 most predictive words from the Naive Bayes algorithm. These are the words that suggest sponsorship. In addition to obvious features such as "code" and "percent," these words include references to "today" (as in "out today"), "having" (as in "you can have this..."), and so on.

A weakness of the supervised classification method is the selection of training data and assumptions we need to label this data. In the training data, we need to assume that all sponsored posts are disclosed. If disclosure is imperfect this means the algorithm will necessarily separate sponsored and non-sponsored posts. If disclosure is at 50%, for example, there will be sponsored posts that we

[17] We do convert the symbol % into the word "percent." We also convert the symbol # into the word "hash."

[18] Run using the Python sci-kit module.

[19] As robustness checks, we applied a Decision Tree and a Random Forest algorithm to classify this data. We find substantial issues with overfitting using these algorithms. Nonetheless, qualitatively, our main results in Section 6 do not change.

will label as non-sponsored in the training data. A comparison between our labelled sponsored and non-sponsored posts may not find many of the relevant words that denote sponsorship. To test for this potential issue, we consider different time periods for the training data. Figure 2 shows that there is an increase over time in the share of disclosure in Germany, peaking around 2018. This could suggest that "full" disclosure only kicks in after some time. We run the algorithm with the full post-regulation data as our training set, and with only 2018 post-regulation German posts as our training set. The results in terms of predicted sponsorship rates or the posts that the algorithm considers to be sponsored do not change.

Another concern with this approach is about the similarity of dictionaries in the training and testing data. This is a concern because we need to translate posts from Spanish into German. There may be a case where German influencers use the word "great" (in German) and Spanish influencers also use the word "great" in Spanish but we translate this word into "awesome" (in German). If these cases are frequent, we would end up with very different dictionaries for Germany and Spain.[20] A more general version of this concern is that language evolves over time, possibly in response to advertising disclosure regulation. Influencers in Germany can change how they talk about advertisers over time, leading to very different sets of words being present in pre- and post-regulation Germany (not to mention Spain).

We address these concerns by comparing the dictionaries (i.e., set of observed words) of pre-regulation Germany and Spain to the most predictive words in the training data (post- regulation Germany). Venn diagrams representing these comparisons are in Figure A1 in Appendix A.5. They show that of the 1,000 most predictive words in post-regulation Germany, over 900 are present in Spain and in the pre- regulation Germany and in Spain. Of the 5,000 most predictive words in post-regulation Germany, around 3,500 are present in Spain and 4,000 are present in pre-regulation Germany. This suggests that our dictionaries are similar enough to detect advertising from post-regulation Germany in the other two datasets.[21]

4.3 Comparison of Approaches

Table 2 compares the two approaches using several metrics from ML literature. In the training (post-regulation Germany) data the Naive Bayes approach beats the manual approach in every metric. The *Precision Score* measures the ratio of true positive predictions to all positive predictions. The Naive Bayes Precision Score is 0.79, as compared to the manual 0.44. The *Recall Score* measures the share of actually disclosed posts that were correctly classified as disclosed. These are closer between the manual and NB classifiers at 0.7 and 0.75, respectively. Last is the *Accuracy Score* that captures the share of correctly classified posts in the training data. The NB approach correctly classifies over 87% of posts, whereas the manual approach only correctly classifies 66% of posts.

We can also test to see how the classifiers perform outside the training data, as we have a number of actually disclosed posts in pre-regulation Germany and in Spain. The Naive Bayes algorithm initially seems to work less effectively than manual detection in uncovering actually disclosed posts outside the training data, with Recall Scores of 0.51 and 0.55. Manual Recall Scores are 0.58 and 0.78. This suggests that there is some overfitting. This does not mean that the manual method is more precise. The total number of predicted disclosed (or sponsored) posts using the manual method is much larger than the NB method. The manual method is likely getting more correct predictions simply because it is less accurate in predicting non-disclosed posts. The Naive Bayes

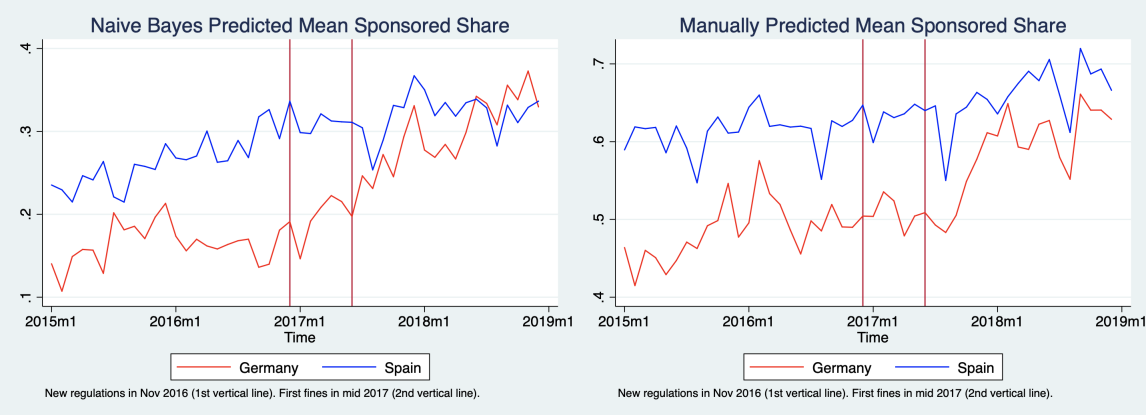
[20] Another approach for dealing with this concern is by not working in word-space but rather with embeddings.

[21] We may still be missing some ads that, for example, use language particular to pre-regulation Germany. This is would mean our estimates are a lower-bound.

Table 2: Comparison of Approaches

	Manual	Naive Bayes
Precision Score (Post-Reg Germany)	0.4364	0.7887
Recall Score (Post-Reg Germany)	0.6978	0.7480
Accuracy Score (Post-Reg Germany)	0.6647	0.8716
Recall Score (Pre-Reg Germany)	0.5798	0.5070
Total Pred. Disclosed (Pre-Reg Germany)	11,667	3,835
Recall Score (Spain)	0.7774	0.5459
Total Pred. Disclosed (Spain)	66,800	29,586

Figure 3: Predicted Sponsored Post Shares in Germany and Spain



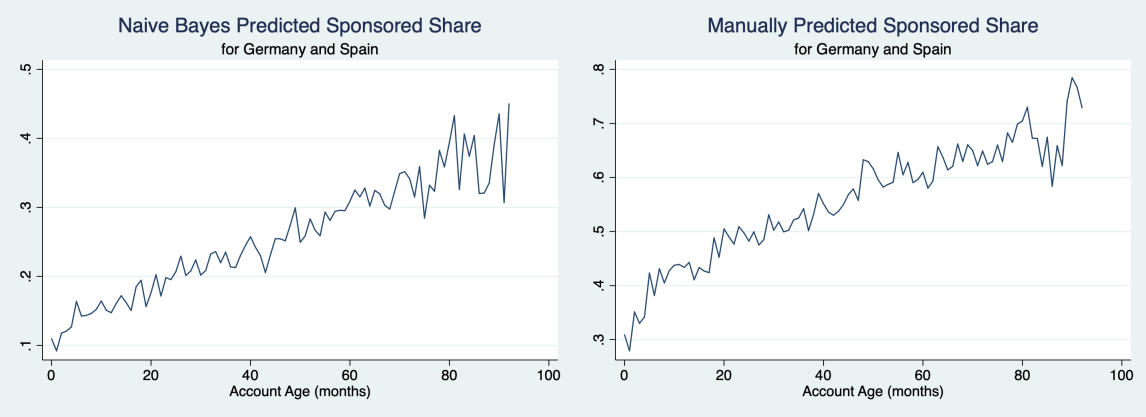
appears to be a more conservative classifier and works well at not classifying non-disclosed posts as disclosed. As long as it is similarly conservative in Germany and in Spain, our estimates should not be affected.

4.4 Sponsorship

In this section we present some features about our predicted sponsorship measure. Figure 3 shows time series of both the Naive Bayes (left panel) and the manual (right panel) sponsored measures for our sample period in Germany and Spain. We track the percentage of all posts in a country in a month labelled as sponsored. There are three descriptive facts that emerge from this data: (1) in both panels Spain has a larger share of sponsored posts than Germany more or less throughout the entire sample period. For example, according to the Naive Bayes predictions, an average 30% of posts in Spain in 2018 are sponsored. This is in contrast to the percentage of posts that are actually disclosed, which in Spain is less than 5%. (2) the share of sponsored posts is increasing in Germany after regulations are introduced. (3) changes in the shares of sponsored posts in Germany and Spain do not perfectly track changes in disclosure from Figure 2. Together, these three facts suggest that we are uncovering something novel about the underlying text data, rather than simply re-stating changes in disclosure.

We also show the evolution of our predicted sponsorship with account age in Figure 4. This figure plots the average share of monthly posts our algorithms label as sponsored (in both Germany

Figure 4: Predicted Sponsored Post Share and Account Age



and Spain) against the age of an influencer's account in months. This picture suggests that the share of sponsored content users post increases over time. This is consistent with Fainmesser and Galeotti (2018) - over time, users become more popular (particularly in our sample), learn how to use Instagram better and make money. This also coincides with theoretical results from Mitchell (2020) - users initially "invest" in higher shares of authentic content and then "cash-out" by posting more sponsored content later on. Moreover, the growth in sponsored content rates is not perfectly smooth which suggests that users alternate periods of more and less sponsored content much as theory predicts.

5 Estimation Methodology

The introduction of regulations in Germany but not in Spain at the end of 2016 suggest a difference-in-differences estimation strategy to identify the effects of disclosure regulations. We compare popular influencers in a country where disclosure regulations were implemented (Germany) to popular influencers in a country where disclosure regulations have not been implemented (Spain) before and after regulation. The intuition for using this methodology in this context is apparent in Figure 2, which shows the share of all posts which are disclosed as advertisements in Germany and Spain from 2015 to 2019. This figure clearly shows that disclosure sharply increases in Germany in the period after regulation and remains constant for Spain. Figure 3 similarly shows that the shares of labelled sponsored posts is stable in Germany and Spain until German regulations kick in.

More formally, we model outcome Y_{it} (i.e., share of sponsored posts) for influencer i at month t as:

$$Y_{it} = \alpha REG_{it} + \beta X_{it} + \delta_i + \delta_t + \epsilon_{it} \tag{1}$$

where REG_{it} is a variable which is equal to 1 if the influencer is in a regulated country after regulation. X_{it} are a set of influencer/time varying controls, such as account age, country characteristics (i.e., popularity of Instagram, GDP per capita) and so on. δ_i and δ_t are influencer and year/month fixed effects which absorb country and "post-regulation period" fixed effects. The key parameter in this regression is α, which captures the average difference between influencers experiencing regulation and influencers who do not experience regulation in the post-regulation period.

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. In fact, the policy itself might be endogenously driven by pre- period influencer behaviour in Germany. We test for this directly in Figure A2. This also helps us directly identify the time at which treatment effects are felt, which helps with determining whether any effects we discover are due to confounders. 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).

FTC regulations on influencers took place in early 2016 and would be clearly captured in the data as differences in pre-trends. There is also no particular reason why FTC regulations would affect influencers in Germany differently than influencers in Spain. Spillovers from German regulations on Spain would only mean that we are getting a lower-bound effect.

There may also be concerns about influencers entering and exiting the sample at different points in time. We flexibly control for influencers' timing of initially joining the platform. We also estimate a balanced panel regression, only having influencers who have been on the platform during the entire sample period.

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

$$Y_{pit} = \alpha REG_{it} + \beta X_{pit} + \delta_i + \delta_t + \epsilon_{pit} \tag{2}$$

The idea is to compare similar posts made by the same influencers over time, and see how their outcomes change in Germany after regulation relative to Spain. In addition to influencer level controls outlined above, we also include post level controls, capturing the type of post this is (video, picture, or series of picture), as well as post length. We also include the number of followers as a control in this regression, as there is an obvious mechanical correlation that needs to be accounted for between liking/commenting intensity and account popularity.

6 Results

Results at the influencer-month level are in Table 3. Column (1) shows estimates for the outcome variable of the share of disclosed monthly influencer posts. Column (2) shows estimates for the share of monthly influencer posts that we manually predict to be sponsored. Column (3) shows estimates for the share of monthly influencer posts that the Naive-Bayes algorithm labels as sponsored. Columns (4) and (5) show disclosure rates for manually sponsored and Naive Bayes sponsored posts as dependent variables. Column (6) shows estimates for the monthly log number of followers. All regressions control for influencer and time fixed effects, as well as flexible influencer account age controls (which allow for different pre-trends depending on when the influencer became active on Instagram). We also include country level controls - population, GDP per capita and a Google Trends search intensity for the term "Instagram" as a control for Instagram's popularity in Germany and Spain. Standard errors are clustered at the influencer level.

Reassuringly, Column (1) results confirm Figure 2 showing that disclosure rates increased at the influencer-level after regulations. Column (2) and (3) results then show that sponsorship itself increased after disclosure regulations in Germany relative to Spain. Using the Naive Bayes labelling the share of sponsored posts for the average influencers increased by 7 percentage points. This is relative to a pre-treatment baseline of 19 percent - a large increase. A similar (though smaller and noisier) increase is seen in the manually labelled sponsored posts. Table A1 in the Appendix shows that this increase in the share of sponsored posts is happening because of the increase in the number of sponsored posts. The total number of posts per-month does not change for influencers in

Germany relative to influencers in Spain, but influencers in Germany end up producing 3 additional Naive Bayes labelled sponsored posts per month. Disclosure rates also increase after regulation. An additional 25% of NB labelled sponsored posts are disclosed in Germany after regulation (Column 5). This is also true for the manually labelled posts. Interestingly, even after regulation comes into effect disclosure rates are far from 100%.

Column (6) shows that the number of followers decreases for German influencers relative to Spanish influencers after regulation. The decrease is on the order of 2 percentage points. This is consistent with predictions from theory about increasing sponsorship rates following regulation reducing consumer engagement.

Timing tests are in Figure A2 in the Appendix. They show that the parallel trends assumption holds and there are no systematic differences between the treatment and control groups in the main outcome variables in the period before treatment takes place. Results from a balanced panel regression are in Table A2. They show that although we lose a relatively large number of observations which increases noise somewhat, the point estimates do not change.

Results at the post-level are in Tables 4 and 5. Column (1) in both tables includes all posts. Columns (2), (3) and (4) break up posts into non-sponsored (Naive-Bayes labelled) and non-disclosed, sponsored and non-disclosed and sponsored and disclosed. In Table 4 we show that the average number of post likes falls in Germany relative to Spain after regulation. This is conditional on influencer and time fixed effects as well as the number of followers of the influencer. A breakdown of which posts are most affected suggests that it is *not* the sponsored posts that are disclosed. Rather followers do not like the undisclosed sponsored posts and the undisclosed non-sponsored posts. This confirms theoretical predictions about changing consumer beliefs about the quality of German influencers following disclosure.

Table 5 shows results for the number of comments. In this set of regressions we control for both the number of followers for the influencer and the number of post likes. We're effectively comparing the difference in the average number of comments for a post with a certain number of likes for a German influencer relative to a similarly popular Spanish influencer before and after regulation. Column (1) results in this table suggest that the number of comments per post increase in Germany relative to Spain after regulations are introduced, *ceteris paribus*. A breakdown by post-type suggests that the only types of posts where this occurs in a way that is statistically different than zero is for non-disclosed sponsored posts. Coefficients for the other types of posts are noisy zero estimates. There is no "dislike" button in Instagram but follower often express their disagreement with a post by commenting on it. Conditional on the number of likes, we interpret the number of comments on posts that are sponsored but not disclosed as an increase in the dislike of such posts. These results then suggest that there is growing dislike of non-disclosed sponsored content after regulation in Germany. This makes sense since consumers are aware of requirements to disclose sponsored content.

7 Discussion and Conclusion

We show that advertising disclosure regulations have real effects. Influencers in Germany increase both the number of posts that are labelled as disclosed and disclosure rates of sponsored posts after regulation is introduced in late 2016. This is an important empirical finding in and of itself given widespread popular skepticism about such regulations (The Guardian). Consistent with previous theoretical work (Mitchell 2020, Fainmesser and Galeotti 2018), we also show that there are potentially adverse effects to these regulations. The number and percentage of sponsored posts increases at the influencer level.

Table 3: Influencer-Month Level Estimates

Outcome Variable:	(1) Disclosed %	(2) Man. Spon. %	(3) NB Spon. %	(4) Man. Spon. Disclosure Rate	(5) NB Spon. Disclosure Rate	(6) ln(N Followers)
Germany	0.115***	0.052*	0.072***	0.159***	0.267***	-0.268**
× Treated Period	(0.028)	(0.030)	(0.026)	(0.038)	(0.061)	(0.130)
User FE	YES	YES	YES	YES	YES	YES
Year-Month FE	YES	YES	YES	YES	YES	YES
User Account Age FE	YES	YES	YES	YES	YES	YES
User Account Age × User First Year FE	YES	YES	YES	YES	YES	YES
Country Controls	YES	YES	YES	YES	YES	YES
Pre-Treatment Mean	0.047	0.400	0.190	0.087	0.118	12.406
Observations	5,067	5,067	5,067	4,744	3,862	3,940
R-squared	0.768	0.636	0.659	0.701	0.736	0.949

Notes: The sample includes monthly observations of Instagram influencers in Germany and Spain with more than 2 posts per month from 2014 to 2019. "Germany × Treated Period" is a variable that is equal to 1 for all German Instagram influencers after November 2016. Additional controls include influencer fixed effects, year-month fixed effects and fixed effects controlling for account age interacted with the first year of the influencer's activity on Instagram. We also include country controls that include log(population), log(GDP per capita), and a measure of country-level Google Trends search intensity for the term "Instagram." Standard errors are clustered at the influencer level. *** p<0.01, ** p<0.05, * p<0.1.

Table 4: Post Level Estimates - Likes

Outcome Variable:	ln(N Post Likes)			
Sample:	All Posts	Disclosed = 0 NB Spon = 0	Disclosed = 0 NB Spon = 1	Disclosed = 1 NB Spon = 1
Germany	-0.230**	-0.241**	-0.197**	0.032
× Treated Period	(0.100)	(0.101)	(0.097)	(0.133)
ln(N Followers)	0.792***	0.778***	0.817***	1.041***
	(0.095)	(0.093)	(0.123)	(0.114)
ln(N Posts)	0.012	0.015	-0.039	-0.028
	(0.030)	(0.032)	(0.041)	(0.041)
ln(Caption Length)	0.006	0.015***	-0.018	-0.015
	(0.005)	(0.005)	(0.012)	(0.012)
Post Type Controls	YES	YES	YES	YES
User FE	YES	YES	YES	YES
Year-Month FE	YES	YES	YES	YES
User Account Age FE	YES	YES	YES	YES
User Account Age × User First Year FE	YES	YES	YES	YES
Observations	103,887	72,489	22,561	6,183
R-squared	0.953	0.956	0.948	0.967

Notes: The sample includes all posts by sampled influencers in Germany and Spain from 2014 to 2019. "Germany × Treated Period" is a variable that is equal to 1 for all German Instagram influencers after November 2016. Additional controls include influencer fixed effects, year-month fixed effects and fixed effects controlling for account age interacted with the first year of the influencer's activity on Instagram. We also include country controls that include log(population), log(GDP per capita), and a measure of country-level Google Trends search intensity for the term "Instagram." We also include dummies capturing whether the post is a photo, a video, or a collection of photos. "ln(Caption Length)" is a measure of the number of characters in the post. Standard errors are clustered at the influencer level. *** p<0.01, ** p<0.05, * p<0.1.

Table 5: Post Level Estimates - Comments

Outcome Variable:	ln(N Post Comments)			
Sample:	All Posts	Disclosed = 0 NB Spon = 0	Disclosed = 0 NB Spon = 1	Disclosed = 1 NB Spon = 1
Germany	0.262**	0.166	0.414***	0.424
× Treated Period	(0.128)	(0.118)	(0.135)	(0.290)
ln(N Likes)	1.138***	1.115***	1.229***	1.050***
	(0.051)	(0.049)	(0.062)	(0.082)
ln(N Followers)	-0.184***	-0.144**	-0.244***	-0.407**
	(0.069)	(0.065)	(0.078)	(0.161)
ln(N Posts)	-0.159***	-0.155***	-0.251***	-0.040
	(0.040)	(0.039)	(0.045)	(0.073)
ln(Caption Length)	0.010	-0.005	0.071***	0.128***
	(0.008)	(0.007)	(0.027)	(0.027)
Post Type Controls	YES	YES	YES	YES
User FE	YES	YES	YES	YES
Year-Month FE	YES	YES	YES	YES
User Account Age FE	YES	YES	YES	YES
User Account Age × User First Year FE	YES	YES	YES	YES
Observations	103,569	72,266	22,481	6,173
R-squared	0.874	0.895	0.836	0.838

Notes: The sample includes all posts by sampled influencers in Germany and Spain from 2014 to 2019. "Germany × Treated Period" is a variable that is equal to 1 for all German Instagram influencers after November 2016. Additional controls include influencer fixed effects, year-month fixed effects and fixed effects controlling for account age interacted with the first year of the influencer's activity on Instagram. We also include country controls that include log(population), log(GDP per capita), and a measure of country-level Google Trends search intensity for the term "Instagram." We also include dummies capturing whether the post is a photo, a video, or a collection of photos. "ln(Caption Length)" is a measure of the number of characters in the post. Standard errors are clustered at the influencer level. *** p<0.01, ** p<0.05, * p<0.1.

There are questions about the welfare implications of these results. If we choose to interpret the number of likes and number of followers as a revealed preference measure of consumer (follower) utility in this market, our findings suggest that consumer welfare falls after regulation. There are fewer followers per influencer (on average), and conditional on the number of followers there is a decrease in the number of likes per post. 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. It is also not clear whether the increase in sponsorship is good or bad for welfare.

Overall, our findings suggest that in markets with no direct compensation mechanisms, regulations that distort indirect compensation mechanisms can have large and unanticipated effects.

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.

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. (2018). The market for influence. *Johns Hopkins Carey Business School Research*, pages 18–13.

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.

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.

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.

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

Mitchell, M. (2020). Free ad(vice): Internet influencers and disclosure regulation.

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*.

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.*

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.[22]

A.2 Manual Keywords Used For "Disclosed"

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

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

[22] 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).

- 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 20 Most Predictive Naive Bayes Features

- 'geht' - going
- 'code'
- 'morgen' - morning
- 'gibt' - give
- 'lieben' - love
- 'fashion'
- 'bio' - reference to "link in bio"
- 'immer' - always
- 'mehr' - more
- 'link'
- 'percent'
- 'einfach' - simple
- 'habt' - have
- 'happy'
- 'new'
- 'ootd' - outfit of the day
- 'mal' - time
- 'heute' - today
- 'love'
- 'schon' - beautiful

A.5 Dictionary Comparison between Pre-Regulation Germany and Spain and Most Predictive Post-Regulation Germany Words

Figure A1: Comparison of Dictionaries

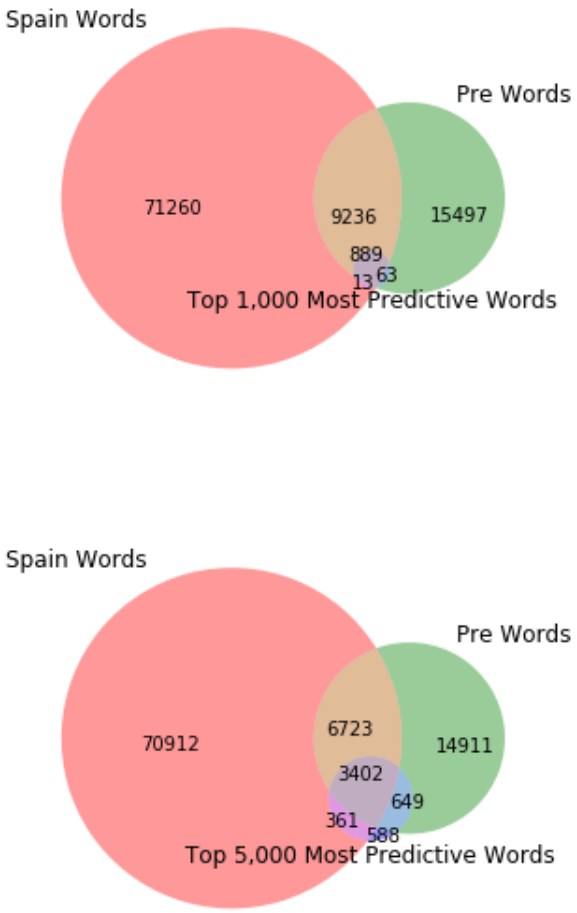


Figure A2: Timing Tests

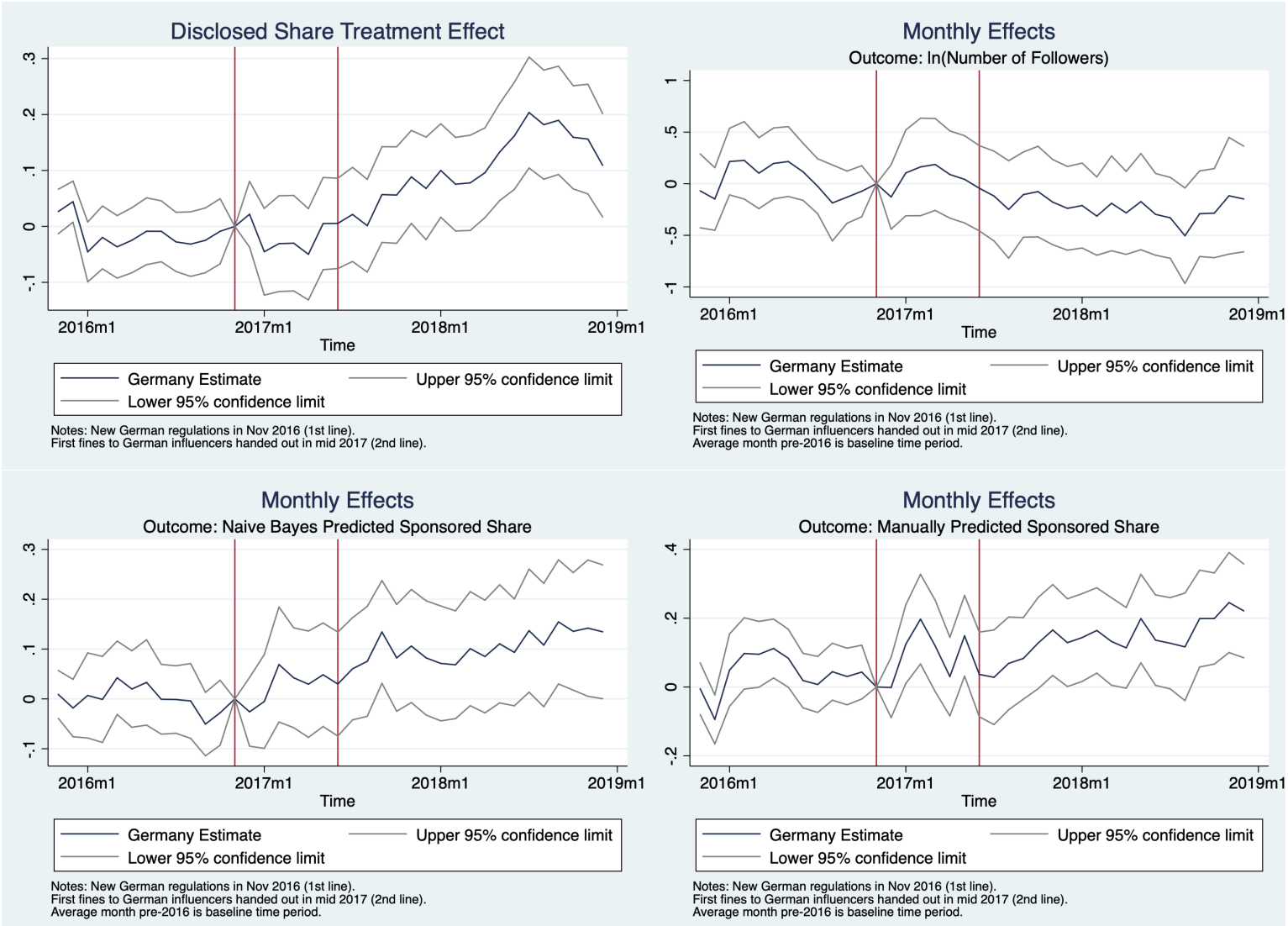


Table A1: Additional Influencer-Month Level Estimates

	(1)	(2)	(3)
Outcome Variable:	N Posts per Month	N Man. Spon. Posts per Month	N. NB Spon. Posts per Month
Germany	4.972	2.973	3.247**
× Treated Period	(3.447)	(2.057)	(1.330)
User FE	YES	YES	YES
Year-Month FE	YES	YES	YES
User Account Age FE	YES	YES	YES
User Account Age × User First Year FE	YES	YES	YES
Country Controls	YES	YES	YES
Observations	5,067	5,067	5,067
R-squared	0.658	0.680	0.677

Notes: The sample includes monthly observations of Instagram influencers in Germany and Spain with more than 2 posts per month from 2014 to 2019. "Germany × Treated Period" is a variable that is equal to 1 for all German Instagram influencers after November 2016. Additional controls include user fixed effects, year-month fixed effects and fixed effects controlling for account age interacted with the first year of the influencer's activity on Instagram. We also include country controls that include log(population), log(GDP per capita), and a measure of country-level Google Trends search intensity for the term "Instagram." Standard errors are clustered at the influencer level. *** p<0.01, ** p<0.05, * p<0.1.

Table A2: Balanced Panel Influencer-Month Level Estimates

	(1)	(2)	(3)
Outcome Variable:	Man. Spon. %	NB Spon. %	ln(N Followers)
Germany	0.053	0.061**	-0.276*
× Treated Period	(0.034)	(0.031)	(0.163)
User FE	YES	YES	YES
Year-Month FE	YES	YES	YES
User Account Age FE	YES	YES	YES
User Account Age × User First Year FE	YES	YES	YES
Country Controls	YES	YES	YES
Observations	3,187	3,187	2,448
R-squared	0.655	0.663	0.957

Notes: The sample includes monthly observations of Instagram influencers in Germany and Spain with more than 2 posts per month from 2014 to 2019 who are present for the entire sample period. "Germany × Treated Period" is a variable that is equal to 1 for all German Instagram influencers after November 2016. Additional controls include user fixed effects, year-month fixed effects and fixed effects controlling for account age interacted with the first year of the influencer's activity on Instagram. We also include country controls that include log(population), log(GDP per capita), and a measure of country-level Google Trends search intensity for the term "Instagram." Standard errors are clustered at the influencer level. *** p<0.01, ** p<0.05, * p<0.1.