Anticipated versus Actual Effects of Platform Design Change: A Case Study of Twitter's Character Limit

KRISTINA GLIGORIĆ, EPFL, Lausanne, Switzerland
JUSTYNA CZĘSTOCHOWSKA, EPFL, Lausanne, Switzerland
ASHTON ANDERSON, University of Toronto, Toronto, Canada
ROBERT WEST, EPFL, Lausanne, Switzerland

The design of online platforms is both critically important and challenging, as any changes may lead to unintended consequences, and it can be hard to predict how users will react. Here we conduct a case study of a particularly important real-world platform design change: Twitter's decision to double the character limit from 140 to 280 characters to soothe users' need to "cram" or "squeeze" their tweets, informed by modeling of historical user behavior. In our analysis, we contrast Twitter's anticipated pre-intervention predictions about user behavior with actual post-intervention user behavior: Did the platform design change lead to the intended user behavior shifts, or did a gap between anticipated and actual behavior emerge? Did different user groups react differently? We find that even though users do not "cram" as much under 280 characters as they used to under 140 characters, emergent "cramming" at the new limit seems to not have been taken into account when designing the platform change. Furthermore, investigating textual features, we find that, although post-intervention "crammed" tweets are longer, their syntactic and semantic characteristics remain similar and indicative of "squeezing". Applying the same approach as Twitter policy-makers, we create updated counterfactual estimates and find that the character limit would need to be increased further to reduce cramming that re-emerged at the new limit. We contribute to the rich literature studying online user behavior with an empirical study that reveals a dynamic interaction between platform design and user behavior, with immediate policy and practical implications for the design of socio-technical systems.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**; **Social media**; **User studies**; **Empirical studies in HCI**.

Additional Key Words and Phrases: platform design; Twitter; user behavior; predictability; demographics

ACM Reference Format:
Kristina Gligorić, Justyna Częstochowska, Ashton Anderson, and Robert West. 2022. Anticipated versus Actual Effects of Platform Design Change: A Case Study of Twitter's Character Limit. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 491 (November 2022), 29 pages. https://doi.org/10.1145/3555659

1 INTRODUCTION

Online platform design is, on the one hand, critically important, given that the online content can reach and affect people worldwide, and, on the other hand, very challenging, given the intrinsic entanglement between the digital environment and user behavior. The design of platforms impacts user behavior [38, 50], and user behavior, in turn, shapes the platforms [87]. It is challenging to

Authors' addresses: Kristina Gligorić, EPFL, Lausanne, Switzerland, kristina.gligoric@epfl.ch; Justyna Częstochowska, EPFL, Lausanne, Switzerland, justyna.czestochowska@epfl.ch; Ashton Anderson, University of Toronto, Toronto, Canada, ashton@cs.toronto.edu; Robert West, EPFL, Lausanne, Switzerland, robert.west@epfl.ch.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.
2573-0142/2022/11-ART491 $15.00
https://doi.org/10.1145/3555659

predict how users will respond to design changes and new platform policies as any changes to large complex socio-technical systems may lead to unintended consequences.

Here we examine an instance of a particularly important real-world platform design change: on 7 November 2017, Twitter suddenly and unexpectedly increased the maximum allowed tweet length from 140 to 280 characters, thus altering its signature feature. According to Twitter, this change, which we henceforth refer to as "the switch", was introduced to give users more space to express their thoughts, as a disproportionately large fraction of tweets had been exactly 140 characters long [28, 72], reflecting users' need to "cram" or "squeeze" their tweets.

The design of this platform change and the choice of the new policy was informed by modeling historical user behavior [39]. When deciding how to design the platform change, Twitter engineers and policy-makers made estimates and predictions based on data from before the intervention. In particular, to estimate users' need to "cram" their tweets, they estimated the number of tweets impacted by the policy, that is, the number of tweets that would be longer than the character limit if they could be longer. The new policy, 280 characters, was selected since it was estimated that, if that limit were enforced [39], a negligibly low fraction of tweets would afterwards be impacted by cramming.

Twitter's decision to double the maximum allowed tweet length is, therefore, a remarkable example of a platform design intervention that was informed by modeling based on publicly available historical user traces. However, it is known that there are factors that render anticipating the impact of real-world interventions challenging. Forecasts by impartial experts and policy-makers often fail to predict the outcomes of behavioral interventions [21, 54]. Users routinely violate expectations and norms, at the individual and the population level [12], interventions can elicit backfire effects [41, 83], and, although practically useful, predictions of online behaviors do not necessarily allow understanding of the phenomena being predicted [80].

Further challenges arise given the inherent limits to the predictability of human behaviors, offline and online, at an individual and at a population level. When it comes to *offline behaviors*, across several domains, there are short-term and long-term limits to predictability [70, 76, 82, 95]. When it comes to *online behaviors*, despite an unprecedented volume of information about users, content, and historical behaviors, there are limits to prediction in the complex social systems we engage with [51]. In studying online behaviors further challenges arise given the complex interaction between humans and the platforms we use. Human behaviors both influence and are influenced by the design and presence of technological platforms [87].

Twitter engineers and policy-makers, by their own account [39], fitted a model to historical user behavior traces and implicitly assumed that the user behavior would not change in response to the implementation of the change. Such a static, "no-response" view might or might not hold. Is it necessary to account for user response, or does the actual user behavior remain faithful to what was anticipated?

Given the aforementioned challenges of modeling online human behaviors and anticipating the impact of policy changes, and the fact that a remarkable platform design intervention intended to impact the behavior of millions of users was made based on predictions informed by the historical user behavior modeling, we perform a case study of Twitter's policy change. We ask: *Did the intervention materialize as intended? Did the predictions regarding anticipated user behavior hold? How did users adapt to this platform change?* While the effectiveness of the implemented intervention can be analyzed in relation to its impact on a wide range of user behaviors, we focus on cramming—the very behavior that led Twitter to change their defining feature by doubling the character limit.

Reducing cramming was important for Twitter policy-makers since it was hypothesized that users' need to cram the tweets to fit the limit was creating friction to post, which in turn was

suspected to be linked with users abandoning their tweets and ultimately leaving the platform [28, 72]. Beyond the importance for Twitter (one of the leading social media platforms [67], with content posted there reaching and affecting billions of people across the globe), examining cramming behavior and the impact of the implemented intervention on cramming is broadly of importance for human–computer interaction and social media studies for three key reasons.

First, when designing socio-technical systems and implementing platform changes, it is important to know whether changes can be implemented based on static analyses (as performed by Twitter) or if dynamical reasoning about how users will respond is necessary. In this regard, the findings and implications from the case of Twitter could be widely applicable to researchers and practitioners designing other policy changes and other platforms that allow the production of textual content.

Second, a deeper understanding of users' cramming behavior on Twitter is necessary, regardless of Twitter's specific policy change motivations. It is known that the policies that a platform imposes affect the audience not only through the content delivered over the platform, but also through the characteristics of the platform itself, or, in a mantra coined by Marshall McLuhan [52], "the medium is the message". On Twitter, the imposed length limit leads to cramming, affecting the linguistic style [42, 96] and the success of the tweets, measured through the received engagement [28–30, 77, 85, 88–90]. Therefore, cramming has implications for the nature and quality of discourse on the platform and the degree to which messages are likely to spread [28]. Nobody in whose research Twitter plays a role should ignore this platform change; all researchers should consider how the new length limit impacts users' cramming behavior, the content, and Twitter as a platform. Since cramming is associated with the linguistic features of the message and its success, when analyzing or modeling the textual content posted on Twitter, researchers need to know whether, after the switch, the retrieved messages are still "compressed" by the users or not.

Third, understanding cramming behavior is consequential beyond Twitter. Cramming is inherent in writing text under any length constraint, online and offline. As such, cramming should be understood when studying any textual content produced under a character limit, given that the limit policy is bound to shape the content [28]. Insights about cramming behavior and its variation across subpopulations of users depending on their languages and devices are relevant to researchers studying any form of textual content produced under a length limit.

1.1 Research questions

In a case study of Twitter's policy change, studying an instance of platform design change, we aim to shed light on how this change has impacted user behavior. Now that the length limit has been changed and it is possible to tweet longer than 140 characters, we can compare estimates made before the intervention with how user behavior has actually evolved. We aim to fill the gaps in the existing literature through the following core research question:

> **RQ:** Did the introduction of the 280-character limit lead to the intended user behavior shifts, or did a gap between anticipated and actual behavior emerge?

In particular, studying user modeling that informed design change on a major platform, we investigate the following dimensions of the above question:

- **RQ1** *Gaps between predictions and emerging user behavior:* Did the platform design intervention address the problem of cramming as intended?
- **RQ2** *Cramming across languages:* How are different user populations affected by the same global platform design change?
- **RQ3** *Fluidity of counterfactual estimates:* How do the predicted policy effects diverge after user behavior shifts?

1.2 Contributions

In this work, we report a novel empirical study of a design decision that shaped a socio-technical system and impacted millions of users. Our results reveal a dynamic interaction between platform design and user behavior, with immediate policy and practical implications for the design of socio-technical systems. We call for the development of cautious approaches that aim to consider multiple factors when designing a platform change, including the dynamic nature of user response and the impact the change will have on different populations depending on users' languages and devices.

1.3 Summary of main findings

Using a 1% sample of all tweets spanning the period from 1 January 2017 to 31 October 2019, we model tweet length over time (illustrated in Fig. 3). We find that gaps between anticipated and actual user behavior emerged after the intervention (RQ1). Initially, the estimated amount of cramming at 140 characters was aligned with actual user behavior (Fig. 4). However, actual user behavior eventually diverged from anticipated user behavior. While cramming at 140 characters sharply decreased after the introduction of 280 characters, cramming, although less drastic, shifted to the new length limit. Furthermore, examining tweet text, syntactic (Fig. 5 and 6) and semantic (Fig. 7) indicators also provide evidence of cramming that emerged at the new length limit. Overall, these gaps between anticipated and actual user behavior are more pronounced on the Web interface compared to mobile devices.

Studying how different user populations are affected by the global platform design change (RQ2), we find that cramming in a language at 140 characters before the switch is correlated with cramming in that language at 280 characters after the switch (Fig. 9), indicating that Twitter is used differently across languages, or that some languages might need more or fewer characters to express the same amount of information.

Finally, we consider hypothetical interventions that would reduce the cramming that emerged post-intervention (RQ3). We find that as user behavior shifts in response to a platform change, the estimated effects of hypothetical policies change as well (Fig. 10). Given that 280 characters were selected to make a vast majority of tweets fit the limit, post-intervention, since the cramming re-emerged, the necessary number of characters would have to be increased further to achieve the same objective (Fig. 11).

1.4 Implications

Our case study has immediate implications for platform design and future platform changes. Our results emphasize a dynamic interaction between platform design and user behaviors. The length limit was doubled in order to reduce users' need to cram their tweets. However, the intervention did not entirely solve the issue, as cramming re-emerged at the new length limit (Fig. 4). Before the intervention, the modeled data was "collected under the policy", with character limit shaping the data that Twitter subsequently based their measurements on. When modeling in such a static regime, the validity of estimates is threatened, and it is complicated to evaluate alternative policies before their deployment. As the new policy is deployed, new behaviors and new data are collected under a different policy which elicits behaviors that are not anticipated—although can be explained after the fact. This fusion of platform design decisions and user behaviors can lead to feedback loops and calls for more cautious approaches that take into account the dynamic nature of user response. These findings highlight the fluidity of online behaviors and have direct implications for large-scale user behavior studies, human–computer interaction, and platform design.

2 BACKGROUND AND RELATED WORK

2.1 Previous work studying Twitter's character limit change

Early existing work that studied Twitter users' attitudes toward the new 280-character limit [71] discovered varying initial reactions ranging from anticipation, surprise, and joy to anger, disappointment, and sadness. Early studies also revealed a low initial prevalence of long tweets immediately following the switch [66] and studied the short-term impact that the length limit had on linguistic features and engagement [7, 28], finding that in response to a length constraint, users write more tersely, use more abbreviations and contracted forms, and use fewer definite articles.

The impact of the switch was also studied in the specific context of political tweets [40], showing that doubling the length limit led to less uncivil, more polite, and more constructive discussions online. Whereas these early studies necessarily had to consider short-term effects, much less is known about the long-term effects of the switch on user behavior and about whether Twitter's intention was achieved or not after the implemented intervention. The present paper constitutes the first attempt to bridge this gap with a long-term study spanning several years.

2.2 Counterfactual policies and user responses

Platform design decisions are often shaped by data-driven user studies [1, 5, 19]. Broadly, one can think about data-driven user studies by considering the degree to which the study is *explanatory* (i.e., focusing on identifying and estimating causal factors of user behavior), and the degree to which the study is *predictive* (i.e., focused on predicting user behavior). Hofman et al. [36] propose integrating predictive and explanatory modeling, for example, by analyzing proposed counterfactual policies and interventions, quantifying impacts on specific behaviors in the short and long run, ahead of their implementation.

Machine learning models are often employed for both predictive and explanatory purposes. However, existing machine learning models typically rely on the assumption that the data, after deployment, resembles the data the model was fit on. As machine learning models are increasingly used to make consequential decisions and users react to the deployed models, this assumption is violated. The reactions create challenges to the deployment of machine learning algorithms in the real world. Studying such reactions, referred to as "strategic feedback", has enabled new avenues of machine learning research [34, 55] that takes into account the feedback of the environment. Behavioral economics aims to understand and model people's strategic responses as well [57]. Twitter's effort to design a real-world platform intervention based on estimates and predictions derived from historical user behavior represents an example of integrative data-driven study that can be potentially be impacted by the users' reactions.

However, despite the advances in understanding and incorporating the feedback of the environment, in practice, a static view is often assumed. Twitter engineers and policy-makers modeled historical user behavior traces, apparently assuming that user behavior would not change in response to the intervention. It is unknown whether assuming such a static view is justified or not. We aim to fill this gap by answering the question of whether it is necessary to account for user responses or not (cf. RQ1). Our analyses of whether the intervention led to intended user behavior shifts will have implications for the future real-world development of integrative user modeling approaches.

2.3 Twitter communication and supporting features: studies of use and conventions

Previous work has extensively studied communication taking place on Twitter and the specific features that support it, most importantly: retweets [8], hashtags [61, 92], quotes [27], and emojis [63]. Previous work has also investigated linguistic conventions on Twitter, the patterns of their

emergence [44, 45], how users align to them in conversations [22], how they diffuse [11, 13], and how they continuously evolve [18]. Additionally, previous work has studied how patterns of adoption of supporting features, as well as the linguistic style used on the platform more broadly, varies across numerous dimensions [79], including gender [15], political leaning [84], age and income [25]; but also within accounts [16].

Still, little is known about how users adapted to the new character limit feature in the long run, as opposed to in the short term. How did the different subpopulations of users change their communication patterns in response to the intervention (cf. RQ2)? We contribute to the literature in usage and conventions by studying how brevity conventions and norms evolve after the introduction of the new character limit feature.

2.4 Message framing and its linguistic features

Focusing specifically on the linguistic feature of the character length, previous work has studied how the imposed length constraint on Twitter and other microblogging platforms affects the dialogues and the linguistic style [42, 96], and the success of the message measured through the received engagement [28–30, 77, 85, 88–90]. More broadly, HCI, philology, communication, marketing, education, and psychology scholars have investigated conciseness and its benefits in many different contexts [47, 81, 86]. The length limit can be thought of as a constraint that determines the format of the content that can be produced. Such constraints are of interest to scholars since they are often thought to have a positive impact on the quality of produced creative content [9, 26, 35, 43, 53, 56].

Beyond length, numerous previous studies have investigated the question of what wording makes messages successful in online social media, often formulated as the task of predicting what makes textual content become popular [6, 32, 48]. In the specific case of Twitter, in addition to characterizing how language is used on the platform in general [23, 37, 49, 59], researchers have investigated the correlation of linguistic signals and user's features with the propagation of tweets [2, 3, 31, 62, 85].

Yet, little is known about the long-term impact of the implemented intervention on the linguistic features of tweets. How did users modify their messages in adaptation to the new character limit feature, and are there still indicators of cramming (cf. RQ1)? How many characters are needed to reduce the cramming, if any, after the intervention (cf. RQ3)? We aim to provide new perspectives on how the new length limit impacts the linguistic characteristics of the messages after the intervention, in the long run. Given this active area of research, it remains important to understand how the shifts in platform design impact the language on the platform.

3 MATERIALS AND METHODS

3.1 Data

We base our user modeling on the publicly available 1% sample of tweets, spanning the period between 1 January 2017 and 31 October 2019, available on the Internet Archive.[1] We consider only original tweets (i.e., we discard retweets).[2] We study the 23 biggest languages: three languages where the switch did not happen (Japanese, Korean, and Chinese), and 20 where it happened, each language with more than 2M tweets in total. The switch did not happen in Japanese, Korean, and Chinese because the 140-character limit was not as restrictive there as in other languages, since more information can be conveyed with the same number of characters [72, 73]. We focus on the

[1] https://archive.org/details/twitterstream

[2] Our main analyses study tweets in isolation. However, Twitter users have long been working around the character limit by splitting a long piece of text into a sequence of length-compliant tweets. We consider threaded tweets in a supplementary analysis (Appendix A).

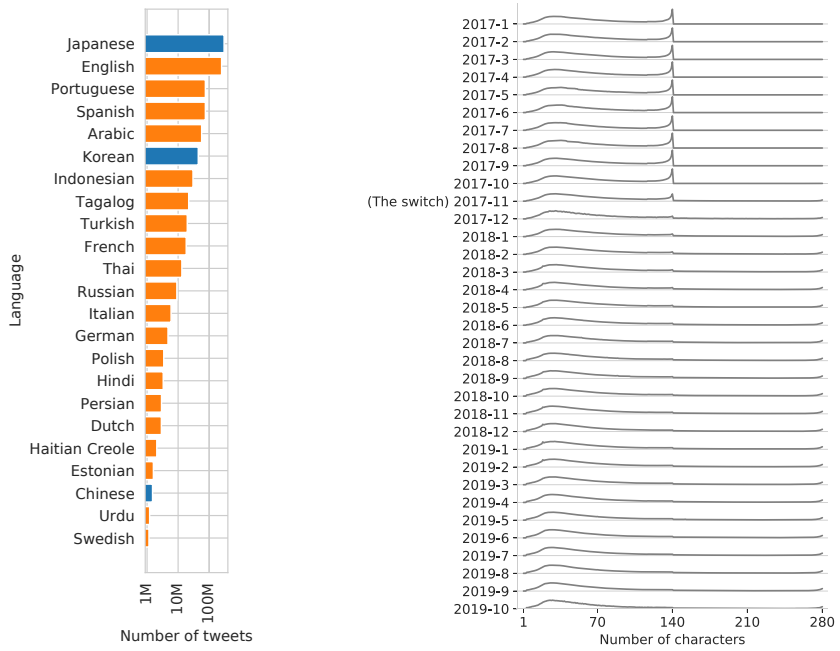


Fig. 1. **Twitter dataset statistics.** (a) Number of original tweets in the 1% sample posted between 1 January 2017 and 31 October 2019, across the 23 studied languages where the switch happened (*orange*) and did not happen (*blue*). (b) Normalized monthly tweet length histograms.

most common sources of tweets: the Web interface and mobile applications, as specified by the *source* field present in the collected tweet objects. We keep tweets posted by Twitter applications for iPhone, Android, iPad, Windows Phone, and their Lite versions and mobile Web clients and desktop Web clients. We disregard tweets posted by all other unofficial and automated sources and third-party applications as a proxy for bots (9.5% of original tweets are discarded on average across studied languages).

With the above, there are between 1M and 1.5M daily original tweets. In Fig. 1a we show the exact number in total across languages. We note that the posts are sampled at the community level. We stay at describing the community-level behaviors as opposed to user-level behaviors since user-level information is incomplete (in expectation, we have 1% of tweets posted by a fixed user).

3.2 Tweet length: counting characters

The focus of our study is the character limit, its change, and the impact on user behaviors, captured via the length of the posted tweets. To that end, we carefully count the number of characters based on the official documentation.[3] Tweet length is counted using the Unicode normalization of the tweet text. The tweet text is selected from the tweet object using the *displayed text range* field, discarding any characters not counted towards the length limit.

The text content of a tweet could contain up to 140 characters (or Unicode glyphs) before the switch, and 280 after the switch. An emoji sequence using multiple combining glyphs counts as

[3]https://developer.twitter.com/en/docs/counting-characters

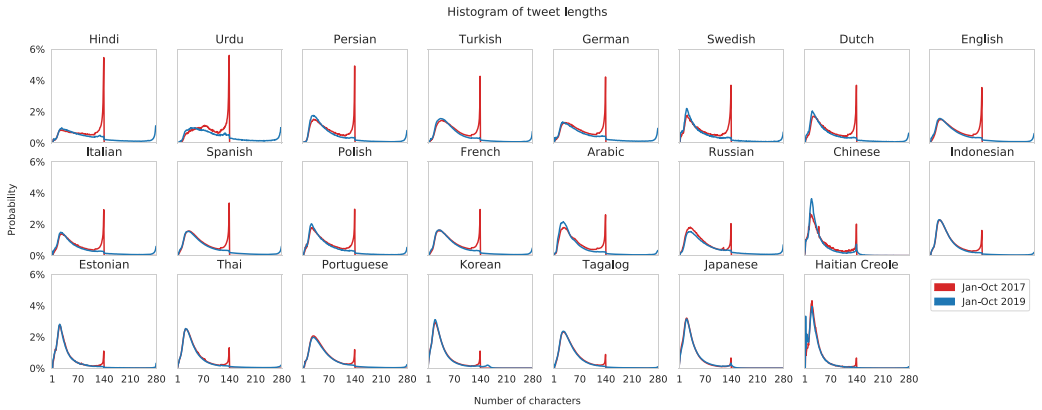


Fig. 2. **Tweet-length distributions for 23 languages,** for the periods before (*red*) and after (*blue*) the 280-character limit was introduced. Languages sorted by prevalence of 140 characters before the switch.

multiple characters. Chinese, Japanese, or Korean glyphs were counted as one character before, and as two characters after, the introduction of the 280-character limit. Therefore, a tweet composed of only Chinese, Japanese, or Korean text can only have a maximum of 140 of these types of glyphs after the switch.

We illustrate the prevalence of different tweet lengths over time (the fraction of tweets of a given number of characters) in Fig. 1b, and for individual languages in Fig. 2. We note that the first peak in the tweet length distribution, consistently between 25 and 30 characters, remains unchanged (i.e., the mode of the distribution is stable). Before the doubling of the character limit, we observe a drastic peak in the prevalence of 140-character tweets, reflecting users' need to cram their tweets [28, 72]. After the character limit doubling, we observe a sharp decline of 140-character tweets and an increase in 280-character tweets. Similar to the overall view, across languages the first peak and the mode of the distribution are constant, and the interesting character length ranges are near 140 and near 280 characters, where tweets are impacted by cramming.

3.3 Tweet length: modeling the impact of the character limit

3.3.1 Background. Before implementing the character length change, Twitter engineers and policy-makers aimed to answer the following questions [39]: "*Are 140 characters an adequate limit for all languages? Do people Tweeting in Japanese have too much space? Do those Tweeting in English not have enough? How often were people trying to craft Tweets which ended up over 140 characters? And by how much?*" Before the intervention, these questions seemed impossible to answer since there were no Tweets in Twitter's database over 140 characters long. However, Twitter policy-makers analyzed historical tweeting behavior to find answers [39].

First, Twitter's analyses revealed that when people write text within a certain length constraint, the resulting text lengths follow a log-normal distribution.

Second, it was discovered that the empirical distribution of tweet lengths deviates from the log-normal distribution near the character limit, since users try to "squeeze" their messages. To quantify the impact of the enforced character limit and other hypothetical character limits on users' behavior, two quantities were introduced: *size of cramming* and *size of run-over*. The estimated size of run-over is a crucial quantity, since it is used to evaluate hypothetical policies, as put by Twitter:

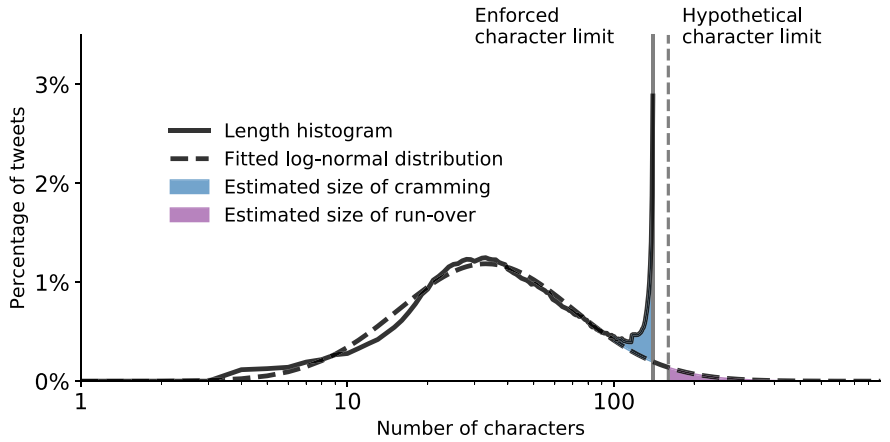


Fig. 3. **Illustration of model for measuring impact of character limit.** The example depicts the length distribution of English tweets across mobile and Web devices, before the 280-character limit was introduced. The solid black line shows the length histogram; the dashed black line marks the fitted log-normal distribution (logarithmic x-axis). The solid gray vertical line marks the enforced character limit (140 characters in the example). The dashed gray vertical line marks a hypothetical character limit (150 characters in the example). *Cramming* (marked as the blue area) is the deviation of the empirical distribution from the underlying log-normal distribution near the character limit. *Run-over* (marked as the purple area) is the part of the fitted log-normal distribution that falls beyond the given number of characters. Cramming can only be used to measure people's attempts to "squeeze" their tweets within the *enforced character limit* (solid vertical line). On the contrary, run-over is used to evaluate the impact of *any hypothetical character limit* (dashed vertical line).

"*9% of English Tweets, and 0.4% of Japanese Tweets, do not make it under 140 characters. And we will need 274 characters to make 99% of English Tweets viable as-is*".

Third, it was assumed that when a tweet ends up being more than 140 characters long, with probability p, a user deletes some characters proportional to the number of characters exceeded. With probability $1 - p$, the user abandons the tweet. Then, with probability q, the truncated tweet constitutes a valid sentence. The user sends the tweet if it is under 140 characters at this point. Otherwise, the process is repeated. Finally, with these factors were put together, Twitter developed a model that accurately approximates the tweet length distribution with cramming [39]. Twitter concluded that more characters are necessary to fit all tweets and that increasing the length limit might increase the fraction of users who post tweets, or, as stated by Twitter [39]: "*It did seem reasonable to expect that people will Tweet more if there is less friction to Tweet. This looked like something worth trying.*"

In the present work, our goal is to study the above user modeling that informed Twitter's design change. To that end, we carefully implement the estimation of the size of cramming and the size of run-over based on the official public documentation outlining the modeling approach [39]. We ensure that Twitter's reported pre-intervention estimates in the case of English and Japanese tweets are reproduced. Then, we monitor the daily temporal evolution of these quantities, across devices and across languages. In what follows, we explain in more detail how these quantities are calculated and what they capture.

3.3.2 Estimating the impact of the enforced character limit: the size of cramming. The enforced character limit impacts users' behavior, pushing people to "squeeze" their tweets within the allowed

number of characters [28]. Due to such "squeezing", the empirical distribution of tweet lengths deviates from the log-normal distribution near the character limit. Cramming at a given number of characters (illustrated in Fig. 3, blue area) is, therefore, the deviation of the actual length distribution from the theoretical log-normal distribution near the character limit. The underlying log-normal distribution is found by fitting a curve to the empirical tweet-length density excluding the tail emerging from cramming, using the least-squares objective. The tail exclusion is done at a point referred to as cramming threshold. The cramming threshold is found heuristically by selecting the rightmost local minimum of the tweet length density curve.

To summarize, cramming measures people's attempts to "squeeze" their tweets within the enforced character limit. Cramming is interpreted as the fraction of tweets impacted by the enforced character limit. Since the character limit was increased in order to reduce cramming [73], now, after the character limit was increased, we monitor the estimated size of cramming to understand the consequences of the implemented intervention.

3.3.3 Estimating the impact of hypothetical character limits: the size of run-over. Cramming is used to measure the impact of the enforced character limit. However, cramming cannot readily be used to measure the impact of hypothetical limits surpassing the currently enforced limit, since the value of the length distribution at hypothetical unsupported tweet lengths is zero. Run-over at a given number of characters (illustrated in Fig. 3, purple area) is, therefore, defined as the part of the fitted log-normal distribution that falls beyond a given number of characters. Run-over captures the fraction of tweets that would have more characters than the hypothetical length limit if it were possible.

To summarize, run-over captures the fraction of tweets that would be impacted by *any hypothetical character limit.* Ahead of the switch, run-over was used by Twitter engineers and policy-makers to evaluate the impact of counterfactual policies and to select the implemented policy (280 characters). Now, after the policy was changed, we monitor the estimated size of run-over to explore the evolution of estimates and the impact of hypothetical character limits.

4 RESULTS

4.1 RQ1: Gaps between predictions and emerging user behavior

Recall our guiding research question: Are there gaps between predicted and emerging user behavior? Concretely, did the platform design intervention address the problem of cramming as intended? Having access only to tweets tweeted before the switch, the fraction of tweets impacted by the limit (i.e., cramming) at 140 characters was estimated. Now, after the switch, we can compare the estimated size of cramming at 140 characters before the switch with the actual fraction of tweets that are longer than 140 characters after the switch. Contrasting these two quantities allows us to measure the gaps between estimates made before the intervention and the actual user behavior after the intervention.

To do so, we fit the model of tweet lengths (Sec. 3) for each day of the studied period (between 1 January 2017 and 31 October 2019) across all tweets in the 20 studied languages where the 28-character limit was introduced.[4] We then monitor how the cramming size estimate evolves over time (Fig. 4). We consider the Web interface and mobile devices separately (11% of tweets are sent from the Web interface vs. 89% from mobile devices).

[4]We take advantage of the fact that the new character limit was not introduced in Chinese, Japanese, and Korean to perform an estimation of the effect of the switch on tweet lengths that accounts for possible global platform-wide changes that are not associated with the doubling of the character limit intervention (Appendix B).

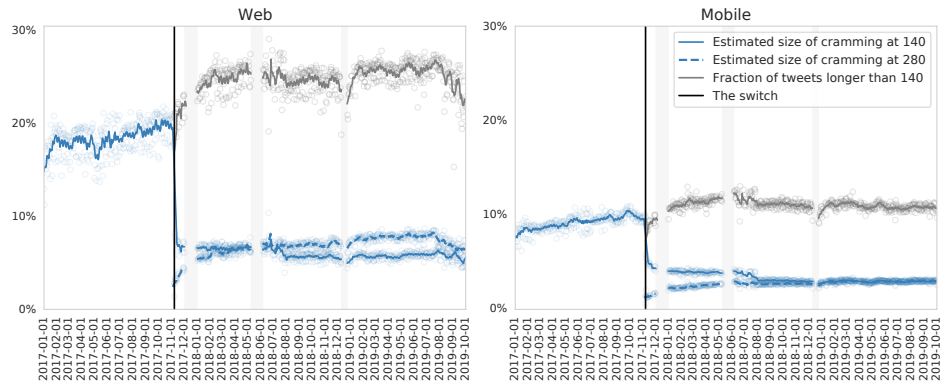


Fig. 4. **Estimated vs. actual fraction of tweets impacted by the length limit.** In blue, the estimated size of cramming, i.e., the estimated fraction of tweets impacted by the length limit, and, in gray, the fraction of tweets longer than 140 characters. Daily quantities are indicated with a circle, and the line marks a 10-day rolling average. The quantities are shown separately for the Web interface (left) and mobile applications (right). The vertical line marks the switch and gray bands mark days with missing data. Immediately after the switch, the fraction of tweets longer than 140 characters (solid gray line) closely mirrored the estimated size of cramming (solid blue line). As time passed, usage of long tweets diverged from estimates made before the intervention. After the switch, cramming at 140 characters drastically decreased (solid blue line), while cramming at 280 characters emerged (dashed blue line).

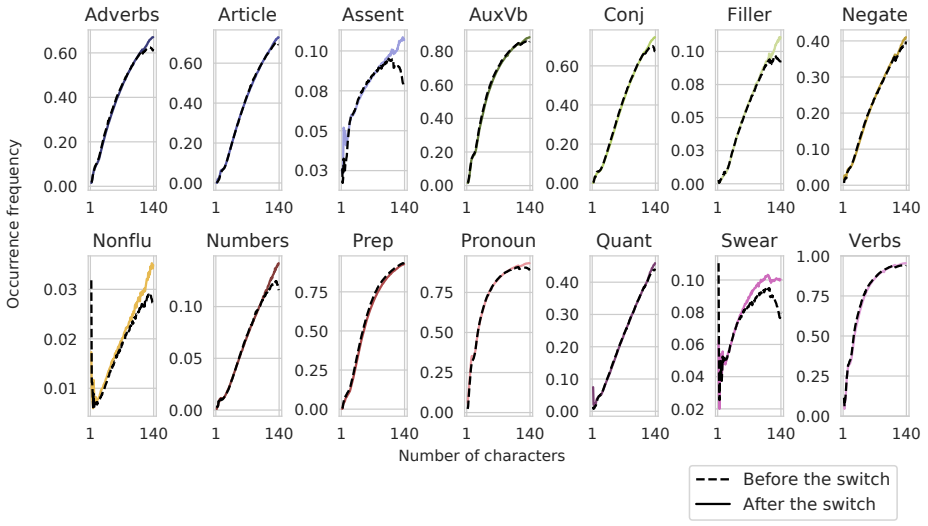
4.1.1 Web interface. Focusing on the Web interface (Fig. 4, left), before the switch, the estimated fraction of tweets impacted by the length limit, i.e., the estimated size of cramming at 140 characters, was, on average across days, 18.35% (95% bootstrapped CI [18.15%, 18.55%]). After the switch, the actual fraction of tweets longer than the previously imposed limit was, on average, 24.81% [24.66%, 24.95%].

Although, in the first weeks after the 280 character limit was introduced, the actual fraction closely mirrored the prior estimates, as time passed, users' tweeting behavior drifted away from the cramming estimates that had informed the policy change, and stabilized after around six months.

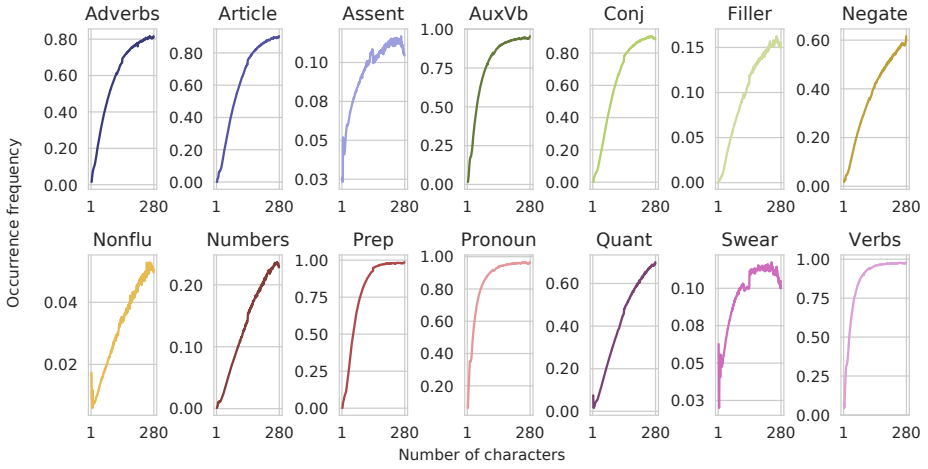
Furthermore, cramming at 140 characters drastically decreased after the switch, while cramming at 280 characters emerged. Before the switch, the estimated size of cramming at 140 characters was, as stated above, 18.35% on average across days. After the switch, the estimated size of cramming at 280 characters is 6.88% [6.8%, 6.96%]. While cramming at 140 characters sharply decreased after the introduction of 280 characters, cramming, although less drastic, shifted to the new length limit. Cramming at the new length limit slowly increased and also stabilized after around six months after the intervention.

The introduction of 280 characters reduced cramming at the respective limit by 62.5% (from 18.35% to 6.88%). However, a substantial fraction of tweets (around 1 in 15) is still impacted by the 280-character limit on the Web interface. By all we know, the emergence of cramming at the new length limit was not anticipated in the modeling approach that informed the platform change (Sec. 3).

4.1.2 Mobile devices. When it comes to mobile devices (Fig. 4, right), we observe that actual user behavior diverged from the cramming estimates made before the switch less drastically than in the case of the Web interface (see above). Similarly, although some cramming emerged at the new limit, it is less pronounced compared to cramming on the Web interface. The estimated size of

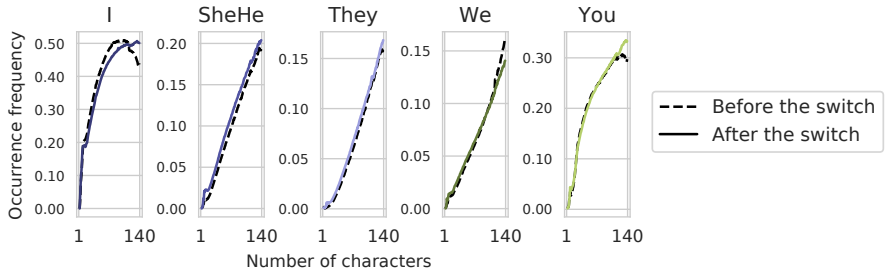


(a) Before vs. after the switch

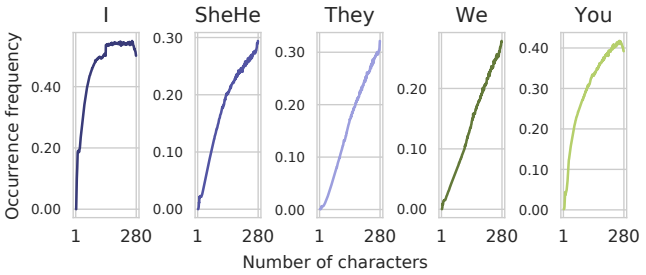


(b) After the switch

Fig. 5. **Syntactic indicators of cramming: part-of-speech (POS) tag frequency across tweet lengths.** (a) Occurrence frequency of POS tags across tweets of different character lengths in the period before the switch vs. after the switch (1–140 characters). The dashed black line represents this quantity across tweets posted in the period before the switch (i.e., under the 140-character limit), and the solid colored line in the period after the switch (i.e., under the 280-character limit). The largest gap is observed for swear words and spoken categories (nonfluencies, fillers, and assent), adverbs, and conjunctions, all nonessential parts of speech that are frequently deleted in the process of "squeezing" a message to fit a length limit. No gap is observed for verbs and negations, essential parts of speech. (b) Occurrence frequency of POS tags across tweets posted in the period after the switch (i.e., under the 280-character limit) of different character lengths (1–280 characters). Among the 280-character tweets after the switch, we observe patterns similar to those associated with 140-character tweets before the switch, e.g., a dip in the frequency of spoken categories, conjunctions, and numbers—traces typical of cramming or "optimizing" a message to fit a length limit. POS tags are sorted alphabetically. Note the different x- and y-axis scales.



(a) Before vs. after the switch



(b) After the switch

Fig. 6. **Syntactic indicators of cramming: personal pronoun frequency across tweet lengths.** (a) Occurrence frequency of personal pronouns across tweets of different character lengths in the period before the switch (1–140 characters). The dashed black line represents this quantity across tweets posted in the period before the switch (i.e., under the 140-character limit), and the solid colored line in the period after the switch (i.e., under the 280-character limit). (b) Occurrence frequency of personal pronouns across tweets posted in the period after the switch (i.e., under the 280-character limit) for different lengths (1–280 characters). Personal pronouns are sorted alphabetically. Note the different y-axis scales.

cramming at 140 characters before the switch was 9.06% [8.97%, 9.15%], whereas after the switch, 10.88% [10.82%, 10.93%] of tweets were longer than the newly imposed limit.

Compared to the Web interface, there was 50% less cramming on the mobile devices before the switch. Similarly, little cramming emerged at the new limit. Only 2.55% [2.52%, 2.58%] (or 1 in 40) tweets are impacted by the 280 character limit on mobile devices. That is, cramming is a smaller problem on mobile devices in general. Whereas on the Web interface cramming was underestimated, and more cramming emerged at the new limit, mobile devices paint a different picture. There, user behavior remained closer to estimates made before the intervention, and little cramming emerged at the new limit.

To summarize, the modeling approach underestimated the fraction of tweets that would be longer than the 140-character limit. Also, cramming now emerging at 280 was apparently not considered in the modeling approach performed before the switch. Cramming was a smaller issue on mobile devices before the intervention, compared to the Web interface, and remained less problematic after the intervention.

4.1.3 Gaps across languages. These insights are robust across languages (Fig. S2 and Fig. S3). The actual usage of long tweets surpassed estimated cramming the most in Hindi, Urdu, and Russian (Fig. S2). Cramming at 280 characters, smaller in size compared to cramming at 140 characters, and

more pronounced on the Web interface compared to the mobile devices, emerged in all languages (Fig. S3).

4.1.4 Syntactic indicators of cramming. In what follows, we aim to understand further the cramming that emerged at 280 characters after the switch by studying its impact on linguistic features of the tweets. Is cramming also evident in the text of tweets? Why do users post long tweets? What are their signature characteristics, and are they indicative of cramming as revealed via length modeling?

To answer these questions, we study tweets in English posted from mobile and Web devices during the pre- (75.56M tweets) and post-switch (65.29M tweets) periods. We annotated the tweets with LIWC's [65] syntactic features (linguistic categories) and semantic features (psychological, biological, and social categories).

First, to characterize syntactic features of tweets, for all tweets with a given number of characters, we measure the occurrence frequency of part of speech (POS) tags among tweets of that length. In Fig. 5a, across all possible tweet lengths in the period before the switch (1–140 characters), we observe the fraction of tweets of that length that have at least one instance of the respective POS tag. The dashed black line represents this quantity across tweets posted in the period before the switch (i.e., under the 140-character limit), and the solid colored line represents this quantity among tweets posted in the period after the switch (i.e., under the 280-character limit).

Comparing the solid colored and dashed black lines across different POS tags lets us isolate the effect of the length limit on the content of tweets. The largest gap is observed for swear words and spoken categories (nonfluencies, fillers, and assent), adverbs, and conjunctions—nonessential parts of speech that are frequently deleted in the process of "squeezing" a message to fit a length limit. No gap is observed for verbs and negations, essential parts of speech that are known to be disproportionately preserved in the cramming process [29].

Fig. 5b represents the same quantities after the switch (i.e., under the 280-character limit) across all possible tweet lengths in this period (1–280 characters). We note that there is no counterfactual observation here, i.e., we do not know what the probability of observing a POS tag among 280-character tweets would be if the 280-character limit was lifted. Nonetheless, among the 280-character tweets after the switch, we do observe patterns similar to those associated with the 140-character tweets before the switch, in particular, a dip in the frequency of the spoken categories, conjunctions, and numbers among the 280-character tweets, traces typical of cramming or "optimizing" a message to fit a length limit.

In Fig. 6a and Fig. 6b we measure the same quantities for fine-grained subtypes of personal pronouns. This suggests that personal pronouns *I* and *you* were most affected by the 140-character limit (i.e., they were most likely to be omitted before the switch), and we observe a similar non-monotonic pattern around 280 characters in Fig. 6b.

To summarize, 280-character tweets from after the switch are *syntactically* similar to 140-character tweets from before the switch, following patterns indicative of cramming [29]. This is evidence indicating that they are generated by similar writing processes as 140-character tweets used to be, and provides further evidence of emerging cramming behavior at the new length limit.

4.1.5 Semantic indicators of cramming. In Fig. 7 we next explore topics of tweets, as measured by LIWC categories describing psychological, biological, and social categories. Across the studied topics, for each allowed tweet length in the pre-switch and post-switch periods, we measure the factor by which the topic is more frequent among tweets of a given length, compared to all tweets across lengths. Categories are sorted by the value of this factor at 140 characters before the switch, and at 280 characters after the switch.

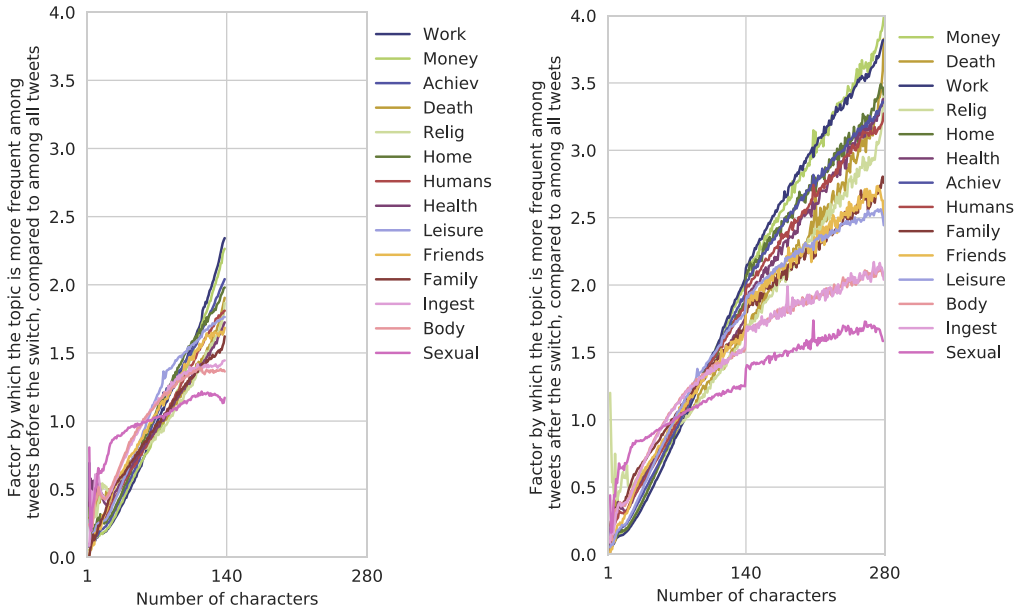


Fig. 7. **Semantic indicators of cramming: topics across tweet lengths.** For each length before the switch (left) and after the switch (right), across 14 topics measured with LIWC categories [65], we monitor the factor by which the topic is more frequent among tweets with that number of characters, compared to tweets across all lengths. Categories are sorted by the value of this factor at 140 characters (left) or 280 characters (right).

We note that the personal-concerns categories were relatively most frequent around 140 characters before the switch: *Work*, *Money*, *Achievement*, *Death*, and *Religion*. The least frequent topics at 140 characters, on the other hand, were the biological categories: *Sexual*, *Body*, *Ingestion*, followed by *Family*, *Friends*, and *Leisure*. An apparent association with the importance of the message emerges: while tweets about topics related to ordinary, overall more prevalent everyday experiences are long the least frequently, topics related to more serious personal concerns are long most frequently. There is a within-topic correlation between usage of 140 characters before the switch, and subsequent usage of 280 character length after the switch, with the ranking of topic usage at the boundary length only slightly changed (Spearman's rank correlation between topics 0.91, $p = 7.30 \times 10^{-6}$), implying that 280-character tweets from after the switch are also semantically similar to 140-tweets from before the switch. This is further evidence indicating that they are generated by similar processes as 140-character tweets used to be.

4.2 RQ2: Cramming across languages

Given that the switch constitutes a global platform design change that affected widely different user populations, we aim to further understand its impact across the globe. In particular, we study the temporal evolution of cramming and contrast how different user populations are affected across languages.

In Fig. 8, we monitor the evolution of cramming at the new limit across languages, for tweets posted from Web and mobile devices. In most of the languages, the cramming seems to have settled,

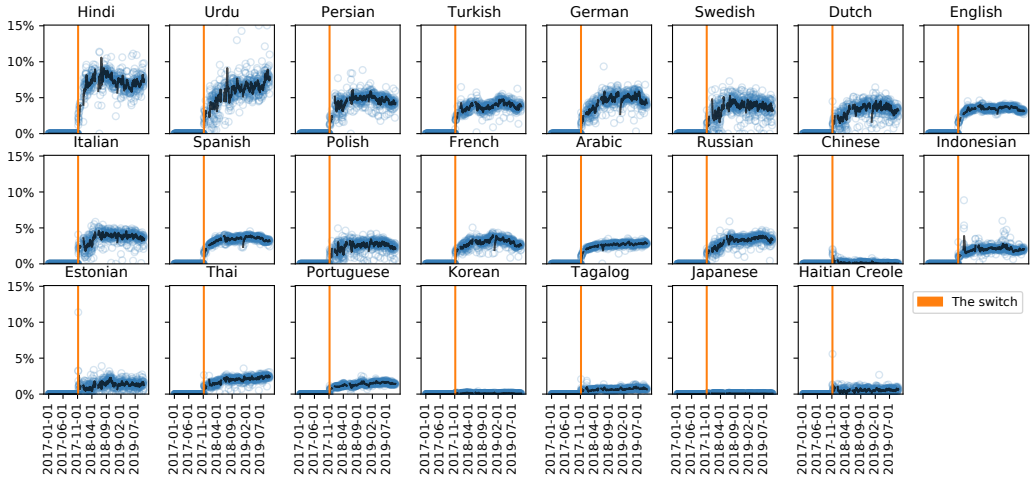


Fig. 8. **Daily evolution of the estimated size of cramming at 280 characters, across languages.** Daily fraction (indicated with a circle) and 10-day rolling average (solid line) of the estimated size of cramming. Languages sorted by size of cramming at 140 characters before the switch.

and is even decreasing again in some languages. Urdu is a notable exception, where the size of cramming is still growing and not yet in a stable state.

In Fig. 9, we show the estimated size of cramming at 280 characters after the switch (x-axis) vs. the estimated size of cramming at 140 characters before the switch (y-axis). We observe that the estimated size of cramming at 140 before the switch in a language is highly correlated with the estimated size of cramming at 280 after the switch: the more cramming there was in a language at 140 before the switch, the more cramming there is at 280 after the switch (Spearman's rank correlation 0.90, $p = 5.47 \times 10^{-9}$). In Hindi and Urdu, there is particularly much cramming at 280 after the intervention, to such an extent that 280 characters is the most frequent tweet length after the switch (Fig. 2).

In summary, we observe that the more cramming there was at 140 in a language before the switch, the more cramming there is at 280 after (Fig. 9). Disaggregation across languages reveals interesting temporal patterns (Fig. 8), whereby in most languages the cramming seems to have settled and is even decreasing again, i.e., the peak of cramming is over.

4.3 RQ3: Fluidity of counterfactual estimates

4.3.1 Tweets that fit the new limit. The new policy (280 characters) was selected because it was estimated that if that limit were enforced, a negligibly low fraction of tweets would not fit the limit [39]. However, as examined up to this point, cramming emerged again at the new limit. Hence, we explore how the estimated impact of the character limit evolves as user behavior evolves, using the same modeling approach applied by Twitter before the switch.

In Fig. 10, we measure the estimated size of run-over at 280 characters. In the case of the Web interface before the switch, 3.76% [3.69%, 3.83%] of tweets were estimated to be longer than 280 characters if it were possible, while after the switch, this number increases to 7.77% [7.68%, 7.86%]. In the case of mobile devices before the switch, it was estimated that 0.97% [0.96%, 0.99%] of tweets would be longer than 280 characters if it were possible. After the switch, this number remains close, at 1.15% [1.14%, 1.16%].

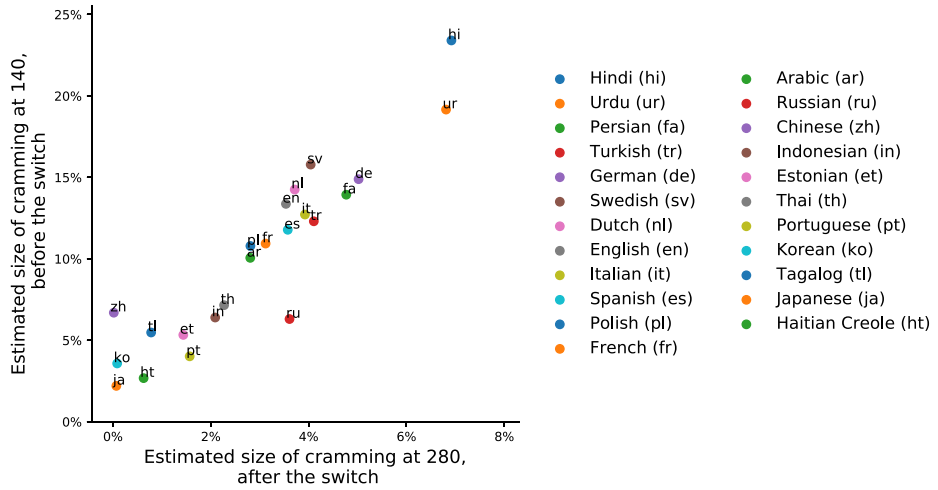


Fig. 9. **Cramming in 23 languages before vs. after the switch.** Estimated size of cramming at 280 characters after the switch (x-axis) vs. estimated size of cramming at 140 characters before the switch (y-axis) (Spearman's rank correlation 0.90, $p = 5.47 \times 10^{-9}$).

In summary, the estimated size of run-over at 280 characters more than doubled on Web interface (+107%) and slightly increased on mobile devices (+19%). Now, more tweets would be longer than 280 if they could be, implying that there is fluidity of estimates due to the fluidity of user behavior—both impacted by the currently imposed limit.

4.3.2 Necessary number of characters to fit all tweets under the new limit. Finally, we apply the same reasoning used before the switch and ask: For a targeted size of run-over, what character limit should be chosen? How many characters are necessary for a targeted fraction of tweets to fit the new limit? What future policies would reduce the cramming that emerged at the new character limit? If a new character limit were to be imposed to reduce cramming at 280, how many characters would it need to be?

In Fig. 11, we show the hypothetical tweet length limits necessary to achieve various targeted sizes of run-over. For instance, to make 95% of tweets fit (that is, to achieve 5% of run-over) on the Web interface before the switch, an estimated 275 (95% CI [273, 277]) characters would be necessary. After the switch, the number of characters would need to be increased to 342 [340, 344]. On mobile devices, there is no significant increase in the necessary number of characters to fit 95% of tweets (178 [177, 179] characters before vs. 175 [174, 177] characters after). Note that these estimates do not take into account the dynamic, fluid nature of user behavior in response to design change, but take the simpler, static view instead, mimicking the view taken by Twitter in their decision-making before the switch.

If, after the switch, 1% run-over (or fitting 99% of tweets) on the Web interface were desired, as many as 628 [623, 633] characters would be necessary. It is likely that, no matter how much one would increase the number of characters, the new limit would again lead to cramming, and the number of characters would have to be further increased to make more tweets fit the new limit.

Cramming is inherent to writing text under a character limit. When anticipating the impact of interventions, it is not sufficient to estimate how many tweets would be longer if the current limit

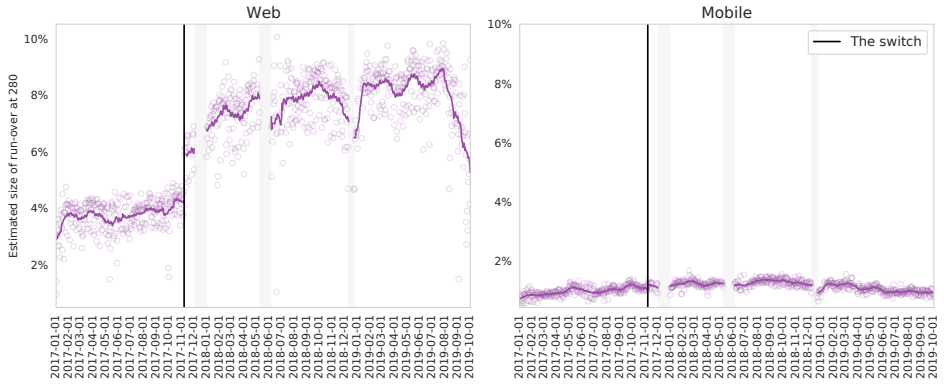


Fig. 10. **Evolution of estimated size of run-over at 280 characters.** Circles indicate the daily estimated fraction of tweets that would be longer than 280 characters if it were possible; the line marks the 10-day rolling average. The quantities are shown separately for the Web interface (left) and mobile applications (right). The vertical line marks the switch, and gray bands mark days with missing data. After the switch, the estimated size of run-over at 280 characters increases sharply on the Web interface, while it increases only slightly on mobile devices.

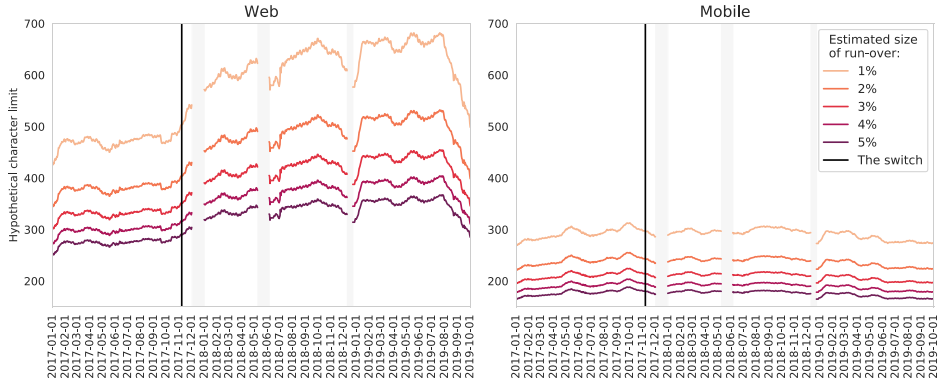


Fig. 11. **Evolution of hypothetical character limit necessary to achieve various targeted sizes of run-over.** Tweet length limits (in the number of characters) are shown on the y-axis; targeted sizes of run-over are marked in colors. Lines mark 10-day rolling averages. The quantities are shown separately for Web interface (left) and mobile applications (right). The vertical line marks the switch and gray bands mark days with missing data. After the switch, the number of characters needed to achieve a targeted run-over increases sharply on the Web interface, while it remains robust on the mobile devices.

did not exist (e.g., by modeling the run-over). It is necessary to also account for new cramming that would emerge given any other hypothetical limit.

5 DISCUSSION

5.1 Implications for the design of socio-technical systems and future platform changes

In this work, we analyze the aftermath of a design decision that shaped a socio-technical system and impacted millions of users. Twitter engineers and policy-makers modeled historical user behavior traces in order to design the platform intervention. By doing so, it was implicitly assumed that the

user behavior would not in other ways change in response to the implementation of the intervention. We find that the actual user behavior diverged from anticipated behavior since cramming emerged at the new character limit, and the usage of long tweets turned out to be higher than predicted (Fig. 4). Such a user response was apparently not taken into account when anticipating the effect of the intervention. Our findings suggest the need to account for user response and highlight the fluidity of online behaviors.

The new length limit (280 characters) was selected since it was estimated that, if that limit were enforced, a negligibly low fraction of tweets would be impacted by cramming [39]. After the intervention, using the same modeling methodology, we estimate that a further increase of the allowed number of characters would be necessary to achieve the same goal (Fig. 11). The evolution of this counterfactual estimate speaks about the fluidity of predictions and the limits to the predictability of user behaviors. The estimated effect of a policy changes as user response is taken into account. These findings fill the gaps between the literature on strategic feedback that takes into account the feedback of the environment [34, 55] and the static user modeling practices outlined in Sec. 2.2. Our findings reveal that although a static view is often assumed in practice, in the case of Twitter's character limit change, it was not justified. The analyses of the user response highlight that, moving forward, it is necessary to account for such responses of the users and the environment by developing novel dynamic and integrative user modeling approaches.

The major challenge of anticipating the impact of platform design changes is considering and incorporating user reactions. To model design change in the presence of users' adapting their behavior in response, more advanced, game-theoretic approaches may be required. Note that even experimental approaches such as A/B tests might fall short if not performed longitudinally over longer periods of time, allowing for users to adapt their behavior in response. An additional promising way forward may be to integrate sensitivity bounds into the estimates, in the spirit of how sensitivity analysis of a causal estimate is performed [64, 74, 75]. It is helpful to consider in advance how severe the users' dynamic response would need to be such that the predictions do not carry any statistical significance anymore. Domain knowledge can then be incorporated to reason about the plausibility of such a sufficiently drastic user response, ahead of the intervention.

5.2 Implications for Twitter research

Although the doubling of the length limit reduced the amount of cramming and eliminated the drastic disproportion of tweets reaching the maximum length (e.g., 9% of English tweets used to be exactly 140 characters long before the switch), our results demonstrate the emergence of a similar, though considerably weaker, effect around 280 characters after the switch. Hence the new character limit can be seen as a less intrusive version of the previous 140 character limit.

These findings have important implications for Twitter-based research, as they show that, although the new limit is "felt less" by users than the old limit, 280 characters still constitutes an impactful length constraint that shapes the nature of Twitter. The evidence suggests that, just as the old 140-character limit [28], the new 280-character limit impacts the writing style (Fig. 5 and Fig. 6) and content of tweets (Fig. 7) via cramming [78]. The length constraint and the resulting tweet-length distribution remain an important dimension to consider in studies using Twitter data [60]. After the switch, the number of characters remains an important variable, correlated with important user features including device, language, and topics. In a nutshell, "280 is the new 140", although it is less intrusive.

5.3 Differences across devices

We observe widely different patterns between Web and mobile devices (Fig. 4). On the Web interface, user behavior widely diverged from what had been anticipated, and considerable cramming

appeared at the new limit. On mobile devices, user behavior remained close to the predictions made before the intervention, and little cramming emerged at the new limit. We highlight this distinction between Web and mobile devices. This bimodal nature of Twitter as a platform should be carefully taken into account in future studies of online platforms.

Cramming and usage of long tweets are particularly prominent on Web clients. The fact that tweets were longer on the Web interface before the switch also indicates a tendency for shorter text on mobile phones. The degree to which the long tweet length is consistent with experiences and needs of users writing tweets can potentially explain the differences between Web and mobile devices. Typing shorter texts remains more compatible with small touchscreens on mobile devices [10]. Hence, users on mobile devices may experience less the need to "squeeze" their message.

5.4 Syntactic and semantic characteristics

Tweets of 280 characters are syntactically similar to 140-character tweets before the switch, following patterns indicative of cramming a message (Figures 5a and 5b). This is evidence indicating that tweets close to the new boundary are generated by similar writing processes as 140-character tweets were before the switch. There is a within-topic correlation between usage of 140 characters before the switch, and subsequent usage of 280 characters after the switch (Fig. 7). Beyond syntax, tweets of 280 characters are also semantically similar to 140-character tweets before the switch. While tweets about topics related to ordinary, overall more prevalent everyday experiences use the longer tweets the least frequently, topics related to more serious personal concerns use them the most. This apparent association with the importance of the message is aligned with previous work that described how Twitter's signature feature—brevity—is particularly pronounced in the context of political expression [69] and mental-health discussions [20], since longer messages are used for more complex sentences and ideas [33]. The fact that "important" topics were affected by cramming more than other topics before the switch, and are affected more after the switch, highlights the need to understand the implications of the implemented character limit intervention. Further considerations of policy changes on social platforms should take into account the effect they have on socially important discourses such as politics [38]. These findings fill gaps in the literature on Twitter communication and supporting features outlined in Sec. 2.3. In particular, while most of the previous work studied how users adapt to new features such as retweets, hashtags, and quotes in the short term [8, 27, 61, 63, 92], our results highlight the importance of analyzing the response in the long run. Our findings also fill gaps in the Twitter communication literature [15, 79, 84] by providing a novel characterization of adoption heterogeneity across topics and user subpopulations.

5.5 Differences across languages

The language-specific findings reveal notable differences between languages that were not a priori expected. Hindi and Urdu are the languages where there was most cramming at 140 characters before the switch, and where there is the most cramming at 280 characters after the switch (Fig. 8). These two languages, additionally, exhibit a different cramming evolution pattern compared to other languages, with the size of cramming still not in a stable state, but growing. This raises the question: What makes the cramming size different in those two languages? It is interesting to note that Hindi and Urdu are mutually intelligible as spoken languages. However, they are written in different scripts: Devanagari and a Perso-Arabic script, completely illegible to readers literate only in one of the two.

A possible explanation could be the fact that some languages might need more or fewer characters to express the same amount of information [17]. The fact that we observe that the cramming at 140 characters before the switch in a language is correlated with cramming at 280 after the switch

potentially points to information density remaining constant in a given language. Beyond possible language-specific reasons, cultural and broader societal values in complex ways influence the use of technology [91]. Future work should understand better the factor driving the observed differences between languages.

5.6 Limitations

This study suffers from limitations that the 1% stream of tweets is known to be susceptible to [93], as certain accounts might be over-represented due to the intentional or unintentional tampering with the Sample API [58, 68]. Due to the nature of Twitter's sampling mechanism, it is possible to deliberately influence the studied sample, the extent and content of any topic. Additionally, technical artifacts can skew Twitter's samples. Therefore, the 1% stream of tweets cannot be regarded as fully random [68]. Additionally, in our study, we focus on the most common sources of tweets: the Web interface and mobile applications. We disregard automated sources and third-party applications as a proxy for bots. However, bot detection can be more reliable using more sophisticated methods detecting bots that use regular applications [14, 46, 94], which we did not consider in this study.

We note that our syntactic and semantic analysis of tweets is limited to tweets in English only, due to the lack of available tools to support annotation consistently across all the studied languages. Future studies should measure these characteristics in other languages, using language-specific tools or machine translation. Finally, we note that in this study, we study Twitter as a platform (tweets are sampled at the community level), as opposed to users, whose timelines are incomplete in the 1% sample.

5.7 Future work

Future work should provide a better understanding of what user-specific features are associated with cramming behavior. Twitter hoped that increasing the length limit would reduce the friction to tweet and that changing the limit might therefore increase the fraction of users who post [39]. Future work should determine whether more users indeed tweeted due to the character limit intervention. Furthermore, future work should develop novel approaches for better modeling how design changes will impact online platforms. For instance, beyond static modeling based on historical data, future work might consider incorporating a game-theoretic analysis in order to anticipate how users might respond to platform design changes.

In the future, one could also focus on the question of age and ask: Are the users who have long been cramming under 140 the same ones who are more likely to cram under 280? Answering these questions requires data beyond the 1% sample, with complete records of users' tweets. However, here we caution against naïve comparisons, as careful quasi-experimental designs are necessary to truly isolate the effect of platform change and the effect of age on specific users [4]. User activity-level and age are correlated with other factors; e.g., users who stay longer on the platform might be in other ways fundamentally different from younger users who joined more recently (i.e., there is "survivor bias" [24]). Finally, our study should be replicated in the future as new platform design changes are implemented.

5.8 Ethical considerations

We perform a population-level study of user behavior that does not focus on any individual user. All analyses are based on highly aggregated daily statistics.

5.9 Code and data

Code and data necessary to reproduce our results at are publicly available at: https://github.com/epfl-dlab/anticipated-vs-actual.

6 CONCLUSION

Reflecting on our main research question, we find evidence that after the introduction of the 280-character limit, a gap between anticipated and actual behavior emerged. After the intervention, users started using long tweets more than anticipated and cramming emerged at the new length limit. Cramming is likely inherent to producing text under any length constraint and, as such, should always be taken into account when designing platforms that allow the production of textual content. The user modeling approach used by Twitter did not consider such shifts in user behavior as a response to the platform change. Moving forward, when anticipating the effect of platform changes, cautious approaches that aim to consider the dynamic interaction between platform design and user behavior, as well as the impact the change will have on different populations depending on users' languages and devices, are necessary.

ACKNOWLEDGMENTS

We are grateful to Léonore Guillain for early help with data analysis. The EPFL Data Science Lab acknowledges support from Microsoft (Swiss Joint Research Center), Swiss National Science Foundation (grant 200021_185043), Collaborative Research on Science and Society (CROSS), European Union (TAILOR, grant 952215), Facebook, and Google. Ashton Anderson acknowledges support from the Natural Sciences and Engineering Research Council.

REFERENCES

[1] Tony Ahn, Seewon Ryu, and Ingoo Han. 2007. The impact of Web quality and playfulness on user acceptance of online retailing. *Information & management* 44, 3 (2007), 263–275.

[2] Yoav Artzi, Patrick Pantel, and Michael Gamon. 2012. Predicting responses to microblog posts. In *Proc. Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*.

[3] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. 2011. Everyone's an influencer: Quantifying influence on Twitter. In *Proc. ACM International Conference on Web Search and Data Mining (WSDM)*.

[4] Samuel Barbosa, Dan Cosley, Amit Sharma, and Roberto M Cesar Jr. 2016. Averaging gone wrong: Using time-aware analyses to better understand behavior. In *Proceedings of the 25th International Conference on World Wide Web (TheWebConf)*.

[5] Elena Belavina, Simone Marinesi, and Gerry Tsoukalas. 2020. Rethinking crowdfunding platform design: mechanisms to deter misconduct and improve efficiency. *Management Science* 66, 11 (2020), 4980–4997.

[6] Jonah Berger and Katherine L Milkman. 2012. What makes online content viral? *Journal of Marketing Research* 49, 2 (2012), 192–205.

[7] Arnout B Boot, Erik Tjong Kim Sang, Katinka Dijkstra, and Rolf A Zwaan. 2019. How character limit affects language usage in tweets. *Palgrave Communications* 5, 1 (2019), 1–13.

[8] Danah Boyd, Scott Golder, and Gilad Lotan. 2010. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *2010 43rd Hawaii International Conference on System Sciences*. 1–10.

[9] Yann Bramoulle and Lorenzo Ductor. 2018. Title length. *Journal of Economic Behavior & Organization* 150 (2018), 311–324.

[10] Daniel Buschek, Alexander De Luca, and Florian Alt. 2015. Improving accuracy, applicability and usability of keystroke biometrics on mobile touchscreen devices. In *Proc ACM Conference on Human Factors in Computing Systems (CHI)*.

[11] Damon Centola, Joshua Becker, Devon Brackbill, and Andrea Baronchelli. 2018. Experimental evidence for tipping points in social convention. *Science* 360, 6393 (2018), 1116–1119.

[12] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proc. ACM Hum.-Comput. Interact. (CSCW)* 2 (2018).

[13] Hsia-Ching Chang. 2010. A new perspective on Twitter hashtag use: Diffusion of innovation theory. *Proceedings of the American Society for Information Science and Technology* 47, 1 (2010), 1–4.

[14] Nikan Chavoshi, Hossein Hamooni, and Abdullah Mueen. 2016. Debot: Twitter bot detection via warped correlation.. In *ICDM*. 817–822.

[15] Morgane Ciot, Morgan Sonderegger, and Derek Ruths. 2013. Gender inference of Twitter users in non-English contexts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[16] Isobelle Clarke and Jack Grieve. 2019. Stylistic variation on the Donald Trump Twitter account: A linguistic analysis of tweets posted between 2009 and 2018. *PloS one* 14, 9 (2019), e0222062.
[17] Christophe Coupé, Yoon Mi Oh, Dan Dediu, and François Pellegrino. 2019. Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances* 5, 9 (2019), eaaw2594.
[18] Evandro Cunha, Gabriel Magno, Giovanni Comarela, Virgilio Almeida, Marcos Andre Goncalves, and Fabricio Benevenuto. 2011. Analyzing the dynamic evolution of hashtags on twitter: a language-based approach. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*. 58–65.
[19] Michael Cusumano. 2010. The evolution of platform thinking. *Commun. ACM* 53, 1 (2010), 32–34.
[20] Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Proc. International AAAI Conference on Weblogs and Social Media (ICWSM)*.
[21] Stefano DellaVigna and Devin Pope. 2018. Predicting experimental results: who knows what? *Journal of Political Economy* 126, 6 (2018), 2410–2456.
[22] Gabriel Doyle, Dan Yurovsky, and Michael C Frank. 2016. A robust framework for estimating linguistic alignment in Twitter conversations. In *Proceedings of the 25th International Conference on World Wide Web (TheWebConf)*.
[23] Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proc. Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*.
[24] Edwin J Elton, Martin J Gruber, and Christopher R Blake. 1996. Survivor bias and mutual fund performance. *The Review of Financial Studies* 9, 4 (1996), 1097–1120.
[25] Lucie Flekova, Daniel Preoţiuc-Pietro, and Lyle Ungar. 2016. Exploring stylistic variation with age and income on twitter. In *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*.
[26] Boris Fritscher and Yves Pigneur. 2009. Supporting business model modelling: A compromise between creativity and constraints. In *Proc. International Workshop on Task Models and Diagrams*.
[27] Kiran Garimella, Ingmar Weber, and Munmun De Choudhury. 2016. Quote RTs on Twitter: usage of the new feature for political discourse. In *Proceedings of the 8th ACM Conference on Web Science (WebSci)*.
[28] Kristina Gligorić, Ashton Anderson, and Robert West. 2018. How constraints affect content: The case of Twitter's switch from 140 to 280 characters. In *Proc. International AAAI Conference on Web and Social Media (ICWSM)*.
[29] Kristina Gligorić, Ashton Anderson, and Robert West. 2019. Causal Effects of Brevity on Style and Success in Social Media. *Proc. ACM Hum.-Comput. Interact. (CSCW)* 3 (Nov. 2019).
[30] Kristina Gligorić, George Lifchits, Robert West, and Ashton Anderson. 2021. Linguistic effects on news headline success: Evidence from thousands of online field experiments (Registered Report Protocol). *Plos one* 16, 9 (2021), e0257091.
[31] Kristina Gligorić, Manoel Horta Ribeiro, Martin Müller, Olesia Altunina, Maxime Peyrard, Marcel Salathé, Giovanni Colavizza, and Robert West. 2020. Experts and authorities receive disproportionate attention on Twitter during the COVID-19 crisis. *arXiv preprint arXiv:2008.08364* (2020).
[32] Marco Guerini, Carlo Strapparava, and Gözde Özbal. 2011. Exploring text virality in social networks. In *Proc. International AAAI Conference on Weblogs and Social Media (ICWSM)*.
[33] Daniel Halpern and Jennifer Gibbs. 2013. Social media as a catalyst for online deliberation? Exploring the affordances of Facebook and YouTube for political expression. *Computers in Human Behavior* 29, 3 (2013), 1159–1168.
[34] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. 2016. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*. 111–122.
[35] Beth A Hennessey. 1989. The effect of extrinsic constraints on children's creativity while using a computer. *Creativity Research Journal* 2, 3 (1989), 151–168.
[36] Jake M Hofman, Duncan J Watts, Susan Athey, Filiz Garip, Thomas L Griffiths, Jon Kleinberg, Helen Margetts, Sendhil Mullainathan, Matthew J Salganik, Simine Vazire, et al. 2021. Integrating explanation and prediction in computational social science. *Nature* 595, 7866 (2021), 181–188.
[37] Yuheng Hu, Kartik Talamadupula, and Subbarao Kambhampati. 2013. Dude, srsly?: The surprisingly formal nature of Twitter's language. In *Proc. International AAAI Conference on Weblogs and Social Media (ICWSM)*.
[38] Ferenc Huszár, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt. 2022. Algorithmic amplification of politics on Twitter. *Proceedings of the National Academy of Sciences (PNAS)* 119, 1 (2022).
[39] Ihara Ikuhiro. 2017. Our discovery of cramming. https://web.archive.org/web/20220428070306/https://blog.twitter.com/engineering/en_us/topics/insights/2017/Our-Discovery-of-Cramming.
[40] Kokil Jaidka, Alvin Zhou, and Yphtach Lelkes. 2019. Brevity is the soul of Twitter: The constraint affordance and political discussion. *Journal of Communication* 69, 4 (2019), 345–372.
[41] Shagun Jhaver, Christian Boylston, Diyi Yang, and Amy Bruckman. 2021. Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter. *Proc. ACM Hum.-Comput. Interact. (CSCW)* 5 (oct 2021).

[42] Huiyuan Jin and Haitao Liu. 2017. How will text size influence the length of its linguistic constituents? *Poznan Studies in Contemporary Linguistics* 53, 2 (2017), 197–225.

[43] Caneel K Joyce. 2009. *The Blank Page: Effects of Constraint on Creativity.* PhD thesis, UC Berkeley.

[44] Farshad Kooti, Winter A Mason, Krishna P Gummadi, and Meeyoung Cha. 2012. Predicting emerging social conventions in online social networks. In *Proc. of the ACM International Conference on Information and Knowledge Management (CIKM).* 445–454.

[45] Farshad Kooti, Haeryun Yang, Meeyoung Cha, P Krishna Gummadi, and Winter A Mason. 2012. The Emergence of Conventions in Online Social Networks.. In *Proc. International AAAI Conference on Weblogs and Social Media (ICWSM).*

[46] Sneha Kudugunta and Emilio Ferrara. 2018. Deep neural networks for bot detection. *Information Sciences* 467 (2018), 312–322.

[47] Nevin Laib. 1990. Conciseness and amplification. *College Composition and Communication* 41, 4 (1990), 443–459.

[48] Sotiris Lamprinidis, Daniel Hardt, and Dirk Hovy. 2018. Predicting news headline popularity with syntactic and semantic knowledge using multi-task learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP).*

[49] Paul Levinson. 2011. The long story about the short medium: Twitter as a communication medium in historical, present, and future context. *Journal of Communication Research* 48 (2011), 7–28.

[50] Momin M Malik and Jürgen Pfeffer. 2016. Identifying platform effects in social media data. In *Proc. International AAAI Conference on Weblogs and Social Media (ICWSM).*

[51] Travis Martin, Jake M Hofman, Amit Sharma, Ashton Anderson, and Duncan J Watts. 2016. Exploring limits to prediction in complex social systems. In *Proceedings of the 25th International Conference on World Wide Web (TheWebConf).*

[52] Marshall McLuhan. 1964. The extensions of man. *New York* (1964).

[53] J McPhee. 2015. Omission: Choosing what to leave out. https://web.archive.org/web/20220710175229/https://www.newyorker.com/magazine/2015/09/14/omission. *The New Yorker* (2015).

[54] Katherine L Milkman, Dena Gromet, Hung Ho, Joseph S Kay, Timothy W Lee, Pepi Pandiloski, Yeji Park, Aneesh Rai, Max Bazerman, John Beshears, et al. 2021. Megastudies improve the impact of applied behavioural science. *Nature* 600, 7889 (2021), 478–483.

[55] John Miller, Smitha Milli, and Moritz Hardt. 2020. Strategic classification is causal modeling in disguise. In *International Conference on Machine Learning (ICML).*

[56] Page C Moreau and Darren W Dahl. 2005. Designing the solution: The impact of constraints on consumers' creativity. *Journal of Consumer Research* 32, 1 (2005), 13–22.

[57] John DW Morecroft. 2015. *Strategic modelling and business dynamics: A feedback systems approach.* John Wiley & Sons.

[58] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. 2013. Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's Firehose. *arXiv preprint arXiv:1306.5204* (2013).

[59] Dhiraj Murthy. 2012. Towards a sociological understanding of social media: Theorizing Twitter. *Sociology* 46, 6 (2012), 1059–1073.

[60] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data* 2 (2019), 13.

[61] Ruth Page. 2012. The linguistics of self-branding and micro-celebrity in Twitter: The role of hashtags. *Discourse & communication* 6, 2 (2012), 181–201.

[62] Ethan Pancer and Maxwell Poole. 2016. The popularity and virality of political social media: hashtags, mentions, and links predict likes and retweets of 2016 U.S. presidential nominees tweets. *Social Influence* 11, 4 (2016), 259–270.

[63] Umashanthi Pavalanathan and Jacob Eisenstein. 2016. More emojis, less:) The competition for paralinguistic function in microblog writing. *First Monday* (2016).

[64] Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect.* Basic books.

[65] James W Pennebaker, Roger J Booth, and Martha E Francis. 2007. LIWC2007: Linguistic inquiry and word count. *Austin, Texas: liwc. net* (2007).

[66] Sarah Perez. 2017. Twitter's doubling of character count from 140 to 280 had little impact on length of tweets. https://cutt.ly/gfoIaY1.

[67] Andrew Perrin and Monica Anderson. 2019. Share of US adults using social media, including Facebook, is mostly unchanged since 2018. *Pew Research Center* 10 (2019).

[68] Jürgen Pfeffer, Katja Mayer, and Fred Morstatter. 2018. Tampering with Twitter's sample API. *EPJ Data Science* 7, 1 (2018), 50.

[69] Shalini Priya, Ryan Sequeira, Joydeep Chandra, and Sourav Kumar Dandapat. 2019. Where should one get news updates: Twitter or reddit. *Online Social Networks and Media* 9 (2019), 17–29.

[70] R.T. Ramos, R.B. Sassi, and J.R.C. Piqueira. 2011. Self-organized criticality and the predictability of human behavior. *New Ideas in Psychology* 29, 1 (2011), 38–48.

[71] Rimjhim and Roshni Chakraborty. 2018. Characterizing User Reactions Towards Twitter's 280 Character Limit. In *Proceedings of the 10th Annual Meeting of the Forum for Information Retrieval Evaluation (FIRE)* (Gandhinagar, India). 48–51.
[72] Aliza Rosen. 2017. Tweeting Made Easier. https://blog.twitter.com/official/en_us/topics/product/2017/tweetingmadeeasier.html.
[73] Aliza Rosen and Ikuhiro Ihara. 2017. Giving you more characters to express yourself. https://blog.twitter.com/official/en_us/topics/product/2017/Giving-you-more-characters-to-express-y\ourself.html.
[74] Paul R Rosenbaum. 2005. Sensitivity analysis in observational studies. *Encyclopedia of statistics in behavioral science* (2005).
[75] Paul R Rosenbaum, PR Rosenbaum, and Briskman. 2010. *Design of observational studies*. Vol. 10. Springer.
[76] Matthew J Salganik, Ian Lundberg, Alexander T Kindel, Caitlin E Ahearn, Khaled Al-Ghoneim, Abdullah Almaatouq, Drew M Altschul, Jennie E Brand, Nicole Bohme Carnegie, Ryan James Compton, et al. 2020. Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences (PNAS)* 117, 15 (2020), 8398–8403.
[77] Carsten D Schultz. 2017. Proposing to your fans: Which brand post characteristics drive consumer engagement activities on social media brand pages? *Electronic Commerce Research and Applications* 26 (2017), 23–34.
[78] Indira Sen, Fabian Floeck, Katrin Weller, Bernd Weiss, and Claudia Wagner. 2019. A total error framework for digital traces of humans. *arXiv preprint arXiv:1907.08228* (2019).
[79] Allison Shapp. 2014. Variation in the use of Twitter hashtags. *New York University* (2014), 1–44.
[80] Benjamin Shulman, Amit Sharma, and Dan Cosley. 2016. Predictability of popularity: Gaps between prediction and understanding. In *Proc. International AAAI Conference on Web and Social Media (ICWSM)*, Vol. 10.
[81] Brenda S Sloane. 2003. Say it straight: Teaching conciseness. *Teaching English in the Two Year College* (2003).
[82] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. 2010. Limits of Predictability in Human Mobility. *Science* 327, 5968 (2010), 1018–1021.
[83] Briony Swire-Thompson, Joseph DeGutis, and David Lazer. 2020. Searching for the backfire effect: Measurement and design considerations. *Journal of Applied Research in Memory and Cognition* (2020).
[84] Karolina Sylwester and Matthew Purver. 2015. Twitter language use reflects psychological differences between democrats and republicans. *PloS one* 10, 9 (2015), e0137422.
[85] Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*.
[86] Amiel D Vardi. 2000. Brevity, conciseness, and compression in Roman poetic criticism and the text of Gellius' Noctes Atticae 19.9. 10. *American Journal of Philology* (2000).
[87] Claudia Wagner, Markus Strohmaier, Alexandra Olteanu, Emre Kıcıman, Noshir Contractor, and Tina Eliassi-Rad. 2021. Measuring algorithmically infused societies. *Nature* 595, 7866 (2021), 197–204.
[88] Timm F Wagner, Christian V Baccarella, and Kai-Ingo Voigt. 2017. Framing social media communication: Investigating the effects of brand post appeals on user interaction. *European Management Journal* 35, 5 (2017), 606–616.
[89] Shuting Ada Wang and Brad N Greenwood. 2020. Does Length Impact Engagement? Length Limits of Posts and Microblogging Behavior. *Length Limits of Posts and Microblogging Behavior (February 12, 2020)* (2020).
[90] Ben S Wasike. 2013. Framing News in 140 Characters: How Social Media Editors Frame the News and Interact with Audiences via Twitter. *Global Media Journal: Canadian Edition* 6, 1 (2013).
[91] Barbara Wejnert. 2002. Integrating models of diffusion of innovations: A conceptual framework. *Annual Review of Sociology* 28, 1 (2002), 297–326.
[92] Peter Wikström. 2014. # srynotfunny: Communicative functions of hashtags on Twitter. *SKY Journal of Linguistics* 27 (2014).
[93] Siqi Wu, Marian-Andrei Rizoiu, and Lexing Xie. 2020. Variation across Scales: Measurement Fidelity under Twitter Data Sampling. In *Proc. International AAAI Conference on Web and Social Media (ICWSM)*, Vol. 14. 715–725.
[94] Kai-Cheng Yang, Onur Varol, Pik-Mai Hui, and Filippo Menczer. 2020. Scalable and generalizable social bot detection through data selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 1096–1103.
[95] Kai Zhao, Denis Khryashchev, and Huy Vo. 2021. Predicting Taxi and Uber Demand in Cities: Approaching the Limit of Predictability. *IEEE Transactions on Knowledge and Data Engineering* 33, 6 (2021), 2723–2736.
[96] Alvin Zhou and Sifan Xu. 2019. Remaking dialogic principles for the digital age: The role of affordances in dialogue and engagement. *SocArXiv* (2019).

APPENDIX: SUPPLEMENTAL MATERIALS

A Threaded tweets

Our main analyses study tweets in isolation. However, Twitter users have long been working around the character limit by splitting a long piece of text into a sequence of length-compliant tweets. These tweet threads are usually connected with the supporting threading feature, or annotated with the position of the tweet in its sequence. They are not intentionally altered to fit the character limit requirement. We investigated user behavior a step beyond the single tweets to estimate the impact of the introduction of 280 character length on the length of threads connecting tweets.

For each thread length k, we estimate the expected number n_k of tweets from threads of length k as the empirical number of tweets ending with the conventional pagination "i/k":

$$n_k = \epsilon \cdot k \cdot m_k, \tag{S1}$$

where $\epsilon = 0.01$ is the sampling rate, and m_k is the total number of threads of length k. With this, we can estimate the total number of threads of length k as

$$m_k = \frac{n_k}{\epsilon \cdot k}. \tag{S2}$$

We then estimate the distribution of thread lengths before and after the 280 character limit was introduced (Figure S1), for tweets tweeted in one of the 20 studied languages where the switch occurred and posted from Web and mobile agents. Tweets without pagination are considered as threads of length $k = 1$. In Figure S1, we depict the estimated distribution of thread lengths, before the 280 limit was introduced, and after.

Although threading is not a prevalent behavior—fewer than 0.1% of all tweets are paginated [28]—, there is evidence of less splitting of long texts into sequences of length-compliant tweets after the introduction of the new limit. We observe that threaded tweets occurred more frequently before the 280 limit was introduced and that long threads were more common before the introduction. However, there are no drastic shifts in the distribution of thread lengths due to the introduction of the 280 limit. Hence, threading is not accounted for when modeling cramming and run-over in our main analyses.

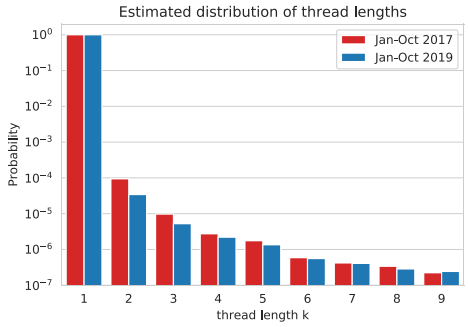


Fig. S1. **Threaded tweets, before vs. after the switch.** Estimated distribution of thread lengths, before the 280 limit was introduced (in red), and after it was introduced (in blue).

B The contrasting case of Chinese, Japanese, and Korean

Additionally, we take advantage of the fact that the new character limit was not introduced in all languages to perform a differences in differences estimation of the effect of the switch on tweet lengths.

To account for possible global platform-wide changes that are not associated with the switch, we use a differences in differences regression estimation, where the tweet lengths in Japanese, Korean and Chinese, languages where the 280 characters were not introduced are the control time series, and the tweet lengths in the other 20 studied languages (Figure 1a) are the treated time series. Both are observed in the pre-switch (Jan-Oct 2017), and post-switch (Jan-Oct 2019) periods. We fit a model

$$y \sim \texttt{treated} * \texttt{period}, \tag{S3}$$

where the dependent variable y is the logarithm of the average tweet length for each studied calendar day, and as independent variables are the following two factors: `treated` (indicates whether the switch was introduced or not in those languages), `period` (indicates whether a calendar day is in year pre-switch or post-switch). $\texttt{treated} * \texttt{period}$ is shorthand notation for $\alpha + \beta\texttt{treated} + \gamma\texttt{period} + \delta\texttt{treated} : \texttt{period} + \epsilon$, where in turn $\texttt{treated} : \texttt{period}$ stands for the interaction of `treated` and `period`.

The interaction term $\texttt{treated} : \texttt{period}$ δ is then the effect of switch on the logarithm of average tweet length. Each studied pre- or post-switch period spans 277 days per condition, amounting to a total of $4 \times 277 = 1108$ data points. The model is multiplicative due to the log. The relative increase over the baseline is then calculated by converting back to the linear scale the fitted coefficient δ. Fitting the model S3, we measure a $e^{\delta} - 1 = e^{0.0598} - 1 = 6.16\%$ (95% CI [5.68%, 6.64%]) increase in tweet lengths in the languages where the switch happened, over the control baseline. We note, however, that tweet lengths increased slightly in the control languages as well. This is likely impacted by the nature of how tweet length is counted at the character level, allowing mixed-character tweets to be longer than 140 characters.

To summarize, we estimate a significant increase in tweet lengths in languages where the new character limit was introduced, compared to the control languages, thus accounting for possible global platform-wide changes that are not associated with the doubling of the character limit intervention.

C Gaps between predictions and emerging user behavior across languages

Finally, for completeness, Figures S2 and S3 summarize the size of estimated cramming and the fraction of tweets longer than 140 characters across languages and devices.

Received January 2022; revised April 2022; accepted August 2022

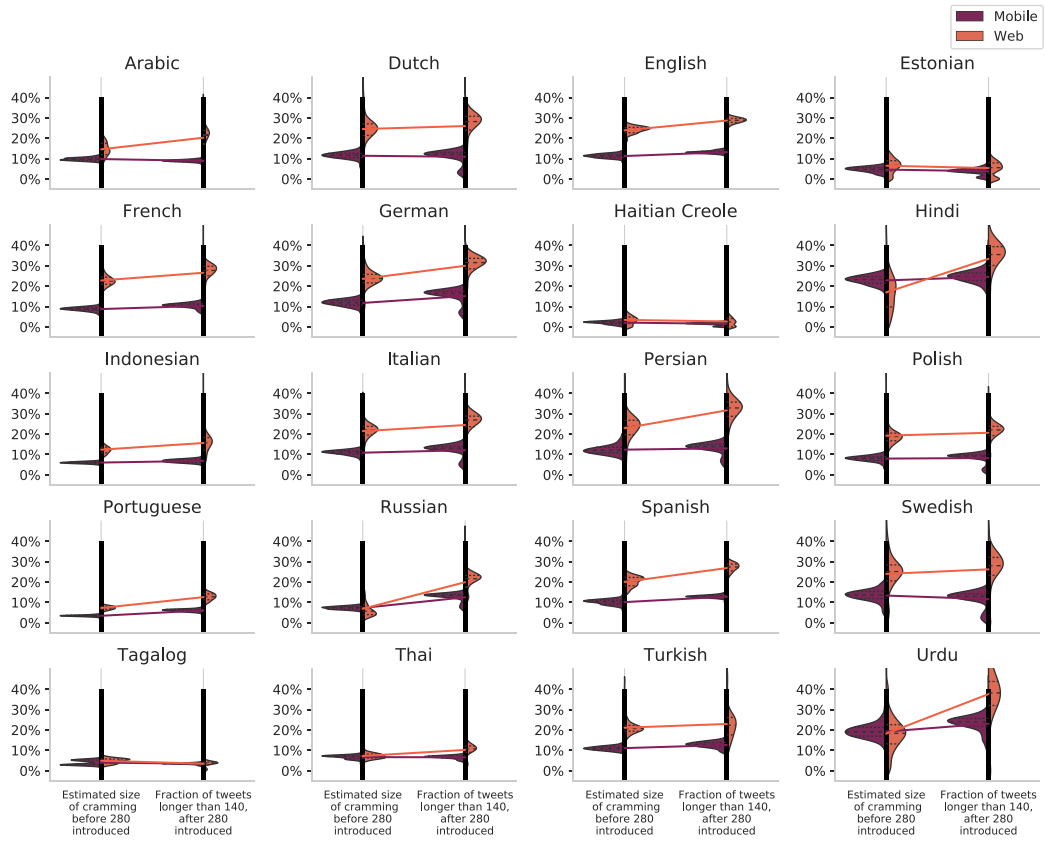


Fig. S2. **Estimated vs. actual fraction of tweets impacted by the length limit, across languages.** Across 20 languages where 280 character limit was introduced, the estimated size of cramming before the intervention (on the left) and the fraction of tweets longer than 140 characters after the intervention (on the right). The point distribution across all days is displayed. Horizontal dashed lines mark the quartiles. The quantities are shown separately for mobile (in purple) and Web (in orange). Lines connect the means for mobile (in purple) and Web (in orange). Languages are sorted alphabetically.

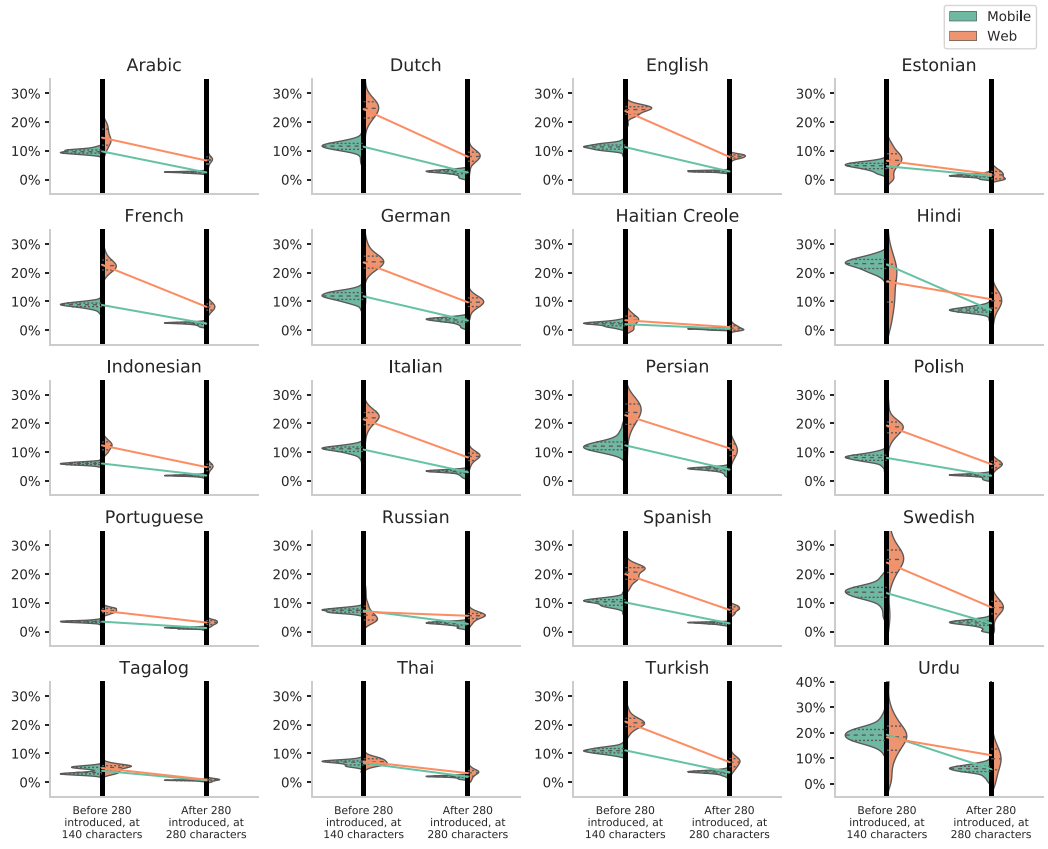


Fig. S3. **Cramming at the enforced character limit, across languages.** Across 20 languages where 280 character limit was introduced, the estimated size of cramming before the intervention at 140 characters (on the left) and the estimated size of cramming after the intervention at 280 characters (on the right). The point distribution across all days is displayed. Horizontal dashed lines mark the quartiles. The quantities are shown separately for mobile (in green) and Web (in orange). Lines connect the means for mobile (in green) and Web (in orange). Languages are sorted alphabetically.