Social Sciences and Humanities
ISSN Print: 3104-4328
ISSN Online: 3104-4336
CD ISSN: 3104-4344
https://doi.org/10.63313/SSH.9126

Early Comments and Later Incivility on YouTube: A Chronological Analysis of Comment Sections

Hongming Miao
College of Publishing, University of Shanghai for Science and Technology, Shanghai 200093, China
Email: kerwinscavo@gmail.com

How to cite this paper: Miao, H. (2026). Early Comments and Later Incivility on YouTube: A Chronological Analysis of Comment Sections. Social Sciences and Humanities, 5(1), 28–36. ISSN Print: 3104-4328, ISSN Online: 3104-4336.
https://doi.org/10.63313/SSH.9126
Published: 2026-08-12

Copyright © 2026 by author(s) and Erytis Publishing Limited.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
CC BY 4.0

Abstract

Online-incivility research often examines isolated comments or aggregate comment-section tone, leaving the temporal ordering of real comment sections less studied. This study tests whether incivility among chronologically early, currently retrievable YouTube comments is associated with later comment tone in the same videos. Using the YouTube Data API and programming scripts for complete pagination, sorting, and variable construction, the study analyzes 39,992 public top-level comments from 30 controversy-oriented videos in three themes. Early incivility is the share of uncivil comments among the first 10 retrievable comments after within-video ordering by publication timestamp. A rule-based coding procedure identifies disrespectful expression while distinguishing it from ordinary criticism. Overall, 17.1% of comments were classified as uncivil. Across videos, early and later incivility were positively correlated (Pearson r = 0.438; Spearman rho = 0.465). In a theme-controlled OLS model, early incivility remained positively associated with later incivility (b = 0.421, robust SE = 0.158, p = .013, 95% CI [0.096, 0.746]). The findings support a temporal, comment-section-level perspective but do not establish causal influence or individual exposure to early comments.

Keywords

YouTube Comments; Online Incivility; Social Norms; Comment Sections; Comment Analysis

1. Introduction

Online comment sections are consequential interaction spaces in which users encounter praise, criticism, humor, corrections, insults, and attacks. Incivility is not equivalent to disagreement or negative evaluation. Robust criticism can remain civil when it addresses ideas, performances, policies, or products without demeaning people or groups. Research on online and political discourse therefore separates civility from mere politeness and treats insults, deception, personal attack, and conduct that shuts down inclusive discussion as distinct dimensions of incivility [1,2]. Content analyses also show that incivility occurs regularly in online news comments and varies with contextual factors [3]. On YouTube, flaming has been

described more narrowly as hostility expressed through insults, swearing, or other offensive language; users report that such behavior is common and partly connected to perceived local norms [4]. These distinctions guide the present study: controversy is not itself coded as uncivil, and strong criticism is classified as uncivil only when it contains disrespectful or attacking expression.

YouTube provides a useful setting because comments are attached to audiovisual content, accumulate over time, and connect topic-specific audiences. A large content analysis demonstrates that YouTube comments serve heterogeneous communicative purposes and can be classified systematically [5]. Comments can also become information for viewers rather than remaining mere reactions. In an experiment concerning nanotechnology, exposure to uncivil blog comments polarized risk perceptions conditional on readers' predispositions [6]. In a YouTube experiment, supportive or derisive comments changed evaluations of an antimarijuana public-service announcement, although effects on attitudes were more conditional [7]. Such studies justify examining comment environments while cautioning against treating every comment effect as uniform.

The platform-specific context also complicates comparison. YouTube comments can respond to a video's argument, a creator's persona, a featured person, a fandom dispute, or a controversy that viewers already encountered elsewhere. A game-review dispute may activate identity and media-trust claims, whereas an AI/creator dispute may activate arguments about labor, originality, or artistic legitimacy. These differences make cross-topic comparison informative, but they also mean that a platform-wide label such as 'YouTube comments' should not be treated as a single normative environment.

Social-norm and social-influence theories offer one explanation for why early comment-section tone may be associated with later tone. Focus theory distinguishes descriptive norms, or perceptions of what others do, from injunctive norms, or perceptions of what others approve [8]. Normative influence concerns conformity for acceptance or to avoid disapproval, whereas informational influence concerns reliance on others as evidence about reality [9]. Online bandwagon experiments likewise show that qualitative cues in comments can shift readers' opinions through perceived public opinion [10]. Applied to a comment section, early uncivil expression could be interpreted as a cue that hostility is common or tolerated. This is a theoretical possibility, not an observed psychological mechanism in the present data, because the study does not measure whether later commenters saw or interpreted any early comment.

More directly, online-incivility research describes discussion tone as dynamic and contextual. Nearby incivility and positive feedback are associated with repeated incivility, supporting a model in which descriptive and injunctive signals interact [11]. Experimental and longitudinal evidence also shows that negative mood and exposure to prior trolling can increase subsequent trolling [12]. Conversely, a large

field experiment found that making community rules visible reduced unruly behavior and increased newcomer participation [13]. Together, these studies motivate attention to temporal sequence while also showing that mood, selection, moderation, and other contextual processes may produce similar patterns.

Measurement introduces a second boundary. Simple lexical rules can confuse offensive terms with hate speech and miss messages whose meaning depends on context [14]. A systematic review of abusive-language datasets documents conceptual inconsistency, annotation bias, domain mismatch, and the frequent omission of conversation context [15]. The current classifier is therefore rule-based and auditable rather than a machine-learning model. Its output is used as an approximate indicator of disrespectful expression, not as a definitive judgment about a commenter's intent or social harm.

Existing studies establish that comments can affect evaluations, that local norms can shape conduct, and that YouTube comments are heterogeneous. Less work examines whether the earliest retrievable portion of real YouTube comment sections is associated with later aggregate tone after preserving within-video time order. This study addresses that gap with a video-level early-later design across three controversy-oriented themes. It asks: RQ1, what is the prevalence of incivility in the collected comments? H1, do videos with more incivility among their first 10 retrievable comments also have more incivility among later comments? RQ2, how do observed rates vary across the three themes? RQ3, is the association robust when early comments are defined as the first 5, 10, or 20 retrievable comments?

2. Data and Methods

2.1. Sample and collection

The analytic sample contains 30 public YouTube videos, with 10 videos in each of three controversy-oriented themes: AI-generated content and creator disputes; comedy, speech-boundary, and cultural controversies; and game and media-review controversies. Videos were purposively selected for open comment sections, English-dominant comment activity, at least 50 public top-level comments, and substantive disagreement rather than purely instructional or uniformly celebratory content. Shorts, livestreams and likely livestream replays, children's content, and obvious spam-heavy sections were excluded. Twenty-four API query combinations yielded 520 initial candidate records; screening considered topical fit, English-comment share, channel diversity, and feasibility of complete pagination.

Candidate screening combined video metadata with preliminary review of comment activity. Additional exclusions covered videos published less than seven days before screening, weak controversy signals, unusually repetitive screening comments, and more than two final videos from the same channel. These rules supported topical relevance and channel diversity, but the result remains a purposive rather than probability sample. It is designed to compare selected controversy environments,

not to estimate population prevalence for YouTube.

Public top-level comments were collected through the YouTube Data API and programming scripts. Every included video was paginated until the API returned no remaining nextPageToken. The final analytic file contains 39,992 comments. Replies were not collected. Four initially selected, exceptionally high-volume videos were replaced because complete pagination was infeasible within the available API quota; replacements came from the same themes, met the original criteria, preserved channel diversity, and were selected without using incivility rates. The analysis concerns this fully paginated final set only.

2.2. Comment order and early windows

Within each video, all collected comments were sorted in ascending order using the publication timestamp returned by the API, and comment_order was assigned from 1. The main early window is the first 10 comments in that local sequence; robustness checks use the first 5 and first 20. Later comments exclude the comments in the relevant early window, so predictor and outcome do not share observations. 'Early comments' means the early public top-level comments that remained retrievable through the API at collection, not every comment originally posted. Deleted, hidden, held-for-review, moderated, filtered, or otherwise inaccessible comments cannot be recovered. API-based chronological order also differs from the ranked order that a user may see in the YouTube interface.

2.3. Incivility coding and manual audit

A rule-based coding procedure classified comments at three levels: 0 for civil or neutral expression, including ordinary criticism without disrespect; 1 for mild sarcasm, belittling, provocation, or disrespect; and 2 for direct insult, personal attack, vulgar abuse, threat, hate expression, or strong hostility. The main binary indicator combines Levels 1 and 2 as uncivil. Transparent lexical and pattern rules covered insults, profanity, demeaning constructions, hostile imperatives, and target cues. The procedure did not estimate machine-learning parameters on the analytic dataset.

Performance was checked against a manually coded sample of 400 comments. For binary incivility, accuracy was 0.890, precision was 0.846, recall was 0.759, and F1 was 0.800; the confusion matrix was TN = 268, FP = 16, FN = 28, and TP = 88. Lower recall than precision indicates that some uncivil comments were probably missed. Sarcasm, implicit hostility, coded attacks, and context-dependent profanity remain difficult for rules to identify.

2.4. Statistical analysis

Analysis was conducted at the video level (N = 30). Pearson and Spearman correlations assessed the early-later relationship for the first-5, first-10, and first-20

definitions. The main OLS model predicted later incivility from first-10 incivility and two theme indicators, using AI/creator disputes as the reference category. HC1 robust standard errors were used because the video-level sample is small and residual variance may differ across videos. All coefficients are interpreted as associations.

Comment-level models were not used as the main specification because treating 39,992 comments as independent would ignore their clustering within only 30 videos. A multilevel model could represent that structure in a larger sample, but the present video-level N is too small for a complex random-effects specification. The analyses therefore match the theoretical unit—the video-level comment section—while reporting both linear and rank-based associations.

3. Results

Across 39,992 comments, 6,820 were classified as uncivil, for an overall rate of 17.1%. Table 1 combines sample composition and theme-level results. AI/creator disputes had an incivility rate of 20.2%, game/media-review controversies 19.7%, and comedy/speech-boundary controversies 7.8%. The corresponding Level 1 and Level 2 rates were 15.2% and 4.9% for AI/creator disputes, 10.9% and 8.8% for game/media-review controversies, and 6.0% and 1.9% for comedy/speech-boundary controversies. These are descriptive results for a purposive sample, not platform-wide estimates.

The composition of the dataset also differed by theme. Game/media-review videos contributed 42.3% of comments, AI/creator videos 34.0%, and comedy/speech-boundary videos 23.7%. Accordingly, the overall comment-level rate is weighted toward videos with more retrievable comments and is not the same quantity as the unweighted average of video-level rates used in the early-later analysis. The theme contrasts describe the collected corpus and do not test whether topic alone accounts for the differences.

Table 1. Sample Composition and Incivility by Theme

Theme	Videos	Comments	Share	Uncivil	Rate
AI/creator disputes	10	13,598	34.0%	2,742	20.2%
Game/media-review	10	16,928	42.3%	3,337	19.7%
Comedy/speech-boundary	10	9,466	23.7%	741	7.8%
Total	**30**	**39,992**	**100.0%**	**6,820**	**17.1%**

For the main first-10 definition, mean early incivility across videos was 10.3% and mean later incivility was 18.3%. The 7.9-percentage-point difference is descriptive; because the two portions of a comment section differ in size and timing, it is not an estimate of an early-comment effect. Early and later rates were moderately positively associated (Pearson r = 0.438; Spearman rho = 0.465).

Table 2 shows that the direction was positive for all windows. The first-5 Pearson

correlation was weaker (r = 0.192), whereas the corresponding rank correlation was 0.385. First-20 correlations were stronger (r = 0.591; rho = 0.615). Five comments provide a noisier estimate of early comment-section tone, while 20 comments may capture tone more stably but are less strict as an 'earliest' window. The first-10 definition remains the prespecified main comparison; the robustness results support a common positive direction, not a claim that the coefficient should be identical across window sizes.

Table 2. Robustness across Early-Comment Definitions

Early window	N	Early mean	Later mean	Difference (pp)	Pearson r	Spearman rho
First 5	30	8.7%	18.2%	9.5	0.192	0.385
First 10	30	10.3%	18.3%	7.9	0.438	0.465
First 20	30	10.7%	18.5%	7.8	0.591	0.615

The theme-controlled OLS model in Table 3 yielded a positive first-10 coefficient (b = 0.421, HC1 SE = 0.158, t = 2.663, p = .013, 95% CI [0.096, 0.746]). Thus, a 10-percentage-point increase in early incivility was associated with about a 4.2-percentage-point increase in later incivility, holding broad theme constant. The model explained 27.2% of the variance (adjusted R2 = 0.188; N = 30; df = 26). Neither theme coefficient was statistically distinguishable from zero in this small model. Their uncertainty should not be read as evidence that themes are equivalent; the model has limited power and uses only broad indicators. H1 therefore received qualified, associational support.

Table 3. Video-Level OLS Regression

Predictor	b	HC1 SE	t	p	95% CI
Intercept	0.148	0.033	4.480	< .001	[0.080, 0.216]
Early incivility rate	0.421	0.158	2.663	.013	[0.096, 0.746]
Comedy/speech-boundary	-0.043	0.036	-1.201	.241	[-0.118, 0.031]
Game/media-review	0.018	0.036	0.486	.631	[-0.057, 0.092]
Model information	R2 = 0.272	Adj. R2 = 0.188	N = 30	df = 26	Reference: AI/creator

4. Discussion

Videos with more incivility among their first 10 retrievable comments tended to have more incivility later, and the positive association persisted after broad theme controls. The result was also positive under first-5 and first-20 windows, although magnitudes varied. This pattern is compatible with social-norm accounts in which observed conduct informs perceptions of what is common or acceptable [8,11]. It is also consistent with evidence that exposure to prior trolling can shape later behavior [12]. Compatibility is the appropriate inference: the present study neither manipulated comment-section tone nor measured whether later commenters

encountered the early comments.

The topic comparison provides context rather than a universal ranking of YouTube communities. AI/creator and game/media-review controversies had higher descriptive rates than the comedy/speech-boundary theme in this purposive sample, while the regression theme indicators were imprecise. Topic cultures, audience composition, creator framing, channel reputation, external controversy, and moderation could affect both early and later tone. The positive early-later coefficient may therefore reflect stable video-level differences or audience selection as well as any local norm process. The small model controls only broad theme and cannot separate these explanations.

Preserving temporal order is nevertheless useful. Treating comments as an unordered corpus discards information about how a retrievable discussion record develops. The early-window design complements studies that measure aggregate prevalence [3], dynamic nearby-comment patterns [11], and experimentally manipulated discussion context [12]. It also clarifies what stronger evidence would require: experiments that manipulate early tone while holding video content constant, quasi-experiments around moderation changes, or designs linking chronological order to actual displayed order and user exposure.

This sequencing perspective also sharpens the distinction between retrievability and exposure. Publication timestamps permit reconstruction of a local chronological record, but the displayed comment section is shaped by ranking, engagement, moderation, and individual sorting choices. Chronologically early comments may be buried, promoted, or removed before later users arrive. A future design that combines timestamps with displayed position, engagement signals, and moderation events would test the proposed mechanism more directly than chronology alone.

The findings have cautious implications for creators, moderators, and platforms. Early comment-section environments may merit monitoring, and visible community standards have reduced unruly behavior in other large online discussions [13]. However, this analysis did not test moderation, pinned comments, creator replies, ranking interventions, or community-guideline treatments. It therefore cannot establish that acting on the first comments will change subsequent behavior. Those interventions should be evaluated directly before operational recommendations are made.

Several limitations define the study's scope. First, its observational design cannot establish causality, and not all video-level confounders were measured. Second, user exposure was not observed; a later commenter may see top-ranked, recent, notified, or no prior comments. Third, API-retrievable chronological order is not the interface display order. Fourth, deleted, hidden, moderated, held-for-review, or otherwise inaccessible comments cannot be recovered. Fifth, replies were excluded even though they may contain direct interpersonal conflict. Sixth, the rule-based coder cannot fully capture sarcasm, implicit attacks, pragmatic context, or nonhostile

profanity, as anticipated by research on automated language classification [14,15]. Seventh, the 30-video, controversy-oriented sample is small and does not represent YouTube generally. Finally, the first 10 comments can span very different clock-time intervals across videos; comment-count robustness checks do not substitute for hour- or day-based windows.

5. Conclusion

In 30 controversy-oriented YouTube videos, incivility among the first 10 currently retrievable public top-level comments was positively associated with incivility among later comments. Alternative early windows and a theme-controlled model preserved the positive direction. The contribution is a comment-section-level temporal perspective that distinguishes early from later retrievable comments. The result is associative, applies only to the purposive sample and rule-based measure, and does not show that later commenters saw or were influenced by early comments.

References

[1] Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. New Media & Society, 6(2), 259-283. https://doi.org/10.1177/1461444804041444

[2] Stryker, R., Conway, B. A., & Danielson, J. T. (2016). What is political incivility? Communication Monographs, 83(4), 535-556. https://doi.org/10.1080/03637751.2016.1201207

[3] Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. Journal of Communication, 64(4), 658-679. https://doi.org/10.1111/jcom.12104

[4] Moor, P. J., Heuvelman, A., & Verleur, R. (2010). Flaming on YouTube. Computers in Human Behavior, 26(6), 1536-1546. https://doi.org/10.1016/j.chb.2010.05.023

[5] Madden, A., Ruthven, I., & McMenemy, D. (2013). A classification scheme for content analyses of YouTube video comments. Journal of Documentation, 69(5), 693-714. https://doi.org/10.1108/JD-06-2012-0078

[6] Anderson, A. A., Brossard, D., Scheufele, D. A., Xenos, M. A., & Ladwig, P. (2014). The 'nasty effect': Online incivility and risk perceptions of emerging technologies. Journal of Computer-Mediated Communication, 19(3), 373-387. https://doi.org/10.1111/jcc4.12009

[7] Walther, J. B., DeAndrea, D., Kim, J., & Anthony, J. C. (2010). The influence of online comments on perceptions of antimarijuana public service announcements on YouTube. Human Communication Research, 36(4), 469-492. https://doi.org/10.1111/j.1468-2958.2010.01384.x

[8] Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. Journal of Personality and Social Psychology, 58(6), 1015-1026. https://doi.org/10.1037/0022-3514.58.6.1015

[9] Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. The Journal of Abnormal and Social Psychology, 51(3), 629-636. https://doi.org/10.1037/h0046408

[10] Lee, S., Atkinson, L., & Sung, Y. H. (2022). Online bandwagon effects: Quantitative versus qualitative cues in online comments sections. New Media & Society, 24(3), 580-599.

https://doi.org/10.1177/1461444820965187

[11] Shmargad, Y., Coe, K., Kenski, K., & Rains, S. A. (2022). Social norms and the dynamics of online incivility. Social Science Computer Review, 40(3), 717-735. https://doi.org/10.1177/0894439320985527

[12] Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2017). Anyone can become a troll: Causes of trolling behavior in online discussions. Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, 1217-1230. https://doi.org/10.1145/2998181.2998213

[13] Matias, J. N. (2019). Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. Proceedings of the National Academy of Sciences, 116(20), 9785-9789. https://doi.org/10.1073/pnas.1813486116

[14] Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. Proceedings of the International AAAI Conference on Web and Social Media, 11(1), 512-515. https://doi.org/10.1609/icwsm.v11i1.14955

[15] Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. PLOS ONE, 15(12), e0243300. https://doi.org/10.1371/journal.pone.0243300