PNAS RESEARCH ARTICLE | APPLIED PHYSICAL SCIENCES
PSYCHOLOGICAL AND COGNITIVE SCIENCES OPEN ACCESS

Patterns of linguistic simplification on social media platforms over time

N. Di Marco[a], Edoardo Loru[b], Anita Bonetti[c], Alessandra Olga Grazia Serra[d], Matteo Cinelli[a], and Walter Quattrociocchi[a,1]

Edited by Susan Fiske, Princeton University, Jamaica, VT; received June 17, 2024; accepted November 7, 2024

Understanding the impact of digital platforms on user behavior presents foundational challenges, including issues related to polarization, misinformation dynamics, and variation in news consumption. Comparative analyses across platforms and over different years can provide critical insights into these phenomena. This study investigates the linguistic characteristics of user comments over 34 y, focusing on their complexity and temporal shifts. Using a dataset of approximately 300 million English comments from eight diverse platforms and topics, we examine user communications' vocabulary size and linguistic richness and their evolution over time. Our findings reveal consistent patterns of complexity across social media platforms and topics, characterized by a nearly universal reduction in text length, diminished lexical richness, and decreased repetitiveness. Despite these trends, users consistently introduce new words into their comments at a nearly constant rate. This analysis underscores that platforms only partially influence the complexity of user comments but, instead, it reflects a broader pattern of linguistic change driven by social triggers, suggesting intrinsic tendencies in users' online interactions comparable to historically recognized linguistic hybridization and contamination processes.

social media | social dynamics | language evolution

Significance

Our study investigates the complexity of user comments across multiple digital platforms and topics, analyzing a large dataset of about 300 million English comments. Findings reveal a consistent trend of decreasing text length and lexical richness, yet users consistently introduce new words at a stable rate in their comments. This pattern suggests that linguistic changes on social media reflect a broader, universal aspect of human behavior rather than platform-specific influences. This insight contributes to understanding how digital environments shape communication norms and user engagement.

The rapid expansion of social media platforms has revolutionized how we connect and communicate, fundamentally altering the landscape of human interactions. These platforms have become integral to our daily lives as primary sources of information, entertainment, and personal communication (1–3). While they offer unprecedented opportunities for connectivity and interaction, they also intertwine entertainment-driven business models with complex social dynamics, raising substantial concerns about their impact on users and society at large (4). The influence of social media on public discourse and individual behavior has become a pressing concern within the scientific community, particularly regarding issues of polarization, misinformation (1, 4–7), and hate speech (8–10). Recent studies have explored their potential effects on user behavior, uncovering complex interactions and identifying key unresolved questions (4, 6, 11–13). They confirm that online users often select information that aligns with their preferences, overlook dissenting information, and form homophilic communities (5), which may influence their belief formation and communication methods. To grasp the evolving social dynamics in these processes, it is crucial to analyze the linguistic features of user-generated content, emphasizing the role of vocabulary and lexicon, which are integral to techniques like sentiment analysis of web texts (14–17). Moreover, the study of lexical variation, supported by empirical data and large corpora of texts, provides valuable insights into the social complexities of communication in digital environments (18).

Vocabulary scope measurements have posed a significant challenge in fields that extend beyond mere linguistic and communicative-discursive realms. This is particularly evident in psycho-cognitive domains, such as reading comprehension and information processing (19, 20).

While previous studies have extensively examined the impact of vocabulary size—e.g., on academic education—revealing substantial individual differences and underlying the pivotal role of lexical knowledge in educational settings (21), there is still a gap in investigating how these dynamics adapt to the digital era. Indeed, linguistics is increasingly focusing on social media, spurred by concerns that internet language may corrupt traditional writing practices and face-to-face communication (22–26).

However, despite extensive discourse, a systematic and metric interpretation of language complexity still needs to be improved. Current research often addresses specific elements of complexity but does not provide a holistic view, making it challenging to develop a unified conceptual and theoretical framework. The definition of complexity, even within the linguistic analysis, is often ambiguous (27), and the role of quantitative

Author affiliations: [a]Department of Computer Science, Sapienza University of Rome, Roma 00161, Italy; [b]Department of Computer, Control and Management Engineering, Sapienza University of Rome, Rome 00185, Italy; [c]Department of Communication and Social Research, Roma CAP 00198, Italia; and [d]Tuscia University - Dipartimento di studi linguistico-letterari, storico-filosofici e giuridici (DISTU) Department of Modern Languages and Literatures, History, Philosophy and Law Studies, Viterbo 01100, Italy

Author contributions: N.D.M. and W.Q. designed research; N.D.M., A.B., A.O.G.S., and W.Q. performed research; N.D.M. and E.L. analyzed data; and N.D.M., E.L., A.B., A.O.G.S., M.C., and W.Q. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: walter.quattrociocchi@uniroma1.it.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2412105121/-/DCSupplemental.

Published December 6, 2024.

metrics in assessing complexity has not been fully explored, with only limited studies addressing this approach (28, 29). For example, one approach rise in social physics, where statistical physics and network science methods are applied to study societal phenomena, could offer valuable insights (30). Scholarly debate on linguistic complexity tends to adopt two distinct perspectives. The first views complexity as a theoretical abstraction with little direct application to real-world linguistic scenarios. The second treats it as an empirical phenomenon that can be quantified and analyzed using theoretical tools. In ref. 31, the author introduces the concepts of "absolute complexity" and "relative complexity." Absolute complexity is considered an intrinsic property of language systems, independent of user interaction. In turn, relative complexity involves the user's perspective, measuring complexity as the cost or difficulty encountered by language users.

Our research primarily considers relative complexity, focusing on how it manifests in user interactions on social media platforms. Specifically, this study explores how users' language has evolved over time in online platforms. Within the globalized web environment, the parameters of intense language contact, dense social networks, large communities, and vast amounts of shared information act as sociolinguistic triggers that initiate and drive a progressive process of linguistic simplification. The analysis centers on the lexical level, because vocabulary serves as a crucial indicator of language evolution. Along this path, we analyze a large dataset of nearly 300 million English comments across eight major social media platforms—Facebook, Twitter, YouTube, Voat, Reddit, Usenet, Gab, and Telegram—covering nearly three decades and several topics. After having provided a linguistic theoretical background capable of explaining the obtained results, we begin by assessing the average vocabulary size of users according to their activity level. We then provide a model describing the evolution of vocabulary used by users, obtaining an estimate of the speed at which they reach their maximum vocabulary, and offering insights into how digital environments shape communication norms and user engagement. Finally, we map text complexity by applying two established lexical richness and repetitiveness metrics and we provide a statistical approach to model their evolution over several years.

Our results suggest that simplification processes may have occurred in online social media. We find consistent patterns of complexity across social media and topics, with a mostly universal reduction in text length, lexical richness, and repetitiveness.

Theoretical Background

Without delving into the evaluative implications of the internet's influence on modern linguistic practices, it is nonetheless important to note that certain effects are both evident and well documented, such as the emergence of novel linguistic forms like abbreviations, phonetic spellings, neologisms, and multimedia elements such as hashtags and emojis (32).

Before examining the central issues of this paper, a recognition of the concept concerning the linguistic simplification processes hereby analyzed is needed. As highlighted by Mühlhäusler (33, 34), simplification is a multifaceted concept encompassing at least three interrelated processes (35): the regularization of irregular patterns; the enhancement of lexical and morphological transparency; the reduction of redundancy. This happens particularly in contexts of language contact involving large community size, dense social networks (33, 36), and large amounts of shared information (37). While simplification and complexification may occur in all languages, there is an imbalance between the two in high-contact situations, with simplification dominant (38). This issue has been extensively studied, particularly in connection with urbanization, colonization, and standardization processes, attesting to the depletion of dialects, vocabulary reduction, and other connected phenomena (35, 39).

In these specific situations, the effect of simplification produces a decrease in the range of verb forms, a more consistent approach to pluralization, and the elimination or reduction of specific syntactical-grammatical categories (35). A pidgin may be partially or fully creolized, as is the case, for example, with English, which is a widespread vehicular language. Contemporary English has a relatively simplified grammar, especially compared to other Germanic languages. Romance languages, too, have undergone significant simplifications compared to the Latin from which they originated. Whether vehicular or not, modern languages seem more uncomplicated than their ancient predecessors (40).

Simplification should not be understood here as an objective concept but rather as a variation that results in communicative outcomes, perceived as more comprehensible by participants in comparison with other linguistic variants. Living languages continuously evolve, discarding some categories and forming new ones (41). This dynamic process often simplifies or reduces less common or "marked" constructs over time. Consequently, multidisciplinary studies have not converged on a single comprehensive definition nor reached a consensus on whether language change predominantly leads to simplification or increased complexity. Nevertheless, lexical depletion is recognized as a typical phenomenon in contemporary language (42).

Trudgill argues that widespread linguistic contact among adults is a relatively recent phenomenon, mainly post-Neolithic and predominantly modern, as is the formation of large, fluid communities (35). Although some languages remain more complex than others, the current diachronic trend points toward a growing number of languages becoming less complex over time: Many features that once illustrated complexification now vanish or are on the brink of disappearing (35). Notably, this trend toward simplification may be interpreted as a move toward leveling linguistic differences to facilitate communication, though it does not necessarily eliminate those variations. Simplification, in fact, often occurs when speakers interact and possess an incomplete understanding of each other's linguistic norms; thus, complexities that hinder communication tend to be discarded (43, 44).

In light of this conceptual premise, the similarities with the current communicative context of the web, particularly in social network interactions, appear evident. These platforms represent environments where all the conditions for dense linguistic contact and the need for reciprocal understanding, as previously discussed, are realized. Besides, complexity has gained significant interest in many disciplines over the past few decades, including linguistics, physics, biology, and mathematics (45–48). Grounded in this theoretical framework, our work aims to explore whether the conditions observed in online social media contribute to the phenomenon of linguistic simplification.

Users Vocabulary in Social Media

In the following sections, we conduct a comparative analysis of 8 different social media platforms: Facebook, Twitter, YouTube, Voat, Reddit, Usenet, Gab, and Telegram. Each social media contains comments related to different topics (see *Data Collection* for further details).

The Vocabulary of Users. To measure the vocabulary size on social media platforms, we aggregate all the comments from each user into one document. We then perform text preprocessing, including tokenization and stemming, to facilitate accurate

word counting (see further details about this procedure in *Materials and Methods*). In this work, we refer to tokens as instances of individual words as they appear in the text, whereas types represent distinct words without repetition. Therefore, we associate a couple of integer values with each user containing the number of tokens (i.e., the number of their total words) and the number of types (i.e., the number of their unique words). In this context, lexical richness refers to the variation in the number of their total words (tokens) and the number of their unique words (types).

Fig. 1 displays the complementary cumulative distribution functions (CCDF) of the number of tokens and types within each user's documents in each dataset.

The distributions display general consistency across different social media platforms and topics, albeit with varying magnitudes. Their behaviors are almost identical, with the primary differences manifesting in their tails, since tokens exhibit longer tails. This pattern aligns with the expectation that shorter texts typically have nearly identical numbers of words and unique words, whereas longer ones can present richer vocabularies. This observation is further supported by *SI Appendix*, Fig. S1, which illustrates the distributions of type-token ratios (TTR). These distributions predominantly peaked at 1, indicating a high similarity between types and tokens in most cases, with only a minority of users exhibiting lower TTR values. Nonetheless, consistently across various topics and social media, we observe a kind of exponential decay in the CCDF, suggesting that most users typically employ up to 10 unique words, i.e. a relatively small vocabulary size.

Obviously, this observation is somewhat dependent on the user activity, which is known to follow a heavy-tailed distribution (11) (i.e., only a few users exhibit high activity, while the majority show very low participation), potentially skewing the observed vocabulary sizes. To disentangle the possible effect of user activity, we categorize users into four classes—low, medium, high, and very high—based on the number of comments they have posted on each specific dataset (more details about the classification criteria are provided in *Materials and Methods*). Fig. 2 illustrates the distribution of types within each class except low, thus considering only users with sufficient activity.

As expected, higher activity levels among users are associated with vocabulary distributions centered around higher values,

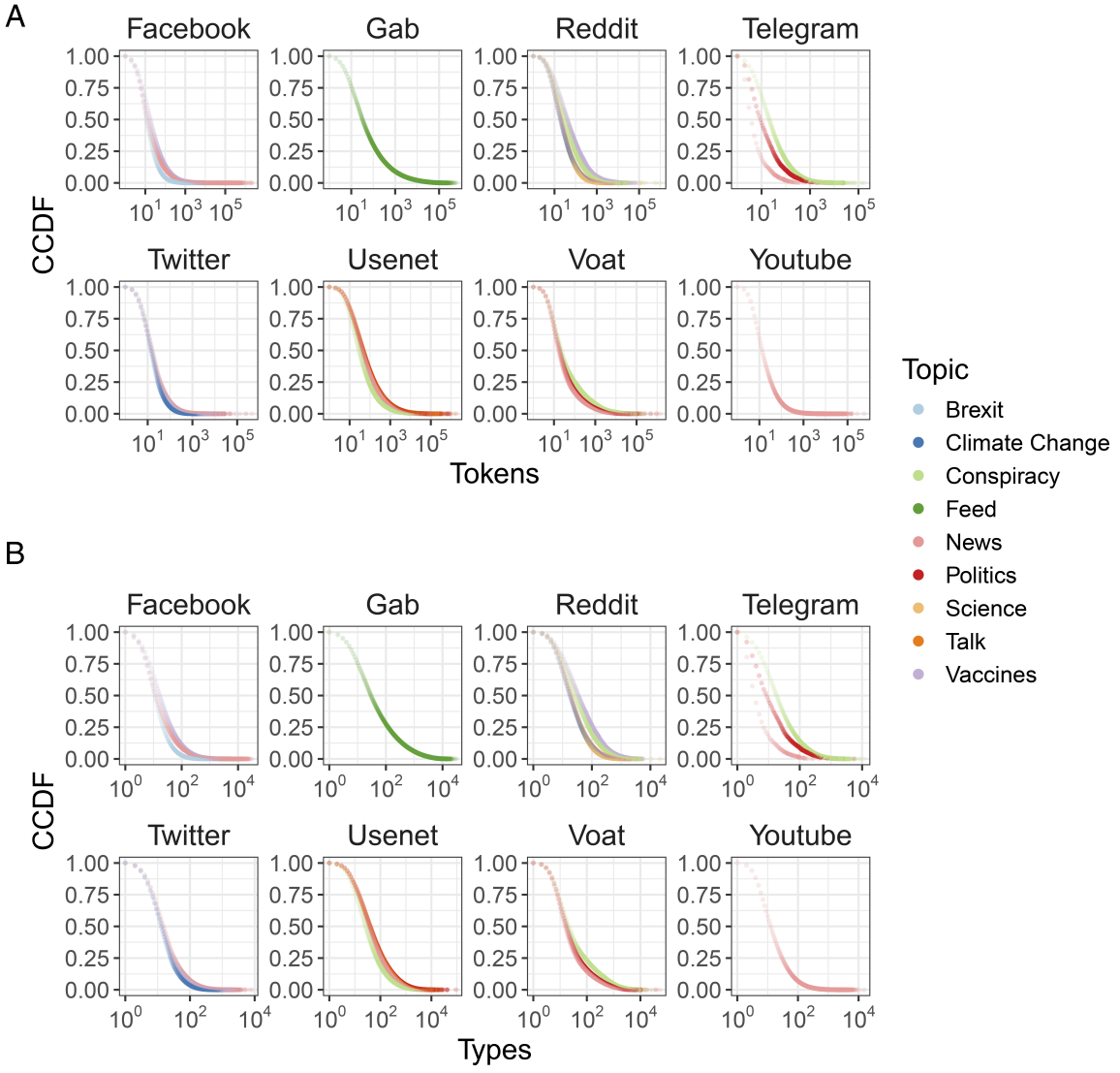

Fig. 1. CCDF of the distributions of number of (*A*) tokens and (*B*) types used by each user. All the distributions exhibit consistency across different social media platforms and topics, with types having shorter tails than tokens.

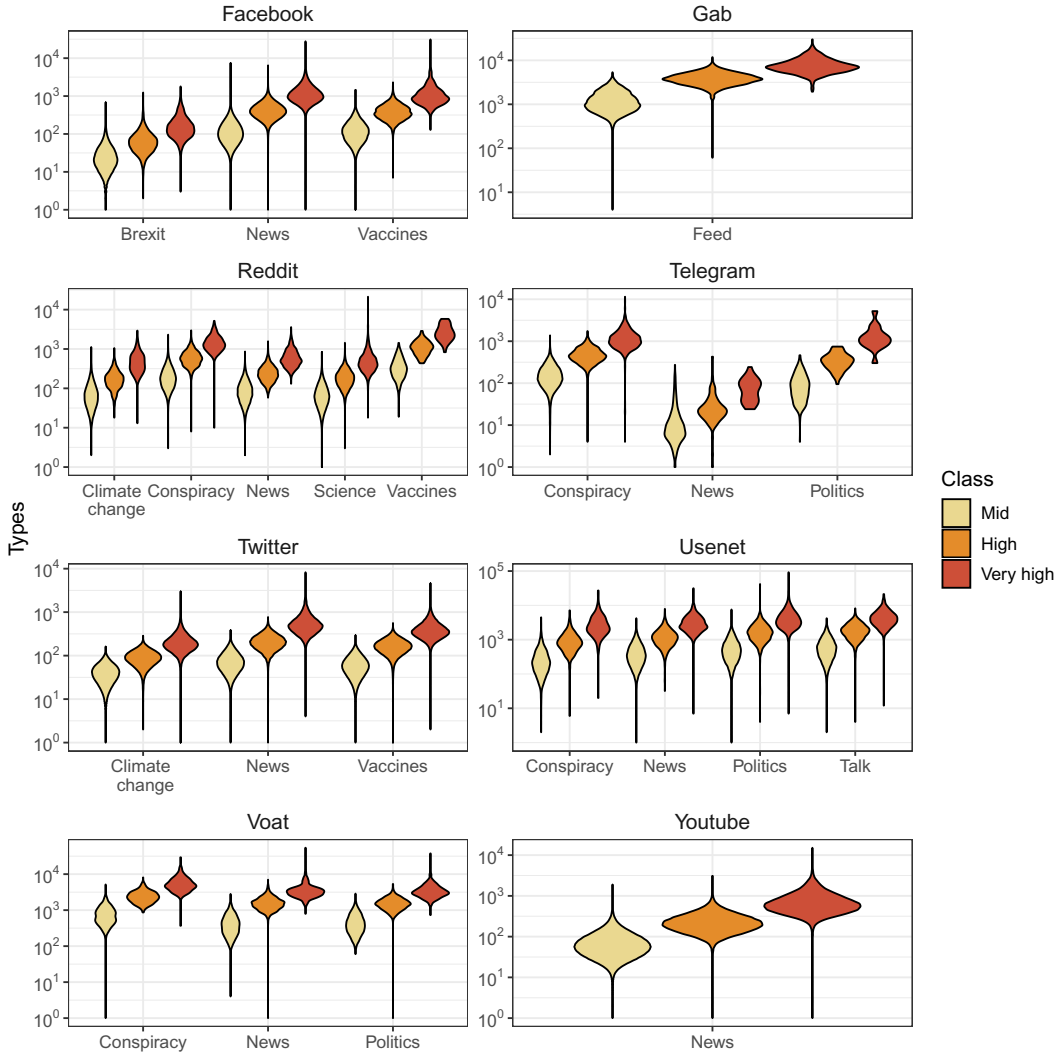

Fig. 2. Distribution of the number of types (i.e., unique words) employed by users according to their activity class, which is determined by the number of comments they left in each specific dataset.

indicating a more diverse lexical richness. Interestingly, while vocabulary size distributions generally show consistency across various social media platforms, it seems specific topics demand larger vocabularies. However, the shifting in the distributions could be a consequence of the different sizes of each dataset, resulting in some users having a larger number of comments and, thus, a larger number of types/tokens. To control for this effect, *SI Appendix*, Fig. S2 represents the same plots but in which activity classes are computed according to the whole distribution of comments of a specific social. Thus, we are able to compare the behavior of users with approximately the same number of comments. The results show that the shifting disappear, providing evidence that neither the topic of discussion can influence the total number of types used by users.

Furthermore, we examine whether users' vocabulary distributions adhere to Zipf's law (49, 50), a principle that suggests a predictable frequency distribution of words in human languages. The results, presented in *Zipf's Law on Comments*, indicate consistent exponents across social media and topics, supporting the hypothesis that the form of Zipf's law is a valid approximation of the frequency term observed in online social media communications.

Vocabulary Evolution. In the previous section, we explored aggregated users' production without examining their evolution. To address this gap, we analyze how the vocabularies of individual users evolve over time by tracking the rate at which they introduce new words in their time-ordered comments.

We chronologically arrange each user's comments and apply the same tokenization process used for our previous analyses. For a user u with n comments, we compute the vector $\mathbf{v}^u \in \mathbb{N}^n$, where each entry i represents the cumulative count of unique words up to the i-th comment. For example, if a user u writes two comments, the first containing the words "politics," "health," "comics" and the second "politics," "left," "right," then $\mathbf{v}^u$ would be (3, 5). We calculate $\mathbf{v}^u$ for all users having 25 to 100 comments (for manageability we consider 50 to 100 comments in Facebook News). This range ensures that we focus on users with a reasonable activity level and excludes accounts that may be malicious or not genuine, often exhibiting excessively high comment counts.

We employ a linear interpolation on the scaled values of $\mathbf{v}$ in $[0, 1]$ to track the evolution of users' vocabularies. Since $v_i \leq v_{i+1}, i = 1, \ldots, N$, a possible measure of the speed at which each user reaches its maximum vocabulary is the area under this curve,

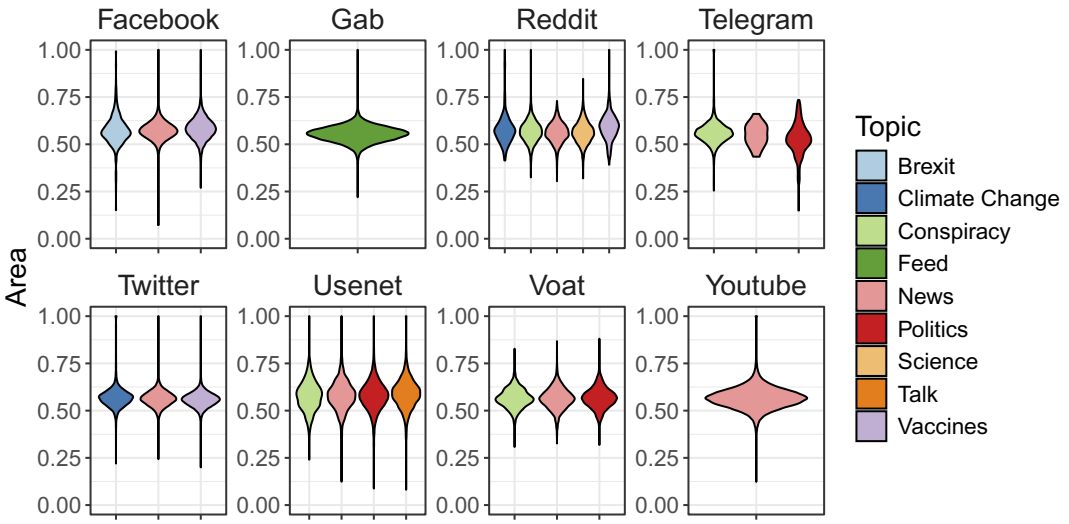

Fig. 3. Distribution of the area under the curves determined by users' progressive exploration of their vocabulary. These results suggest a continuous but modest addition of new words independently of topic and social.

similar to methodologies used in previous studies (51). This value, which ranges from 0 to 1, provides insights into vocabulary usage patterns: Values close to 0 indicate a late expansion of vocabulary, values close to 1 suggest rapid saturation of vocabulary usage early on, and values around 0.5 indicate a steady increase in vocabulary across comments. To clarify our procedure, in *Evolution of User's Vocabulary*, we provide some examples of these curves.

Fig. 3 illustrates the distribution of these measurements across different topics and social media platforms.

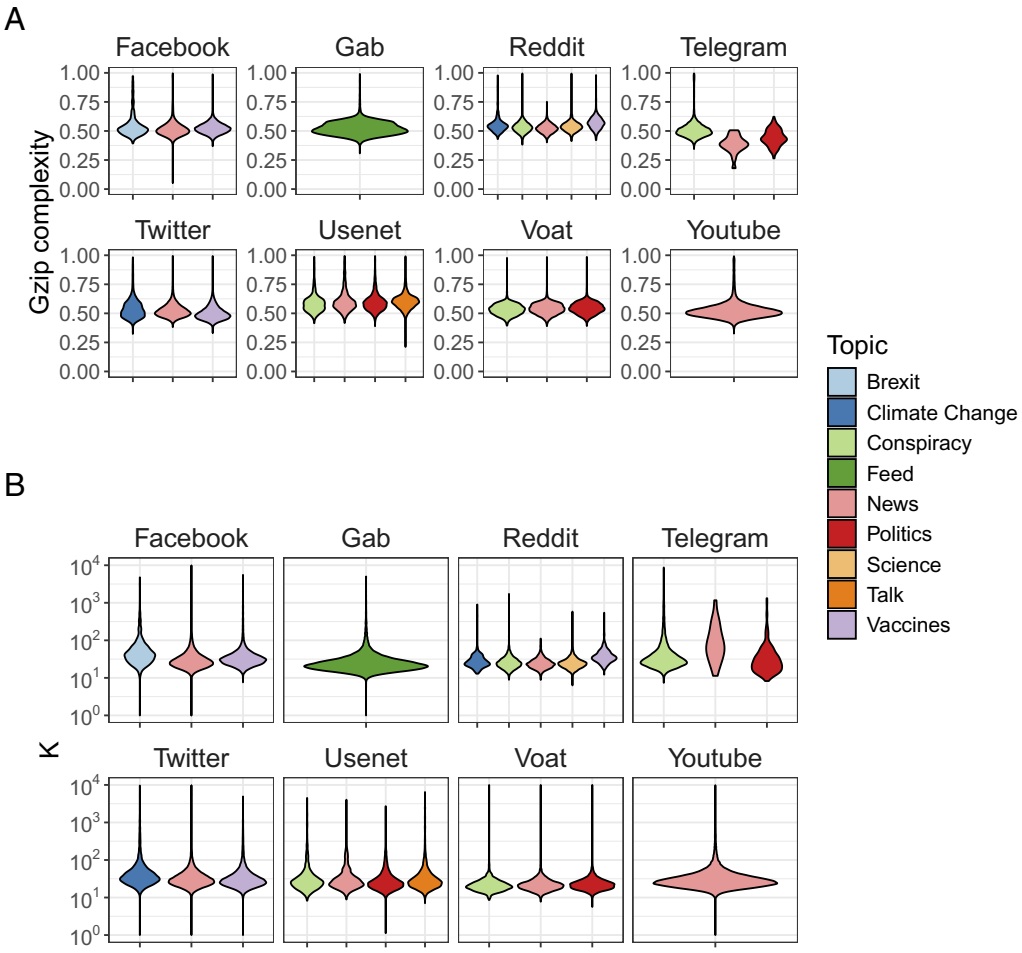

Fig. 4. Distribution of (*A*) gzip complexity and (*B*) $K-$complexity for users having at least 20 comments. For the larger dataset, we selected a sample of 50,000 users to compute $K-$complexity. For visual reasons, we add 1 to all the values of $K-$complexity.

We observe a general consistency across all distributions, peaking at values around 0.6; this suggests, on average, a continuous but modest addition of new words to a user's vocabulary with a minority of users reaching their whole vocabulary in the first comments.

The consistency of the findings underscores universal behaviors that appear to be largely independent of the specific platforms and topics involved.

Comments Complexity. Beyond the size of users' vocabularies, exploring the general complexity of comments adds a significant dimension to our analysis. The complexity of texts can be approached from various perspectives, and the literature made numerous metrics available (49, 52–57). After a careful review of the available measures and previous works, we decided to rely upon two of them that are able to provide a somewhat orthogonal perspective in terms of text complexity, namely Yule's K-complexity and gzip complexity g. The detailed methodologies are described in *Materials and Methods*. We recall that a high K suggests a slight lexical complexity, whereas it assumes its minimum (i.e., $K = 0$) for a text in which only distinct words are used. On the other hand, values of g close to 0 (or even negative) suggest texts with low repetitiveness, while values close to 1 indicate texts with high repetitive patterns.

As with our previous analysis, we compile the complete set of comments from each user into single documents. K-complexity is calculated following the previously outlined preprocessing steps, whereas gzip complexity is assessed using the raw texts, as done by previous works (58). Fig. 4 illustrates the distribution of both complexity measures among users who have posted at least 20 comments. For datasets comprising over 50,000 users, Yule's K-complexity was calculated on a sample of 50,000 users to ensure manageability and computational efficiency.

We observe consistency in the distributions of both K and g on different social media platforms and topics. Generally, the texts produced by users display moderate lexical complexity and repetitiveness. However, a minority of users produce highly repetitive texts and exhibit low complexity, which may indicate the presence of automated or coordinated accounts (58–63). Notably, gzip complexity seems to detail better than K the smaller level of complexity of the Conspiracy topic with respect to News and Politics, suggesting higher repetitive texts in that topic.

In *Null Model of Complexity Measures*, we present an analysis where each user's comments are randomized before being aggregated into documents, i.e. the document associated with a user is made of random comments from other users. The results show distributions with a lower variance, suggesting much more uniform behaviors. Although distributions look similar,

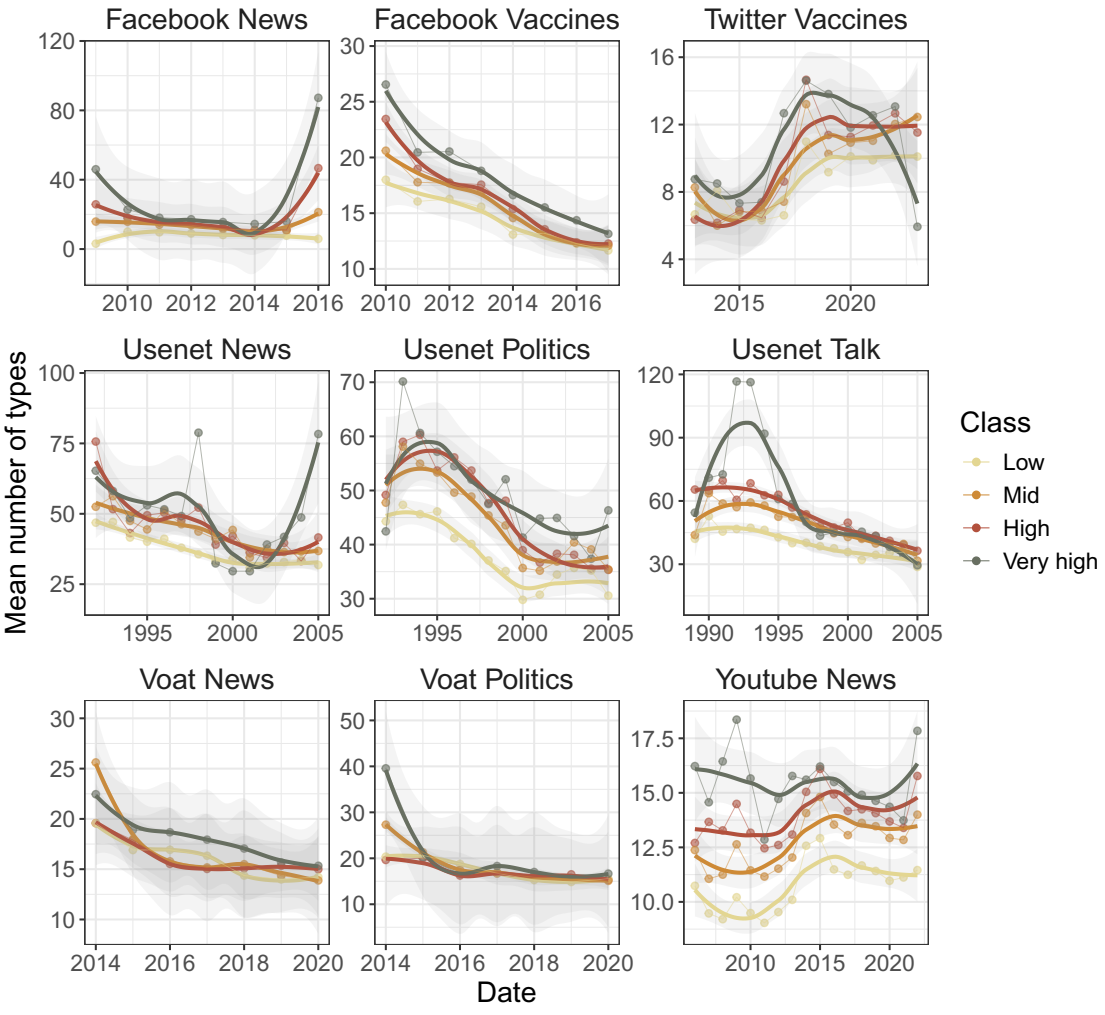

Fig. 5. Evolution of the mean number of types in each dataset. The smooth curves are obtained using a loess regression. In almost all cases, we observe a decrease in the mean number of types used.

Mann–Whitney tests detect that the real and null distributions are different in almost all cases, suggesting that users adopt their vocabulary that is not replicable with a random assignment of comments.

Evolution of Complexity. In previous sections, we have explored the complexity of texts written by users online. However, an interesting point to investigate regards the evolution of the complexity of comments over time as a proxy for the use of social media as a public square for opinion sharing and active debate. Thus, our focus now shifts to determining whether the complexity of comments has changed over time, possibly revealing some simplification processes. We select subsets of datasets with a sufficiently broad time span to achieve this. Specifically, we include Facebook News, Facebook Vaccines, Twitter Vaccines, Usenet News, Usenet Politics, Usenet Talk, Voat News, Voat Politics, and YouTube News, and we analyze the whole set of comments without aggregating by user.

First, we examine the evolution of the number of types (i.e., unique words), being the most straightforward complexity measure. To mitigate any potential bias from users' activity levels, for each year we classify each user into one of four activity classes—low, mid, high, or very high—as previously established. Fig. 5 illustrates the annual progression of the average number of types across all activity classes, computed for all years with a minimum number of 100 comments.

The number of unique words appears to decrease in all user classes except for those on Twitter and YouTube. Additionally, *SI Appendix*, Fig. S5 reveals that the TTR remains relatively stable across all platforms, suggesting a concurrent reduction in the total number of words (i.e., tokens) used. This trend also reflects a decrease in user activity, characterized by shorter comments that contain fewer unique words.

To further quantify the relationship between time and text complexity, we implement a regression model with interaction terms to account for the specific social media platform where each comment was posted. For this analysis, we use a sample of 6,000 comments per year from each platform and topic, ignoring all years having less than this number of comments.

We provide a detailed discussion of the model in *Materials and Methods*. Moreover, *SI Appendix*, Table S1 contains a breakdown of the dataset used for this experiment.

We recall here that our regression model employs complexity measures as regressors and the year of the comments as the dependent variable. To allow for a better comparison between measures, we first normalize each regressor and, after detecting a heteroskedasticity problem, we correct SEs with their robust version. The model, whose coefficients are detailed in Table 1, achieves an adjusted R^2 value of approximately 0.87.

We use Facebook as the reference category, meaning that the estimates β_i refer to the model obtained using Facebook comments. On the other hand, the "Total Estimate" column reports the sum of the coefficient with its relative baseline, i.e. $\beta_{k,j} + \beta_k$, thus describing the change in y after a unit change of regressor k if all the other variables are fixed and the comment comes from social j. In simpler terms, $\beta_{k,j}$ is the deviation from the behavior observed on Facebook.

The results suggest that the number of unique words is positively correlated with more recent years on Twitter and YouTube, as previously observed. Conversely, there is a negative relation on other platforms. Regarding Yule's K-complexity, higher values are associated with more recent years, thus suggesting that comments exhibit a decrease in lexical complexity over time.

Table 1. Estimates of coefficient of the regression model

Variables	Estimate	Total estimate	SE	*P*
β_0	2,013.091	2,013.091	0.010	< 0.001
β_1	−0.367	−0.367	0.034	< 0.001
β_2	0.079	0.079	0.007	< 0.001
β_3	−0.228	−0.228	0.007	< 0.001
$\beta_{0,tw}$	7.344	2,020.435	0.036	< 0.001
$\beta_{0,un}$	−13.644	1,999.447	0.013	< 0.001
$\beta_{0,vt}$	4.376	2,017.467	0.013	< 0.001
$\beta_{0,yt}$	2.076	2,015.167	0.021	< 0.001
$\beta_{1,tw}$	4.548	4.181	0.136	< 0.001
$\beta_{1,un}$	0.314	−0.054	0.035	< 0.001
$\beta_{1,vt}$	0.357	−0.011	0.039	< 0.001
$\beta_{1,yt}$	1.031	0.664	0.064	< 0.001
$\beta_{2,tw}$	−0.071	0.008	0.009	< 0.001
$\beta_{2,un}$	−0.067	0.012	0.012	< 0.001
$\beta_{2,vt}$	−0.067	0.012	0.010	< 0.001
$\beta_{2,yt}$	−0.070	0.009	0.019	< 0.001
$\beta_{3,tw}$	−0.098	−0.326	0.032	0.002
$\beta_{3,un}$	−0.540	−0.769	0.015	< 0.001
$\beta_{0,vt}$	0.047	−0.181	0.012	< 0.001
$\beta_{0,yt}$	0.169	−0.059	0.021	< 0.001

The results suggest that comments exhibit lower lexical complexity and repetitiveness as time progresses.

Finally, *g*-complexity consistently shows negative associations with time across all platforms, indicating that recent comments tend to have lower repetitiveness, despite their lower lexical complexity.

To check the robustness of the results, we conduct a similar analysis in *Logistic Regression* using a logistic regression. Notably, despite using a different methodology, the results show consistency.

Overall, the data indicate that comments across platforms exhibit lower lexical complexity and repetitiveness as time progresses. Furthermore, in nearly all cases, comments become shorter and contain fewer unique words. These findings are consistent with the existing literature on linguistic simplification, which describes the process as involving increased lexical and morphological transparency, as well as the reduction of redundancy.

One factor contributing to the observed reduction in text length could be the evolution of language itself within digital environments, where new words and shorthand expressions continuously emerge. For instance, abbreviations and neologisms such as "ttyl" or terms like "ghosting" reduce the need for longer, more descriptive language. This phenomenon reflects the dynamic nature of language adaptation to the rapid, informal communication typical of social media platforms. While these linguistic innovations enable users to convey complex ideas more succinctly, they also contribute to the overall trend of linguistic simplification and may play a role in the shortening of messages over time.

Conclusions

Our comprehensive analysis across eight major social media platforms reveals consistent patterns in user behavior and language complexity. This study clarifies how users' linguistic behavior has adapted to the digital era and evolved over nearly three decades of internet use. Despite the diversity of topics and

platforms analyzed, encompassing approximately 300 million English comments over nearly three decades, we find a general decrease in the length of comments and a reduction in lexical richness. Notably, the observed simplification aligns with the increased transparency and reduction of redundancy noted in the literature. We have also provided a model to explore how individual users evolve their vocabulary over time, noticing that most users gradually introduce new words, with distributions peaking at relatively low levels of vocabulary expansion.

This linguistic convergence mirrors long-standing processes of linguistic hybridization observed in language contact scenarios. However, this process is not only accelerated and intensified within the globalized, entertainment-driven context of social media, but also intersects with the broader societal dynamics analyzed through the lens of network science and social physics (30). On social media platforms, designed primarily for engagement, simplified and emotionally resonant language is favored by algorithms prioritizing viral content. As a result, the homogenization of language could occur more rapidly, reducing linguistic diversity. Following the current diachronic trend that points toward a growing number of languages becoming less complex over time, the convergence on social media narrows discourse, aligning users around simplified expressions that could further entrench polarization. The entertainment-focused nature of these platforms thus reinforces uniformity in communication, possibly limiting the range of ideas and perspectives shared.

Finally, our findings, based on a multiplatform analysis, support the hypothesis that language complexity is not influenced by the platform but rather reflects a broader aspect of human communication, highlighting a general trend of language evolution influenced by social dynamics, revealing inherent patterns in users' online communication that parallel long-established processes of linguistic hybridization and contamination.

Materials and Methods

Data Collection. Below, we report the detailed procedure for gathering each dataset.

Facebook. We utilize datasets from prior studies on discussions concerning Vaccines (64), News (65), and Brexit (66). For the vaccine topic, the dataset comprises approximately 2 million comments from public groups and pages collected over the period from January 2, 2010 to July 17, 2017. For the News topic, we selected pages from the Europe Media Monitor that reported the news in English, resulting in a dataset containing roughly 362 million comments between September 9, 2009, and August 18, 2016. Additionally, this dataset includes approximately 4.5 billion likes associated with posts and comments on these pages. Last, for the Brexit topic, the dataset encompasses around 460,000 comments from December 31, 2015 to July 29, 2016.

Gab. We collected data from the Pushshift.io archive (https://files.pushshift.io/gab/) on discussions from the platform's inception on August 10, 2016, until it temporarily went offline on October 29, 2018, following the Pittsburgh shooting (67). The dataset includes approximately 14 million comments.

Reddit. Data was collected from the Pushshift.io archive (https://pushshift.io/) covering the period from January 1, 2018 to December 31, 2022. We manually identified and selected subreddits for each topic that best represented the targeted discussions. From this process, we gathered approximately 800,000 comments from the *r/conspiracy* subreddit for the Conspiracy topic. For the Vaccines topic, we collected about 70,000 comments from the *r/DebateVaccines* subreddit, focusing on the COVID-19 vaccine debate. The *r/News* subreddit provided roughly 400,000 comments for the News topic. From the *r/environment* subreddit, we obtained approximately 70,000 comments related to the Climate Change topic. Last, the *r/science* subreddit yielded about 550,000 comments for the Science topic.

Telegram. We compiled a list of 14 channels, each linked to one of the study's topics. We manually collected messages and their associated comments from each channel. From the 4 channels related to the News topic (*news_notiziae, news_ultimora, news_edizionestraordinaria, news_covidultimora*), we gathered approximately 724,000 comments from posts dated between April 9, 2018 and December 20, 2022. For the Politics topic, the 2 channels (*politics_besttimeline, politics_polmemes*) yielded a total of about 490,000 comments from the period between August 4, 2017 and December 19, 2022. Last, the 8 channels focused on the Conspiracy topic (*conspiracy_bennyjhonson, conspiracy_tommyrobinsonnews, conspiracy_britainsfirst, conspiracy_loomeredofficial, conspiracy_thetrumpistgroup, conspiracy_trumpjr, conspiracy_pauljwatson, conspiracy_iononmivaccino*) produced approximately 1.4 million comments from August 30, 2019 to December 20, 2022.

Twitter. We utilized datasets from previous research that include discussions on Vaccines (68), Climate Change (7), and News (69). We collected approximately 50 million comments on the Vaccines topic from January 23, 2010 to January 25, 2023. For the News topic, we expanded the dataset used in ref. 69 by including all threads with fewer than 20 comments, resulting in a total of approximately 9.5 million comments collected from January 1, 2020 to November 29, 2022. Last, we gathered about 9.7 million comments on the Climate Change topic from January 1, 2020 to January 10, 2023.

Usenet. We collected data from the Usenet discussion system using the Usenet Archive (https://archive.org/details/usenet?tab=about). We identified a range of topics, including extensive, broad, and heterogeneous discussions within active and populated newsgroups. As a result, we selected conspiracy, politics, news, and talk as the topic candidates for our analysis. We gathered approximately 280,000 comments from the *alt for the Conspiracy topic.conspiracy* newsgroup, from September 1, 1994 to December 30, 2005. About 2.6 million comments were collected from the *alt for the Politics topic.politics* newsgroup between June 29, 1992 and December 31, 2005. We obtained approximately 620,000 comments for the News topic from the *alt.news* newsgroup, from December 5, 1992 to December 31, 2005. Finally, we collected all discussions from the *alt for the Talk topic.talk* newsgroup, totaling about 2.1 million contents, from February 13, 1989 to December 31, 2005.

Voat. We utilized a dataset described in ref. 70 that encompasses the entire lifespan of the platform from January 9, 2018 to December 25, 2020. This dataset includes approximately 16.2 million posts and comments from around 113,000 users across approximately 7,100 subverses (Voat's equivalent of a subreddit). Similarly to previous platforms, we linked topics to specific subverses. As a result, for the Conspiracy topic, we collected about 1 million comments from the *greatawakening* subverse during the platform's operational period. For the Politics topic, we gathered roughly 1 million comments from the *politics* subverse between June 16, 2014 and December 25, 2020. Last, we amassed approximately 1.4 million comments from the *news* subverse for the News topic between November 21, 2013 and December 25, 2020.

YouTube. We utilized a dataset referenced in previous research that initially focused on Climate Change discussions (7). This dataset has been expanded to include conversations on Vaccines and News topics, following the approach used for other platforms. The data collection for YouTube was conducted using the YouTube Data API (https://developers.google.com/youtube/v3). For the Climate Change topic, we collected approximately 840,000 comments from March 16, 2014 to February 28, 2022. For the Vaccines topic, we gathered comments between January 31, 2020 and October 24, 2021, that include keywords related to COVID-19 vaccines, such as *Sinopharm, CanSino, Janssen, Johnson&Johnson, Novavax, CureVac, Pfizer, BioNTech, AstraZeneca,* and *Moderna*, resulting in about 2.6 million comments. Finally, for the News topic, we collected roughly 20 million comments from February 13, 2006 to February 8, 2022, including videos and comments from a list of UK-based news outlets, provided by Newsguard, a fact-checking agency.

We select only English text from these data as detected by the cld3 package of R (71). Table 2 shows a breakdown of the resulting dataset.

Preprocessing of Comments. Before the main analysis, all comments were preprocessed to keep only the significant part of the texts and avoid spurious results. In particular, using the quanteda package of R (72, 73), we follow these steps:

1. we tokenize the comments using the tokens function, removing punctuation, symbols, numbers, URLs, hashtags, and English stopwords;

Table 2. Data breakdown of the dataset

Dataset	Time range	Comments	Users
Facebook brexit	2015-12-31 2016-07-29	322,365	171,054
Facebook vaccines	2010-01-02 2017-07-17	1,590,907	304,706
Facebook news	2009-09-09 2016-08-17	229,915,622	36,096,691
Gab feed	2016-08-10 2018-10-29	10,799,968	126,351
Reddit climate change	2018-01-01 2022-12-12	60,113	22,822
Reddit conspiracy	2018-01-01 2022-10-31	649,054	82,733
Reddit news	2018-01-01 2018-12-31	358,594	102,982
Reddit science	2018-01-01 2022-12-11	488,963	192,675
Reddit vaccines	2018-01-21 2022-11-06	59,980	4,866
Telegram conspiracy	2019-08-30 2022-12-20	1,111,479	107,009
Telegram news	2018-05-24 2022-12-16	4,493	1,674
Telegram politics	2017-08-05 2022-12-19	20,851	2,013
Twitter climate change	2020-01-01 2023-01-10	3,562,447	1,467,783
Twitter news	2020-01-01 2022-11-29	5,882,655	1,237,679
Twitter vaccines	2010-01-23 2023-01-25	17,682,887	4,916,617
Usenet conspiracy	1994-09-01 2005-12-30	160,632	30,223
Usenet news	1992-12-05 2005-12-30	385,404	51,204
Usenet politics	1992-06-29 2005-12-30	1,541,803	142,469
Usenet talk	1989-02-13 2005-12-30	1,390,574	128,321
Voat conspiracy	2018-01-09 2020-12-25	828,018	24,666
Voat news	2013-11-21 2020-12-25	1,164,549	78,199
Voat politics	2014-06-19 2020-12-25	902,419	59,442
YouTube news	2006-02-13 2022-02-10	20,946,472	5,623,730

We consider approximately 300M comments wrote by 50M users.

2. we apply a stemmer to the resulting tokens, reducing them to their root form, using the tokens_wordstem function set with english language;
3. we set all tokens to lowercase and remove specific tokens seen in social media, such as RT (retweet).

Note that, after these steps, a text may remain empty (for example if it contains only tags or hashtags). Therefore, we remove all the comments containing 0 tokens.

Measures of Text Complexity. In linguistics, special attention has always been paid to developing measures capable of detecting text complexity. In particular, a text can be considered complex from a variety of points of view, such as lexical, readability, or repetitiveness. Many measures have been proposed [see ref. 73 for a collection of measures provided by quanteda.textstats package in R (72)]. Here, we have focused on lexical complexity and repetitiveness measures, using Yule's K and gzip complexity.

For what concerns the former, previous works have highlighted that many lexical complexity measures are incapable of being independent of text length (54, 57). The same studies also highlight that, even if with some limitations, the well-known Yule's K-complexity (53) seems to be almost independent of text length. Given a text of length N with V unique words, Yule's K is defined as

$$K = 10^4 \cdot \left[-\frac{1}{N} + \sum_{i=1}^{V} V(i, N) \left(\frac{i}{N} \right)^2 \right], \qquad [1]$$

where $V(i, N)$ denotes the number of words appearing i times in the text. K is lower bounded by 0, a value obtained only when the text contains distinct words (i.e., $V(1, N) = N$). In general, the larger K is, the less rich the vocabulary is, even if it is not possible to find an upper bound to its value.

For what concerns the repetitiveness of text, we use an approach akin to previous works (74, 75), i.e. we compress the raw texts using gzip and compare their dimensions with the original ones. We define the gzip complexity g as

$$g = \frac{s_{raw} - s_{compressed}}{s_{raw}}, \qquad [2]$$

where s_{raw} is the size of the raw text and $s_{compressed}$ is the size of the compressed text. Note that if the text is highly repetitive $s_{compressed} \ll s_{raw}$ and therefore $g \approx 1$. On the other hand, low values characterize texts with low repetitiveness. In particular, for very short texts the compressor may increase the size, therefore g can also assume negative values.

Classification of Users. We adopt a nonparametric method developed by Glänzel and Schubert (76) to partition heavy-tailed distributions, and recently employed in different domains (3, 77, 78) to divide users into classes according to the number of comments they left. In detail, we consider four classes of activity, namely low, mid, high, very high, and adopt the following procedure: First, we compute the mean number of comments $\bar{x}$ and assign to class low all users that have left less than $\bar{x}$ comments; then, we delete these users from the distribution and recursively repeat the procedure until each user is assigned to one of the four classes low, mid, high, very high.

Regression Model. Since we are interested in obtaining a unique model capable of detecting the overall relationship between time, social media, and the complexity of comments, we consider

$$y_i \sim [1\ w_i\ K_i\ g_i] \cdot \mathbf{B} \cdot \begin{bmatrix} 1 \\ tw_i \\ vt_i \\ yt_i \\ un_i \end{bmatrix}, \qquad [3]$$

where $\cdot$ is the standard matrix product and

- w_i is the number of types of comment i;
- K_i is the K—complexity of comment i;
- g_i is the g-complexity of comment i;
- y_i is the year in which comment i has been created;
- tw_i, vt_i, yt_i, and un_i are dummy variables that are equal to one if and only comment i has been written in Twitter, Voat, YouTube, or Usenet, respectively.

Finally, **B** is the matrix of the coefficients, defined as follows:

$$B = \begin{bmatrix} \beta_0 & \beta_{0,tw} & \beta_{0,vt} & \beta_{0,yt} & \beta_{0,un} \\ \beta_1 & \beta_{1,tw} & \beta_{1,vt} & \beta_{1,yt} & \beta_{1,un} \\ \beta_2 & \beta_{2,tw} & \beta_{2,vt} & \beta_{2,yt} & \beta_{2,un} \\ \beta_3 & \beta_{3,tw} & \beta_{3,vt} & \beta_{3,yt} & \beta_{3,un} \end{bmatrix}$$

We use Facebook as the baseline, i.e. the estimates of coefficients β_i refer to Facebook comments. Therefore, the generic coefficient $\beta_k + \beta_{k,j}$ describes the change in y due to a unit increase of regressor k if all the other variables are kept constant and comment i comes from a social j different from Facebook. To obtain more interpretable estimates, we first normalize all the regressors. Therefore, the parameters quantify changes in the dependent variable in SDs. Moreover, since we detect heteroskedasticity, we correct the errors and tests using the *sandwich* package (79, 80) in R.

Ethics Statement. This study did not require ethical approval, as it involved only the analysis of anonymized data gathered from social media platforms. No direct interaction with human subjects was conducted, and all data were handled in compliance with relevant ethical guidelines. The study design ensured that individual identities could not be traced or inferred, thus protecting privacy and minimizing any ethical risks associated with the research.

Data, Materials, and Software Availability. Previously published data were used for this work (81).

ACKNOWLEDGMENTS. The work is supported by IRIS Infodemic Coalition (UK government, grant no. SCH-00001-3391), SERICS (PE00000014) under the National Recovery and Resilience Plan Ministry of University and Research (MUR) program funded by the European Union (EU)–NextGenerationEU, project CRESP from the Italian Ministry of Health under the program Climate Change Mitigation 2022, Programma Operativo Nazionale project "Ricerca e Innovazione" 2014–2020, and Progetti di Rilevante Interesse Nazionale Project MUSMA for Italian Ministry of University and Research (MUR) through the PRIN 2022. This work was supported by the PRIN 2022 "MUSMA"–CUP G53D23002930006–Funded by EU–Next-Generation EU–M4 C2 I1.1.

1. J. Tucker *et al*., Social media, political polarization, and political disinformation: A review of the scientific literature. *SSRN Electron. J.* (2018). https://dx.doi.org/10.2139/ssrn.3144139. Accessed 20 November 2024.
2. T. Aichner, M. Grünfelder, O. Maurer, D. Jegeni, Twenty-five years of social media: A review of social media applications and definitions from 1994 to 2019. *Cyberpsychol. Behav. Soc. Netw.* **24**, 215–222 (2021).
3. N. Di Marco, M. Cinelli, S. Alipour, W. Quattrociocchi, Users volatility on reddit and voat. *IEEE Trans. Comput. Soc. Syst.* **11**, 1–9 (2024).
4. A. M. Guess *et al*., How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science* **381**, 398–404 (2023).
5. M. Cinelli, G. De Francisci Morales, A. Galeazzi, W. Quattrociocchi, M. Starnini, The echo chamber effect on social media. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2023301118 (2021).
6. S. González-Bailón *et al*., Asymmetric ideological segregation in exposure to political news on Facebook. *Science* **381**, 392–398 (2023).
7. M. Falkenberg *et al*., Growing polarization around climate change on social media. *Nat. Clim. Chang.* **12**, 1114–1121 (2022).
8. S. A. Castaño-Pulgarín, N. Suárez-Betancur, L. M. T. Vega, H. M. H. López, Internet, social media and online hate speech. Systematic review. *Aggress. Violent. Behav.* **58**, 101608 (2021).
9. Y. Lupu *et al*., Offline events and online hate. *PLoS One* **18**, e0278511 (2023).
10. A. A. Siegel, "Online hate speech" in *Social Media and Democracy: The State of the Field, Prospects for Reform*, N. Persily, J. A. Tucker, Eds. (Cambridge University Press, 2020), pp. 56–88.
11. M. Avalle *et al*., Persistent interaction patterns across social media platforms and over time. *Nature* **628**, 582–589 (2024).
12. B. Nyhan *et al*., Like-minded sources on Facebook are prevalent but not polarizing. *Nature* **620**, 137–144 (2023).
13. A. M. Guess *et al*., Reshares on social media amplify political news but do not detectably affect beliefs or opinions. *Science* **381**, 404–408 (2023).
14. F. Greaves *et al*., Use of sentiment analysis for capturing patient experience from free-text comments posted online. *J. Med. Internet Res.* **15**, e2721 (2013).
15. G. Xu, Y. Meng, X. Qiu, Z. Yu, X. Wu, Sentiment analysis of comment texts based on BiLSTM. *IEEE Access* **7**, 51522–51532 (2019).
16. A. Alrumaih, A. Al-Sabbagh, R. Alsabah, H. Kharrufa, J. Baldwin, Sentiment analysis of comments in social media. *Int. J. Electr. Comput. Eng.* **10**, 2088–8708 (2020).
17. S. Volkova, Y. Bachrach, M. Armstrong, V. Sharma, Inferring latent user properties from texts published in social media. *Proc. AAAI Conf. Artif. Intell.* **29**, 4296–4297 (2015).
18. M. Perc, Evolution of the most common English words and phrases over the centuries. *J. R. Soc. Interface* **9**, 3323–3328 (2012).
19. M. K. Smith, Measurement of the size of general English vocabulary through the elementary grades and high school. *Genet. Psychol. Monogr.* **24**, 311–345 (1941).
20. E. Tschirner, Breadth of vocabulary and advanced English study: An empirical investigation. *Electron. J. Foreign Lang. Teach.* **1**, 27–39 (2004).
21. J. Milton, J. Treffers-Daller, Vocabulary size revisited: The link between vocabulary size and academic achievement. *Appl. Linguist. Rev.* **4**, 151–172 (2013).
22. N. S. Baron, *Always On: Language in an Online and Mobile World* (Oxford University Press, 2008).
23. G. McCulloch, *Because Internet: Understanding the New Rules of Language* (Penguin, 2020).
24. M. Zappavigna, *Discourse of Twitter and Social Media: How We Use Language to Create Affiliation on the Web* (Bloomsbury Academic, 2012).
25. D. Crystal, *Internet Linguistics: A Student Guide* (Routledge, 2011).
26. N. S. Baron, *Words Onscreen: The Fate of Reading in a Digital World* (Oxford University Press, 2015).
27. T. Gong, C. Coupé, A report on the workshop on complexity in language: Developmental and evolutionary perspectives. *Biolinguistics* **5**, 370–380 (2011).
28. K. Ehret, A. Berdicevskis, C. Bentz, A. Blumenthal-Dramé, Measuring language complexity: Challenges and opportunities. *Linguist. Vanguard* **9**, 1–8 (2023).
29. M. Mirzapour, J. P. Prost, C. Retoré, "Measuring linguistic complexity: Introducing a new categorial metric" in *Logic and Algorithms in Computational Linguistics 2018*, R. Loukanova, Ed. (Springer International Publishing, Cham, 2020), pp. 95–123.
30. M. Jusup *et al*., Social physics. *Phys. Rep.* **948**, 1–148 (2022).
31. M. Miestamo, *Grammatical Complexity in Cross-Linguistic Perspective* (John Benjamins Publishing Company, 2008), pp. 23–41.
32. J. Eisenstein, B. O'Connor, N. A. Smith, E. P. Xing, Diffusion of lexical change in social media. *PLoS One* **9**, e113114 (2014).
33. P. Muhlhausler, *Pidgin & Creole Linguistics* (University of Westminster Press, 1986).
34. P. Mühlhäusler, *Language of Environment, Environment of Language: A Course in Ecolinguistics* (Battlebridge Publications, 2003).
35. P. Trudgill, "Sociolinguistic typology and complexification" in *Language Complexity As an Evolving Variable* (Oxford University Press, Oxford, 2009), vol. 13, pp. 98–109.
36. J. B. Pride, L. Milroy, "Language and social networks" in *Language* (Linguistic Society of America, 1982), p. 231.
37. R. D. Perkins, *Deixis, Grammar, and Culture* (John Benjamins Publishing Company, 1992).
38. P. Trudgill, *Sociolinguistic Variation and Change* (Edinburgh University Press, 2019).
39. S. Vejdemo, T. Hörberg, Semantic factors predict the rate of lexical replacement of content words. *PLoS One* **11**, e0147924 (2016).
40. J. McWhorter, *The Power of Babel: A Natural History of Language* (Random House, 2011).
41. J. Aitchison, "Language change" in *The Routledge Companion to Semiotics and Linguistics* (Routledge, 2005), pp. 111–120.
42. A. M. Petersen, J. N. Tenenbaum, S. Havlin, H. E. Stanley, M. Perc, Languages cool as they expand: Allometric scaling and the decreasing need for new words. *Sci. Rep.* **2**, 943 (2012).
43. W. Labov, *Principles of Linguistic Change, Volume 3: Cognitive and Cultural Factors* (John Wiley & Sons, 2011), vol. 3.
44. N. Schilling-Estes, "Linguistic variation as social practice: The linguistic construction of identity in belten high" in *Language* (Linguistic Society of America, 2001), pp. 575–577.
45. M. M. Waldrop, *Complexity: The Emerging Science at the Edge of Order and Chaos* (Simon and Schuster, 1993).
46. H. A. Simon, "The architecture of complexity" in *Facets of Systems Science* (Springer, Boston, MA, 1991), pp. 457–476.
47. O. Dahl, *The Growth and Maintenance of Linguistic Complexity* (John Benjamins Publishing Company, 2004).
48. S. S. Mufwene, C. Coupé, F. Pellegrino, *Complexity in Language: Developmental and Evolutionary Perspectives* (Cambridge University Press, 2017).
49. G. K. Zipf, *The Psycho-Biology of Language: An Introduction to Dynamic Philology* (Routledge, 2013).
50. G. K. Zipf, *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology* (Ravenio books, 2016).
51. G. Etta *et al*., Characterizing engagement dynamics across topics on Facebook. *PLoS One* **18**, 1–12 (2023).
52. F. Golcher, "A new text statistical measure and its application to stylometry" in *Proceedings of Corpus Linguistics* (2007).
53. C. U. Yule, *The Statistical Study of Literary Vocabulary* (Cambridge University Press, 2014).
54. F. J. Tweedie, R. H. Baayen, How variable may a constant be? Measures of lexical richness in perspective. *Comput. Humanit.* **32**, 323–352 (1998).
55. D. Dugast, *Vocabulaire et stylistique* (Slatkine, 1979), vol. 8.
56. A. Rényi, "On measures of entropy and information" in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, J. Neyman, Ed. (University of California Press, 1961), vol. 4, pp. 547–562.
57. K. Tanaka-Ishii, S. Aihara, Computational constancy measures of texts–yule's skand rényi's entropy. *Comput. Linguist.* **41**, 481–502 (2015).
58. K. Lee, B. Eoff, J. Caverlee, Seven months with the devils: A long-term study of content polluters on twitter. *Proc. Int. AAAI Conf. Web Soc. Media* **5**, 185–192 (2021).
59. N. Di Marco, S. Brunetti, M. Cinelli, W. Quattrociocchi, Post-hoc evaluation of nodes influence in information cascades: The case of coordinated accounts. *ACM Trans. Web* (2024). https://doi.org/10.1145/3700644. Accessed 20 November 2024.
60. M. Cinelli, S. Cresci, W. Quattrociocchi, M. Tesconi, P. Zola, Coordinated inauthentic behavior and information spreading on twitter. *Decis. Support Syst.* **160**, 113819 (2022).
61. A. C. Nwala, A. Flammini, F. Menczer, A language framework for modeling social media account behavior. *EPJ Data Sci.* **12**, 33 (2023).

62. F. B. Keller, D. Schoch, S. Stier, J. Yang, Political astroturfing on twitter: How to coordinate a disinformation campaign. *Polit. Commun.* **37**, 256–280 (2019).
63. D. Pacheco *et al*., Uncovering coordinated networks on social media: Methods and case studies. *Proc. Int. AAAI Conf. Web Soc. Media* **15**, 455–466 (2021).
64. A. L. Schmidt, F. Zollo, A. Scala, C. Betsch, W. Quattrociocchi, Polarization of the vaccination debate on Facebook. *Vaccine* **36**, 3606–3612 (2018).
65. A. L. Schmidt *et al*., Anatomy of news consumption on Facebook. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 3035–3039 (2017).
66. M. Del Vicario, F. Zollo, G. Caldarelli, A. Scala, W. Quattrociocchi, Mapping social dynamics on Facebook: The Brexit debate. *Soc. Netw.* **50**, 6–16 (2017).
67. T. Hunnicutt, P. Dave, Gab.com goes offline after Pittsburgh synagogue shooting. Reuters (2018). https://www.reuters.com/article/world/gab-com-goes-offline-after-pittsburgh-synagogue-shooting-idUSKCN1N20QN/. Accessed 20 November 2024.
68. C. M. Valensise, et al., Lack of evidence for correlation between Covid-19 infodemic and vaccine acceptance. arXiv [Preprint] (2021). https://arxiv.org/abs/2107.07946 (Accessed 20 November 2024).
69. A. Quattrociocchi, G. Etta, M. Avalle, M. Cinelli, W. Quattrociocchi, "Reliability of news and toxicity in twitter conversations" in *Social Informatics*, F. Hopfgartner, K. Jaidka, P. Mayr, J. Jose, J. Breitsohl, Eds. (Springer International Publishing, Cham, 2022), pp. 245–256.
70. A. Mekacher, A. Papasavva, "i can't keep it up" a dataset from the defunct voat.co news aggregator. *Proc. Int. AAAI Conf. Web Soc. Media* **16**, 1302–1311 (2022).
71. J. Ooms, *cld3: Google's Compact Language Detector 3*, (2024). R package version 1.6.0.
72. K. Benoit *et al*., quanteda: An R package for the quantitative analysis of textual data. *J. Open Sour. Softw.* **3**, 774 (2018).
73. K. Benoit *et al*., quanteda: An R package for the quantitative analysis of textual data. *J. Open Source Softw.* **3**, 774 (2018).
74. E. Parada-Cabaleiro *et al*., Song lyrics have become simpler and more repetitive over the last five decades. *Sci. Rep.* **14**, 5531 (2024).
75. A. Desiderio, A. Mancini, G. Cimini, R. Di Clemente, Recurring patterns in online social media interactions during highly engaging events. arXiv [Preprint] (2023). http://arxiv.org/abs/2306.14735 (Accessed 20 November 2024).
76. W. Glänzel, A. Schubert, Characteristic scores and scales in assessing citation impact. *J. Inf. Sci.* **14**, 123–127 (1988).
77. G. Abramo, C. A. D'Angelo, A. Soldatenkova, An investigation on the skewness patterns and fractal nature of research productivity distributions at field and discipline level. *J. Informet.* **11**, 324–335 (2017).
78. M. Cinelli, Ambiguity of network outcomes. *J. Bus. Res.* **129**, 555–561 (2021).
79. A. Zeileis, S. Köll, N. Graham, Various versatile variances: An object-oriented implementation of clustered covariances in R. *J. Stat. Softw.* **95**, 1–36 (2020).
80. A. Zeileis, Econometric computing with HC and HAC covariance matrix estimators. *J. Stat. Softw.* **11**, 1–17 (2004).
81. M. Avalle *et al*., Persistent interaction patterns across social media platforms and over time. OSF. https://osf.io/fq5dy/. Accessed 20 November 2024.