The Evolution of Language in Social Media Comments

By: Edoardo Loru, Niccolò Di Marco, Alessandra Olga Grazia Serra

Referent – Stefanie Chuqiao Li
Credit - 6 credits
Course - CrossTemporal NLP

Introduction

- Evolution of language in social media comments across platforms and decades
- Dataset scale: **~300M English comments**, **~50M users**, **8 platforms**, time span up to **~34 years**
- Focus: measurable lexical behavior (vocabulary size, vocabulary growth, lexical richness, repetitiveness) rather than anecdotal “language decay” claims
- Aim: test whether changes are **platform-driven** or reflect **universal human constraints**

Data overview

1. Platforms:
    a. Facebook, Twitter, YouTube, Voat, Reddit, Usenet, Gab, Telegram

2. Topics include News, Politics, Vaccines, Climate change, Conspiracy, Science, Talk, Brexit, Feed (varies by platform)

3. Time spans differ Usenet provides long historical baseline
    → modern platforms give recent high-volume data

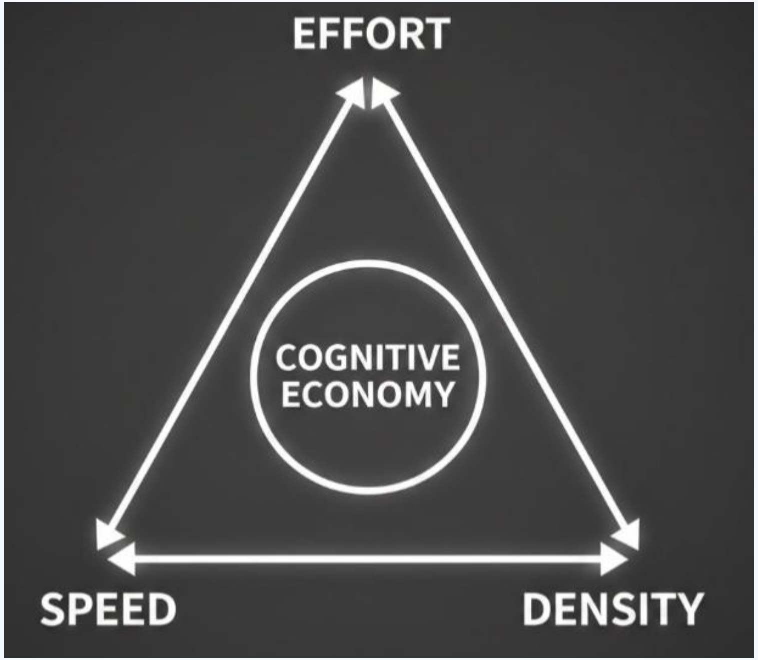
4. Key benefit
    → cross-platform comparison reduces overfitting conclusions to one community or one event cycle

Cognitive Economy

Speakers often optimize for:

Cognitive economy
→ Say as much as needed, with as little cost as possible.

Important distinction
→ Efficiency ≠ simplification
→ Fewer words ≠ less information



Trade-off-model

trade-off model (Ferrer-i-Cancho & Solé (2003)) suggest that:

Communication optimized by balancing:
→ **speaker effort** (shorter, easier production)
→ **listener effort** (clarity, reduced ambiguity)

→ Social media plausibily shifts trade-offs (time pressure, fragmentation of context)
→ to maximize viewer retention
  - → shorter + less lexically rich, but also less repetitive
  - → “compressed expression”

language complexity

Absolute Complexity	Relative Complexity
Intrinsic Property → Internal characteristic of the language itself **User independence** → Exist independently of any interaction of usage by individuals **Theoretical Nature** → Often categorized as a theoretical abstraction with little real-world application	**User Perspective** → How is the language experienced by the person using it **Cost and Difficulty** → Based on effort and difficulty a user encounters when communicating Affected by factors like reading comprehension skill and processing speed → **More suitable**

Types and Tokens

Token (word token) = word amount in a text
→ If a user writes: “**this** is **this**”, that is **3 tokens** (“this”, “is”, “this”)
→ Tokens mainly measure **how much language is produced**

Type (word type) = *each distinct word form* (iIn “this is this”, the distinct words are {“this”, “is”} → **2 types**
→ Types mainly measure **how diverse the vocabulary is** (breadth of lexicon used)

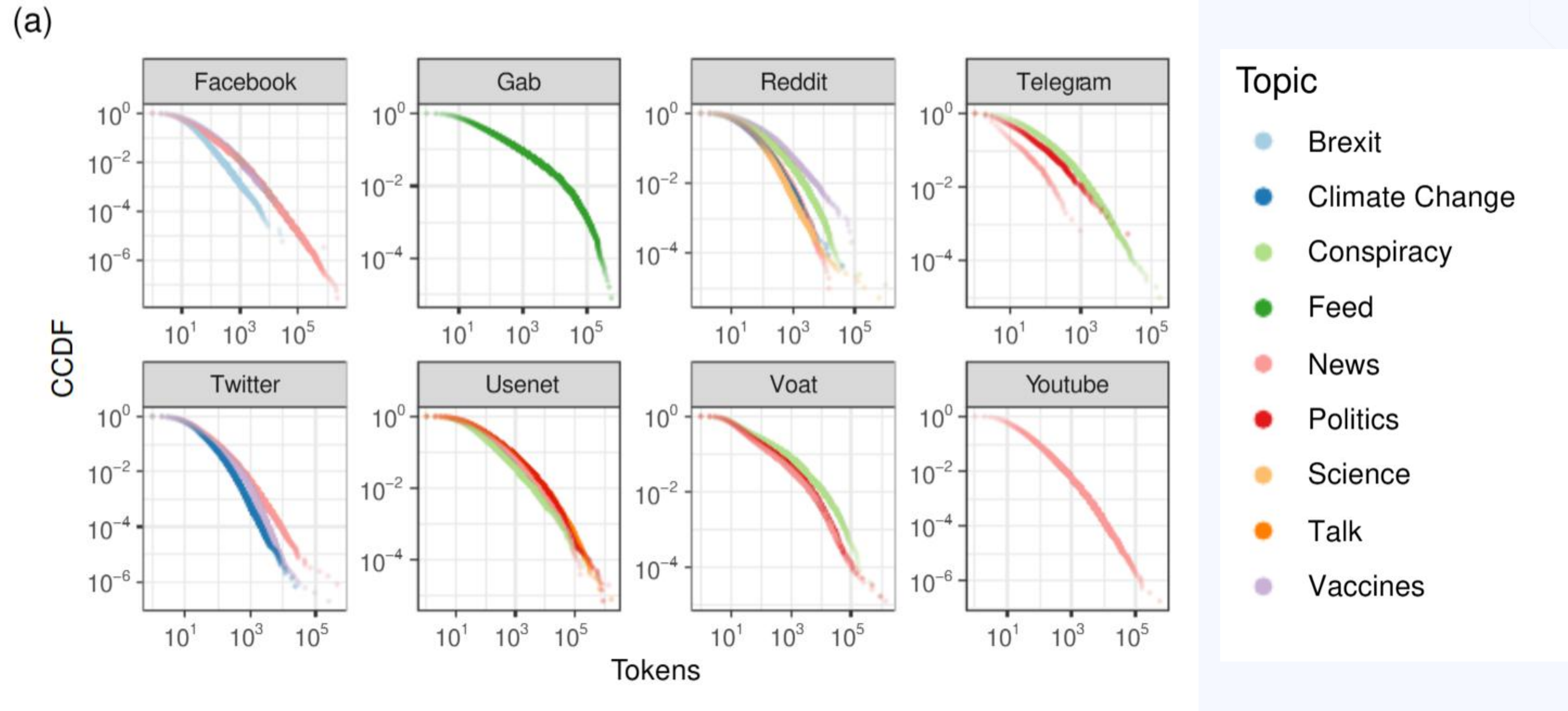
For every user all comments get Tokenized and then appended to one singular dokument

TTR of measure of complexity

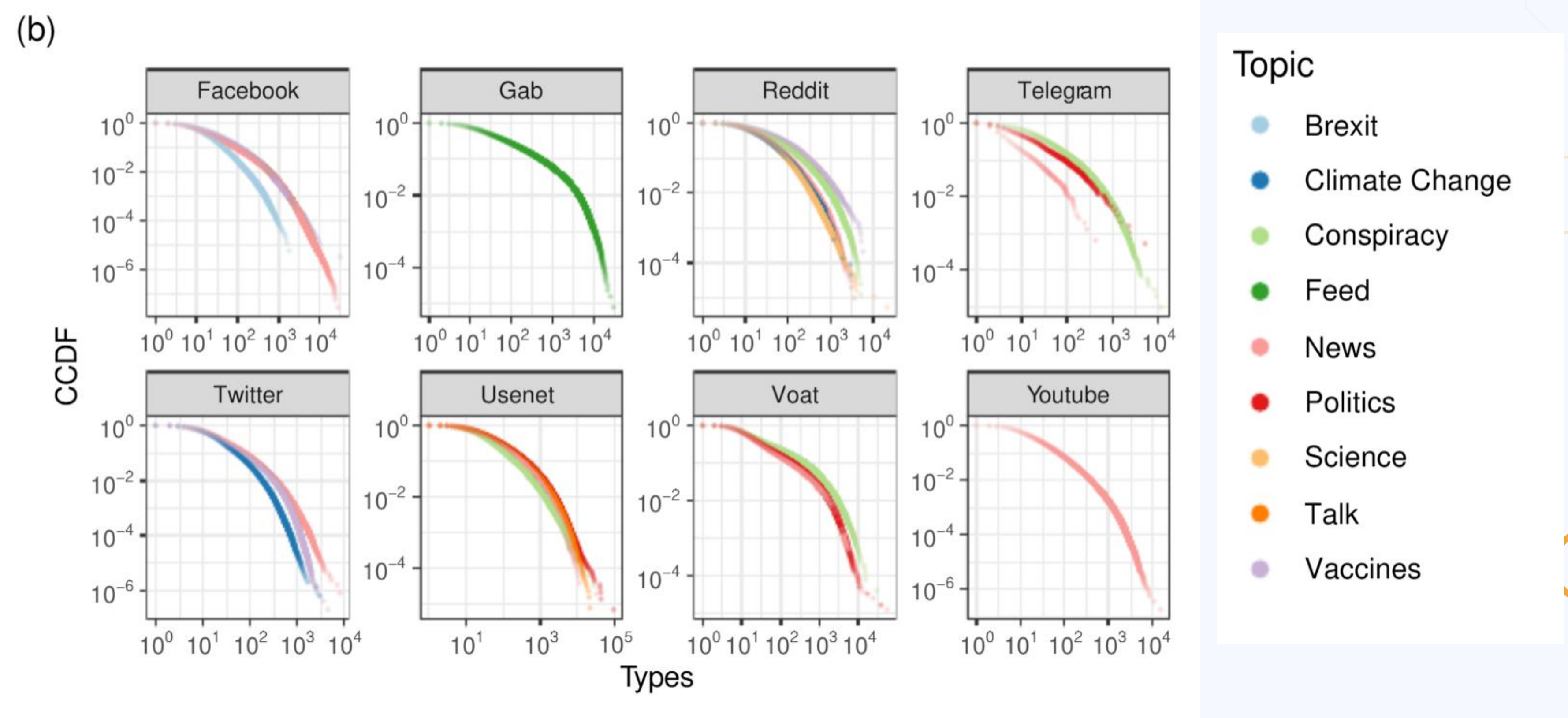
$$\text{TTR} = \frac{\text{Types}}{\text{Token}}$$

→ not as reliable
Short comments usually have types ≈ tokens
→longer comments = more repitition

Distribution of Tokens and Types



Distribution of Tokens and Types



CCDF

Complementary Comulative distribution Function

$CCDF(x)=P(X\geq x)$ → probalitiy that a variable is atleast as large as x

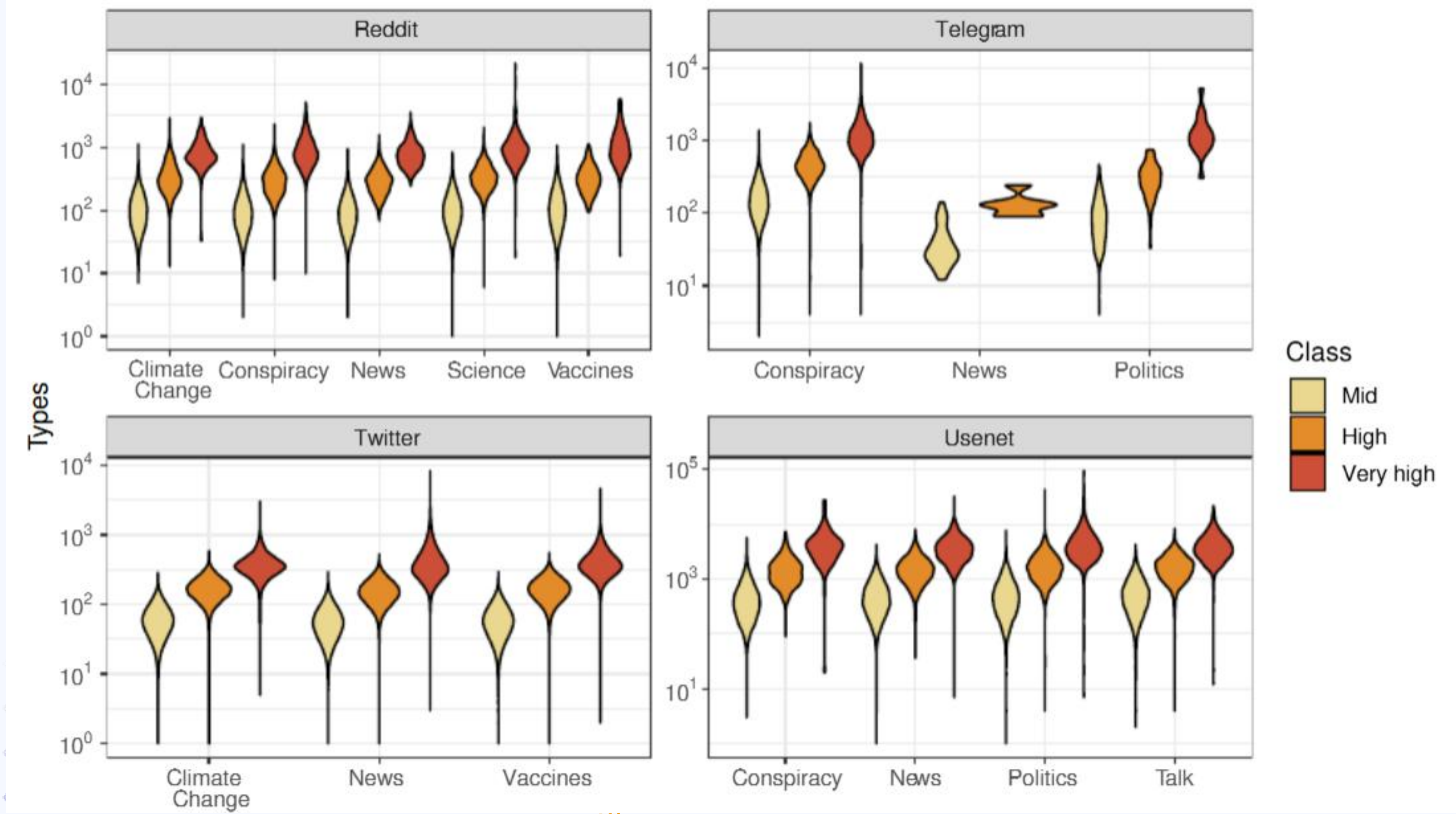
- Token distributions have longer tails than type distributions → vocabulary expansion is slower than production volume
- Cross-platform similarity in shapes suggests “platform mechanics” change scale but not distributional form
- Implication: any complexity comparison must control for activity skew; otherwise platform differences may be compositional

Activity Classes

1. Users are grouped by **activity level**:
   a. Low
   b. Medium
   c. High
   d. Very High
2. Heavy-tailed → “**average user**” is misleading
3. Each class is analyzed **separately** to remove volume bias

4. Curves for different platforms **overlap within each activity class**
   a. Only minor topic-dependent shifts (e.g. politics vs vaccines)
   b. Shifts disappear when controlling for volume

Activity Classes



Coherence with Zipf's law

Zipf's law: small number of words account for a large fraction of usage
→ Not a claim about meaning

$$f(r) \sim r^{-\propto}$$

→ r = rank of word type after sorting by frequency
→ f(r) = frequency of word at rank r
→ $\propto$ = controls how quickly frequency decays:
- → Larger $\propto$ means top words dominate more strongly

If alpha is stable across platforms, then platform design will not affect the fundamental frequency structure of lexical choice

Vocabulary evolution

Two competing patterns

→ early saturation: quickly reaches max vocabulary, then repeats

→ gradual expansion: continues adding words across comments

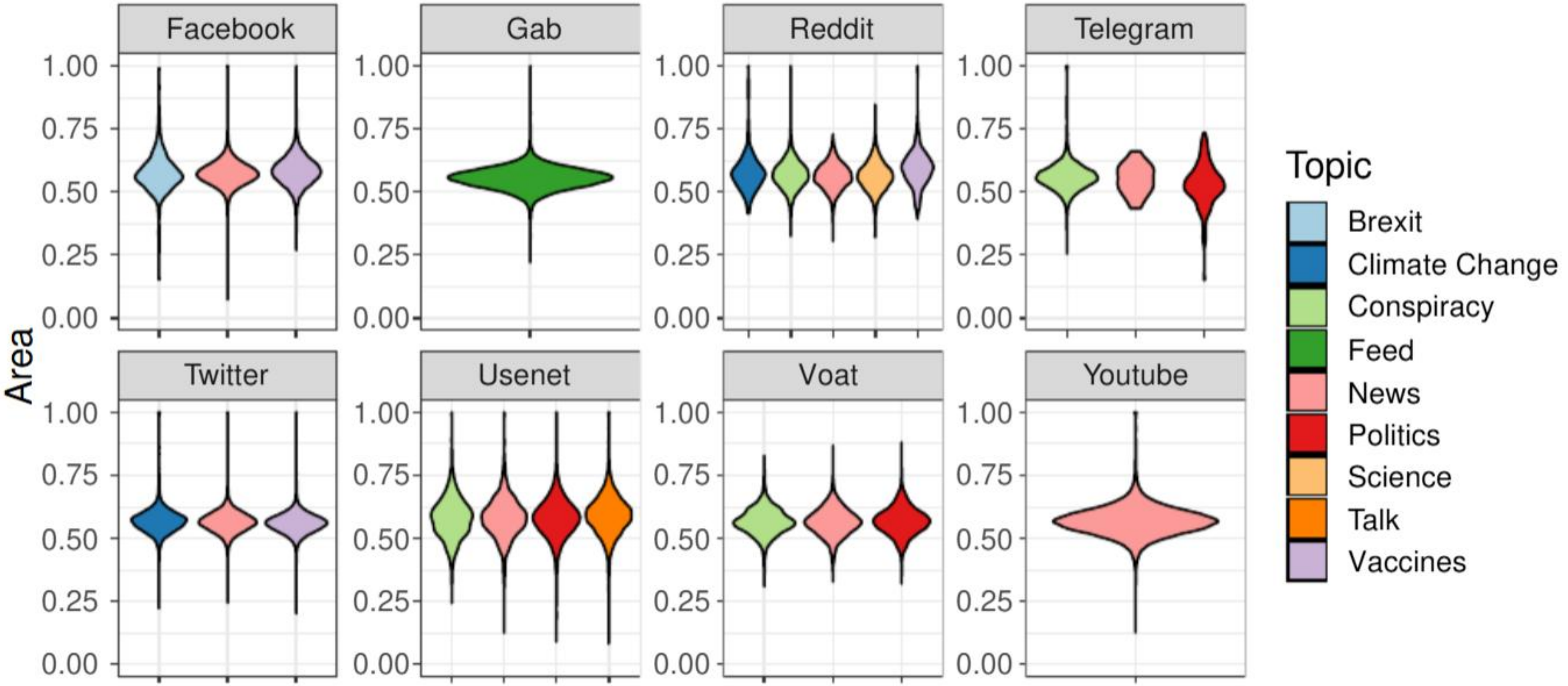
Comments ordered chronologically per user

→ Only uses with 25 -100 comments

→ avoid outliers and non active users

→ $V_u(i) =$ vector describing unique words used up to comment i

→ E.g. “hello, how are you?”, “are you going?” = $(4, 5)$



- Normalized to [0,1] → change from first to last comment
- Moderate increase → majority does not “freeze” vocabulary

Methods for complexity estimation

"Complexity" is **not directly observable**

1. No single scalar captures:
   a. vocabulary richness
   b. Repetition
   c. distributional inequality

Strategy of the paper: Use of a distribution-based metric

→ How unevenly words are used
   - **Gzip complexity** → inequality of word usage

→ How repetitive the text is
   - **Yule's K** → lexical repetition / richness

Two metrics in orthoganal dimensions → prevents one dimention

Gzip complexity

Gzip complexity measures how compressible a text is

→ high compressibility = high repetitiveness = low information density

→ Measure of redundance

$$g = \frac{S_{\text{raw}} - S_{\text{compressed}}}{S_{\text{raw}}}$$

1. S_{raw} = size of the original text
2. $S_{\text{compressed}}$ = size of the text after gzip compression

interpretation

→ Higher values for g = greater redundancy ?

→ Lower values for g = less repetitive and more information-dense

Yule’s K

High Yule’s K

→ few word types used many times

→ high repetition, low lexical richness

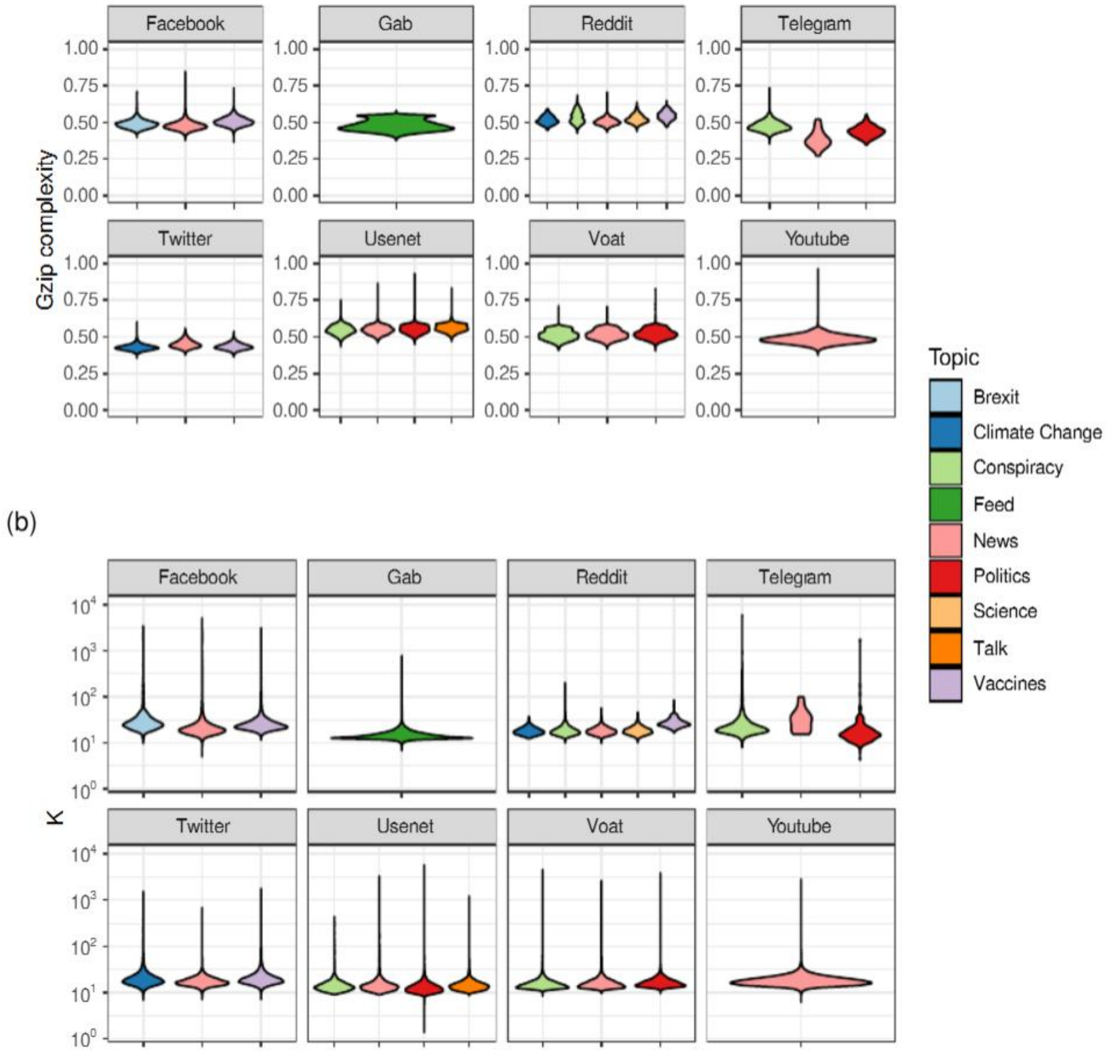
Low Yule’s K

→ Many different word types

→ Low repetition, high lexical richness

Unlike simple counts (e.g. number of words), Yule’s K reflects **structure**, not length

Two texts of equal length can differ strongly in Yule’s K depending on **how varied** their vocabulary is



Cross-platform similarity:

Distributions are broadly similar across platforms
→ **platform architecture is not the main driver** of linguistic complexity

Topic effects are secondary:
→ Different topics shift distributions slightly
→ **within-platform variation dominates**

Mann Whitney test

Temporal evolution:

User-aggregated docs cannot show how 1995 vs. 2020 comments differ

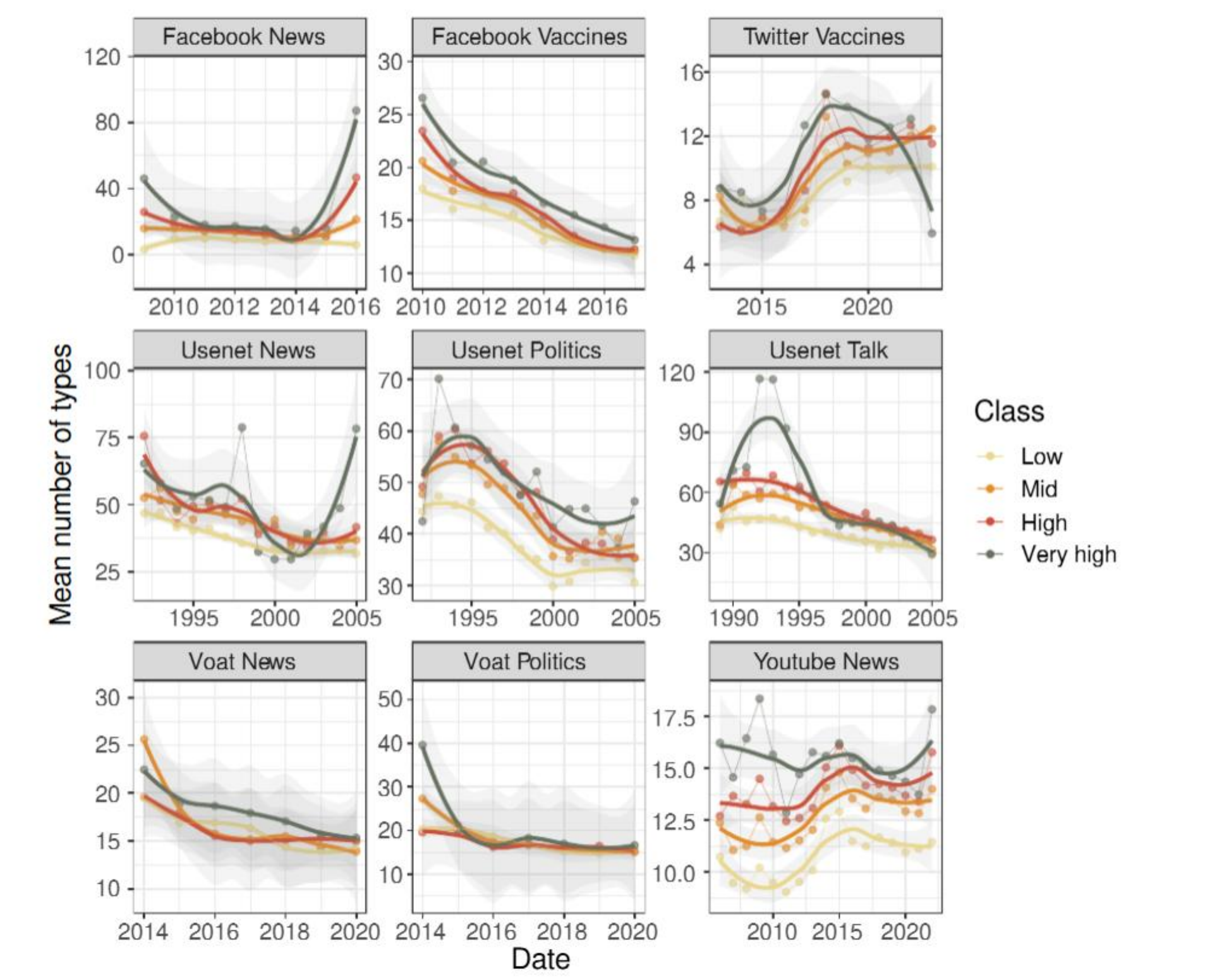
→ Paper analyzes yearly trends

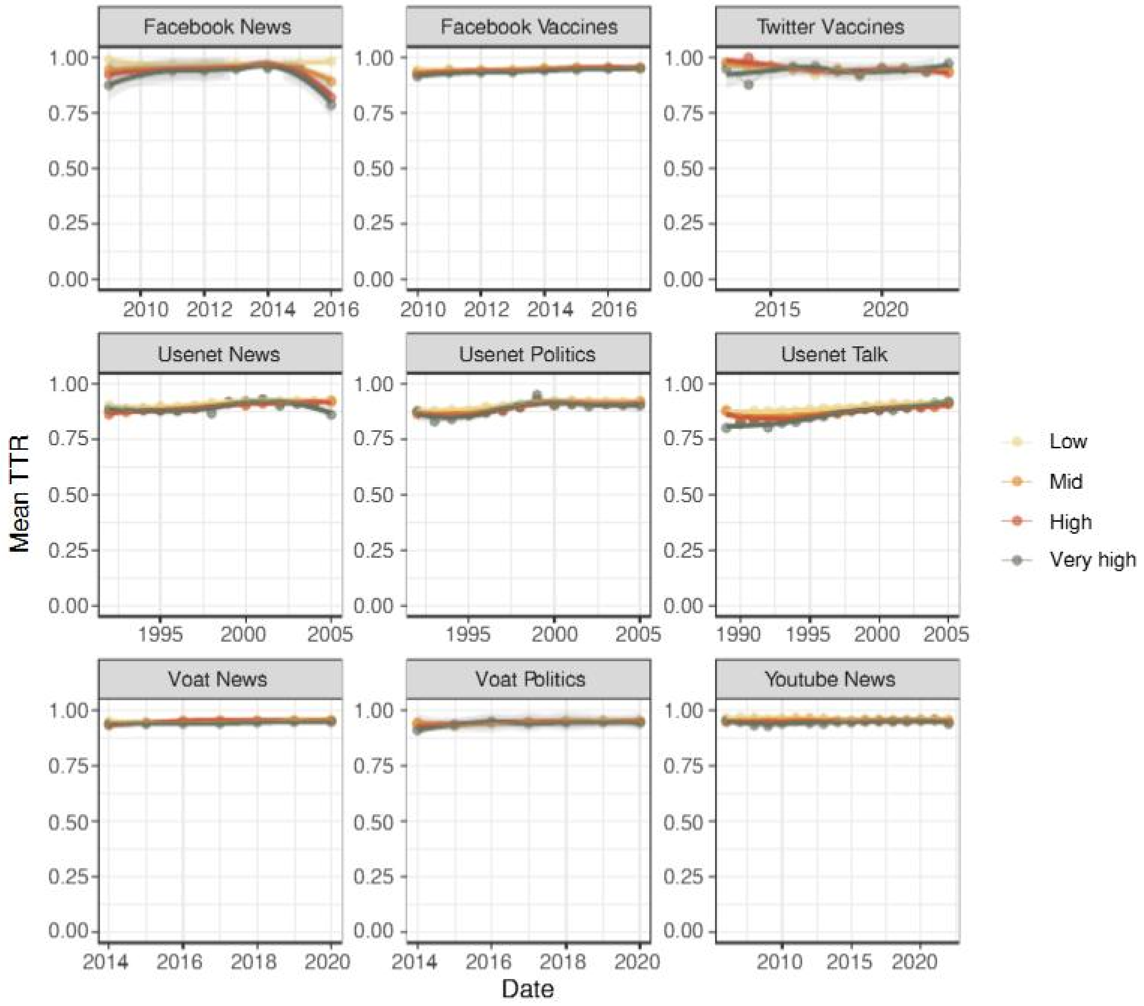
→ Control for activity

  - Each year, users are re-classified into activity classes to deuce bias

→ First temporal proxy

  - Mean number of types per comment across classes and years


Facebook News
Facebook Vaccines
Twitter Vaccines
Usenet News
Usenet Politics
Usenet Talk
Voat News
Voat Politics
Youtube News
Mean number of types
Date
Class
Low
Mid
High
Very high


Facebook News
Facebook Vaccines
Twitter Vaccines
Usenet News
Usenet Politics
Usenet Talk
Voat News
Voat Politics
Youtube News
Mean TTR
Date
Low
Mid
High
Very high
1.00
0.75
0.50
0.25
0.00
2010 2012 2014 2016
2015 2020
1990 1995 2000 2005
2014 2016 2018 2020

Interpretation

What does and does not change over time:

→ mean unique words decline
  → supports “reduced lexical richness over time

→ TTR calculation stay relatively stable
  → suggests types decline partly because comments are shorter (tokens also decline)

→ gzip complexity tends to decrease over time
  → recent comments less repetitive, despite reduced lexical richness

→ “simpler” is multi-dimensional
  → shorter and lexically poorer, but also less templated

Regression model

Goal: estimate overall relationship between year and text measure
→ Combining time + platform + complexity
→ Model using standardized predictors:

- number of types w ii
- Yule's $K^{`}_{ii}$
- gzip g ii

$$Y_i \sim [1\ w_i\ K_i\ g_i] \times B \times \begin{matrix} 1 \\ tw_i \\ vt_i \\ yt_i \\ un_i \end{matrix}$$

→ Platform effects: same linguistic features → different temporal trends
→ Unified comparison → All platforms are estimated in one model
→ Sign of coefficients directly tells direction of change over time

Regression Model

First vector:

→ $[1\ w_i\ K_i\ g_i]$, 1 – baseline year offset

Second vector:

	1	= Facebook (baseline)
	tw_i	= Twitter
→	vt_i	= Voat
	yt_i	= YouTube
	un_i	= Usenet

Each dummy is 1 if comment i is from that platform
0 otherwise

The coefficient matrix B

Rows:

1. Intercept
2. $w \quad {}_{ii}$ (types)
3. $K \quad {}_{ii}$
4. $g\beta_{ii}$

$$
B = \begin{matrix} \beta_0 & \beta_0, tw & \beta_0, vt & \beta_0, yt & \beta_0, un \\ \beta_1 & \beta_1, tw & \beta_1, vt & \beta_1, yt & \beta_1, un \\ \beta_2 & \beta_2, tw & \beta_2, vt & \beta_2, yt & \beta_2, un \\ \beta_3 & \beta_3, tw & \beta_3, vt & \beta_3, yt & \beta_3, un \end{matrix}
$$

Columns:

→ Platforms

e.g. for a Twitter comment $y_i =$

$(ß_0 + ß_0, tw) + (ß_1 + ß_1, tw)\, w\_i + (ß_2 + ß_2, tw)\, K\_i + (ß_3 + ß_3, tw)\, g\_$

Variables	Estimate	Total estimate	Standard error	p
β_0	2013.091	2013.091	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	2020.435	0.036	< 0.001
$\beta_{0,un}$	-13.644	1999.447	0.013	< 0.001
$\beta_{0,vt}$	4.376	2017.467	0.013	< 0.001
$\beta_{0,yt}$	2.076	2015.167	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

Table 1: Results of regression model.

β0

→ Roughly the **center year** of Facebook comments after normalization

β1 (types)

→ More unique words in **earlier years**

β2 (Yules K)

→ Higher K in **later years**

β3 (gzip complexity)

→ More compressible text in **later year**

Implication

Results argue against a single “social media ruins language” claim; the pattern is more specific:

→ comments become shorter

→ lexical richness declines

→ repetitiveness also declines

Platform influence is partial: distributions are strikingly similar across very different communities and topics

Conceptual implication

→ online language can be seen as communication under intensified constraints (speed, attention, interface)

Limitations

English-only dataset
→ morphological/syntactic complexity may behave differently in other languages

Metrics are lexical/compression-based
→ they do not directly capture syntax, semantics, argument quality, or pragmatic depth

Preprocessing removes emojis/hashtags and stems tokens
→ improves comparability but removes social-media-native signals

Platform datasets come from different collection pipelines
→ residual sampling bias and topic selection effects remain plausible

Conclusion

Across decades and platforms, lexical behavior shows strong cross-context regularities
→ evidence for universal patterns in online language production

Comments become:
→ shorter and less lexically rich
→ less repetitive
suggesting compressed expression rather than templated copying

Users keep adding new words steadily (AUC peak ~0.6)
→ consistent with continued lexical exploration rather than early saturation

Theoretical framing: Zipf + efficiency + trade-off models help interpret these trends as constraint-driven adaptation, not linguistic collapse

References

1. https://www.researchgate.net/publication/381485530_The_Evolution_of_Language_in_Social_Media_Comments#pf17