Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

What Makes a Review Read as Positive or Negative?
Linguistic Signatures of Valence in Online Consumer and Cultural-Product Reviews

Senthil kumar Anantharaman

senthilkumara1851@gmail.com

Independent Researcher https://orcid.org/0009-0002-8527-3288

Research Article

Keywords: Computational Linguistics, Sentiment Analysis, Electronic Word-of-Mouth (eWOM), Readability Metrics, Consumer Reviews, Linguistic Signatures, Cross-Category Calibration

Posted Date: July 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10425058/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Automated sentiment tools increasingly decide which online reviews get surfaced, flagged, or acted on but the linguistic markers these tools rely on may not transfer across product categories. This study tests whether six measurable text features (sentiment intensity, readability, length, lexical diversity, affective punctuation, average word length) reliably distinguish positive from negative reviews across two independent corpora: 2,000 film reviews (Pang & Lee, 2004) and 294 consumer-electronics reviews (Hu & Liu, 2004). Using t-tests, Mann-Whitney U tests, point-biserial correlations, and multivariate logistic regression (with VIF checks), we find sentiment intensity, readability, and lexical diversity distinguish valence consistently in both domains safe, portable signals for cross-category models. Affective punctuation, however, reverses direction entirely between domains: exclamation marks mark negativity in film reviews but positivity in product reviews, meaning punctuation-based sentiment rules cannot be deployed unvalidated across categories. The full feature set explains a modest-to-moderate share of variance (pseudo-R^2 = .12, .26), with sentiment intensity the only predictor robust to joint modelling in both domains. We outline concrete implications for review-triage systems, category-aware sentiment calibration, and readability-based ranking heuristics.

1. Introduction

Online reviews are now a primary input into consumer decision-making and, correspondingly, into firm strategy: review volume and valence significantly impact sales. (Chevalier & Mayzlin, 2006). Review characteristics predict how helpful other consumers find a review (Mudambi & Schuff, 2010), and the emotional register of a piece of content predicts whether it is shared at all (Berger & Milkman, 2012). Firms, platforms, and marketplaces therefore have a direct commercial interest in being able to characterize review valence, not only to summarize customer sentiment at scale, but to triage reviews for moderation, prioritize customer-service response, and understand which categories of product elicit which kinds of evaluative language.

In practice, most automated systems for characterizing review valence lean heavily on sentiment lexicons and classifiers, treating valence as a property that a sentiment score alone can recover. This study tests whether that reliance is well founded, by asking whether textual features beyond sentiment, readability, length, lexical diversity, affective punctuation, and average word length carry independent, non-redundant signal about a review's valence, and whether the relationships between these features and valence are stable across product category or are instead category-specific. Both questions matter operationally: a feature that predicts valence in one domain but reverses in another is not a safe input for a one-size-fits-all classification or moderation rule.

We test this using two independent, publicly available review corpora spanning different domains, film reviews and consumer-electronics reviews that lets us distinguish linguistic signatures of valence that generalize across category from those that are category-specific. We operationalize review "linguistic

profile" through seven computationally tractable features and test whether they predict self- or annotator-derived valence individually and jointly, across both corpora.

2. Theoretical Background and Hypotheses

Three literatures motivate the present hypotheses. First, research on electronic word-of-mouth (eWOM) and review helpfulness establishes that reviews vary systematically in structural and stylistic properties i.e, length, extremity, depth and that these properties carry commercial consequences independent of star rating alone (Chevalier & Mayzlin, 2006; Mudambi & Schuff, 2010). This literature motivates treating a review's linguistic profile, not just its numeric rating, as a variable of managerial interest. Second, the Elaboration Likelihood Model (ELM) proposes that message features associated with deeper cognitive processing greater length, lower readability, more distinctive vocabulary are more likely to accompany strongly held, elaborated attitudes, whether positive or negative (Petty & Cacioppo, 1986); applied to reviews, this predicts that valence should leave a detectable signature in structural complexity, not only in sentiment words. Third, negativity bias research suggests that negative evaluations recruit more vivid, differentiated, and emotionally marked language than positive ones (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Rozin & Royzman, 2001), and work on viral and shared content shows that the relationship between emotion and audience response is category- and context-dependent rather than fixed (Berger & Milkman, 2012) motivating an explicit test of whether linguistic-valence relationships generalize across product category or are category-specific. Accordingly, we advance the following hypotheses, each stated with its original(null) hypothesis

H1 — Sentiment intensity

H1: Reviews expressing positive valence will show significantly higher sentiment polarity (VADER compound score) than reviews expressing negative valence. $H1_0$: mean compound scores do not differ by valence.

H2 — Readability

H2: Readability (Flesch Reading Ease) will differ significantly between positive- and negative-valence reviews. $H2_0$: readability does not differ by valence.

H3 — Length

H3: Positive-valence reviews will be significantly longer (word count) than negative-valence reviews. $H3_0$: length does not differ by valence.

H4 — Lexical diversity

H4: Lexical diversity (type–token ratio) will differ significantly between positive- and negative-valence reviews. $H4_0$: lexical diversity does not differ by valence.

H5 — Affective punctuation

H5: The frequency of exclamation marks will differ significantly between positive- and negative-valence reviews. $H5_0$: exclamation-mark frequency does not differ by valence.

H6 — Combined predictive model

H6: A linear combination of linguistic features (sentiment, readability, length, lexical diversity, punctuation, word length) will predict review valence significantly better than an intercept-only model — i.e., structural and stylistic features carry incremental signal beyond sentiment alone. $H6_0$: the full model does not improve fit over the null model.

3. Method

3.1 Corpora

Corpus 1 (film reviews). Data were drawn from the Sentiment Polarity Dataset v2.0 (Pang & Lee, 2004), a public corpus of 2,000 English-language film reviews originally posted to the rec.arts.movies.reviews newsgroup and archived via the Internet Movie Database. The corpus is balanced by construction: 1,000 reviews labeled positive and 1,000 labeled negative, with valence extracted algorithmically from each reviewer's own star or letter-grade rating at the time of posting.

Corpus 2 (product reviews). To test whether the Corpus 1 pattern generalized to a commercially distinct review category, we added the Hu and Liu (2004) customer-review corpus: 313 English-language consumer reviews of five electronics products (a DVD player, a digital camera, an MP3 player, a second digital camera, and a mobile phone), collected from Amazon.com and C|net. Because this corpus scores sentiment strength (− 3 to + 3) at the product-feature level rather than providing a document-level star rating, we derived a document-level valence label by summing feature-sentiment scores per review and classifying reviews with a positive net sum as positive-valence and a negative net sum as negative-valence; reviews with no scored features (n = 16), a net sum of exactly zero (n = 8), or fewer than 15 words after exclusion were dropped, yielding 294 usable reviews (187 positive, 107 negative).

3.2 Measures

For each review in both corpora, we computed: (a) sentiment intensity, the compound polarity score from the VADER rule-based sentiment lexicon (Hutto & Gilbert, 2014); (b) readability, indexed by the Flesch Reading Ease and Flesch–Kincaid Grade Level formulas (Flesch, 1948); (c) length, total word count; (d) lexical diversity, the type–token ratio; (e) affective punctuation, the raw count of exclamation marks; and (f) average word length in characters, a common proxy for lexical/orthographic complexity (Biber, 1988). For Corpus 2, review text was reconstructed by concatenating the corpus's pre-tokenized sentences; readability and sentiment scores computed on reconstructed text should be treated as close approximations rather than exact values.

3.3 Analytic strategy

Group differences (positive vs. negative valence) were tested separately in each corpus with Welch's independent-samples t-test and cross-validated with the non-parametric Mann-Whitney U test; standardized effect sizes are reported as Cohen's d. Point-biserial correlations were computed between each feature and valence (coded 0/1). To test H6, a multivariate binary logistic regression predicted valence from all six standardized features simultaneously, run independently within each corpus, with variance inflation factors (VIFs) computed for every regression. Analyses were conducted in Python (SciPy and statsmodels); α was set at .05, two-tailed. Corpus 2 results are treated as a replication/generalization check on the Corpus 1 findings rather than an independent confirmatory test of identical effect sizes, given its smaller sample and different valence-derivation method.

4. Results

4.1 Corpus 1 (film reviews, N = 2,000)

Descriptive statistics and univariate test results for all seven features are reported in Table 1. Every feature differed significantly between positive- and negative-valence reviews (all $p < .001$; $|d| = 0.20–0.67$).

Table 1
Descriptive Statistics and Univariate Tests, Corpus 1 (Film Reviews, N = 2,000)

Feature	Positive M (SD)	Negative M (SD)	t (df ≈ 1998)	p	Cohen's d
VADER sentiment compound	0.646 (0.702)	0.107 (0.904)	14.89	< .001	0.67
Flesch Reading Ease	57.18 (9.14)	59.67 (8.22)	−6.40	< .001	−0.29
Flesch–Kincaid grade level	10.46 (2.25)	9.89 (2.07)	6.00	< .001	0.27
Review length (words)	832.6 (372.3)	751.3 (315.4)	5.27	< .001	0.24
Lexical diversity (TTR)	0.520 (0.066)	0.538 (0.071)	−6.01	< .001	−0.27
Exclamation marks (count)	0.66 (1.63)	1.06 (2.37)	−4.39	< .001	−0.20
Average word length	4.51 (0.22)	4.45 (0.20)	6.23	< .001	0.28

Sentiment intensity showed the largest effect (H1 supported, d = 0.67). Positive reviews were reliably less readable — lower Flesch Reading Ease and higher Flesch–Kincaid grade level — than negative reviews (H2 supported, in a direction that a simple processing-fluency account would not predict). Positive reviews were also significantly longer (H3 supported, d = 0.24) and less lexically diverse than

negative reviews (H4 supported). Negative reviews used exclamation marks nearly twice as often as positive reviews (H5 supported, d = − 0.20).

Point-biserial correlations were consistent in sign and significance with the t-test results (sentiment r = .32; readability r = − .14; grade level r = .13; length r = .12; lexical diversity r = − .13; exclamation marks r = − .10; word length r = .14; all p < .001).

The multivariate logistic regression (H6) is reported in Table 3. The full model significantly improved on the intercept-only model, $\chi^2(6) = 331.4$, $p < .001$, pseudo-$R^2 = .12$. With all predictors entered simultaneously, sentiment compound score (B = 0.67, p < .001), lexical diversity (B = − 0.62, p < .001), average word length (B = 0.40, p < .001), review length (B = − 0.20, p = .024), and exclamation-mark frequency (B = − 0.15, p = .007) remained significant unique predictors of valence; Flesch Reading Ease did not retain unique predictive value once the other features were controlled (B = − 0.12, p = .167). All VIFs were below 3.4 (range 1.01–3.37).

Table 3
Multivariate Logistic Regression Predicting Valence, Corpus 1 (Film Reviews)

Predictor	B	SE	z	P	95% CI
Constant	−0.023	0.049	−0.47	.637	[− 0.119, 0.073]
Sentiment compound (VADER)	0.673	0.051	13.22	< .001	[0.573, 0.772]
Flesch Reading Ease	−0.116	0.084	−1.38	.167	[− 0.280, 0.048]
Lexical diversity (TTR)	−0.620	0.095	−6.50	< .001	[− 0.806, − 0.433]
Review length (words)	−0.202	0.089	−2.26	.024	[− 0.376, − 0.027]
Exclamation marks	−0.151	0.056	−2.71	.007	[− 0.260, − 0.042]
Average word length	0.398	0.086	4.66	< .001	[0.231, 0.566]

4.2 Corpus 2 (product reviews, N = 294) — replication and generalization check

Table 2 reports the equivalent results for the product-review corpus. The sentiment effect replicated closely (H1 supported, d = 1.33). Readability and lexical-diversity effects also replicated in direction and significance: positive reviews were again less readable (H2 supported, d = − 0.29) and less lexically diverse (H4 supported, d = − 0.41) than negative reviews. Review length was directionally consistent but did not reach significance (H3 not supported at α = .05; point-biserial p = .095). Two effects diverged from Corpus 1: exclamation marks were more frequent in positive reviews here (d = 0.29), the reverse of the film-corpus pattern (H5 not supported in the same direction), and average word length showed no reliable difference (d = 0.08, ns).

Table 2
Descriptive Statistics and Univariate Tests, Corpus 2 (Product Reviews, N = 294)

Feature	Positive M (SD)	Negative M (SD)	t	p	Cohen's d
VADER sentiment compound	0.872 (0.245)	0.274 (0.674)	8.85	< .001	1.33
Flesch Reading Ease	70.68 (9.35)	73.46 (9.81)	−2.38	.018	−0.29
Flesch–Kincaid grade level	7.62 (2.30)	6.84 (2.31)	2.79	.006	0.34
Review length (words)	215.9 (178.2)	177.7 (204.7)	1.61	.108	0.20
Lexical diversity (TTR)	0.638 (0.121)	0.689 (0.134)	−3.25	.001	−0.41
Exclamation marks (count)	0.97 (1.47)	0.58 (1.14)	2.53	.012	0.29
Average word length	4.16 (0.27)	4.14 (0.27)	0.63	.529	0.08

The multivariate logistic regression for Corpus 2 (Table 4) also reached significance overall, $\chi^2(6) = 101.2$, $p < .001$, pseudo-$R^2 = .26$. Once all predictors were entered simultaneously, only sentiment compound score remained an independent, significant predictor ($B = 1.59$, $p < .001$); readability, lexical diversity, length, exclamation marks, and word length did not retain unique significance (all $p > .06$). All VIFs remained below 3.6 (range 1.07–3.57).

Table 4
Multivariate Logistic Regression Predicting Valence, Corpus 2 (Product Reviews)

Predictor	B	SE	z	P	95% CI
Constant	0.509	0.157	3.24	.001	[0.201, 0.817]
Sentiment compound (VADER)	1.591	0.255	6.25	< .001	[1.092, 2.090]
Flesch Reading Ease	−0.234	0.198	−1.18	.238	[− 0.623, 0.155]
Lexical diversity (TTR)	−0.195	0.272	−0.72	.473	[− 0.729, 0.338]
Review length (words)	−0.442	0.243	−1.82	.069	[− 0.918, 0.034]
Exclamation marks	0.178	0.165	1.08	.282	[− 0.146, 0.501]
Average word length	0.077	0.198	0.39	.697	[− 0.311, 0.465]

4.3 Cross-corpus comparison

Table 5 summarizes direction and significance of each effect across both corpora. Three of seven features — sentiment intensity, readability, and lexical diversity — replicated in both direction and statistical reliability across a different product category, sample size, and valence-elicitation method: the strongest evidence here for a category-general linguistic signature of valence. Two features (length, average word length) showed weaker or non-significant effects in the smaller corpus, consistent with reduced power. One feature — exclamation-mark frequency — reversed direction entirely, indicating a

category-dependent rather than universal relationship to valence: in film reviews, exclamation marks track indignant emphasis in negative reviews; in product reviews, they track enthusiasm in positive ones.

Table 5
Cross-Corpus Direction-of-Effect Summary

Feature	Corpus 1 (film)	Corpus 2 (product)	Converges?
Sentiment (VADER compound)	Positive > Negative***	Positive > Negative***	Yes
Flesch Reading Ease	Positive < Negative***	Positive < Negative*	Yes
Flesch–Kincaid grade	Positive > Negative***	Positive > Negative**	Yes
Review length	Positive > Negative***	Positive > Negative (ns)	Partial
Lexical diversity (TTR)	Positive < Negative***	Positive < Negative**	Yes
Exclamation marks	Positive < Negative***	Positive > Negative*	No — reversed
Average word length	Positive > Negative***	Positive ≈ Negative (ns)	No — attenuated

5. Discussion

Three findings matter for how firms and platforms treat review text as a data source. First, a core subset of linguistic features — sentiment intensity, readability, and lexical diversity — showed the same direction and statistical reliability in both a large film-review corpus and a much smaller, differently constructed product-review corpus, supporting their use as category-general signals rather than artifacts of one dataset. Second, the direction of the readability and diversity effects runs counter to an intuitive processing-fluency account in both corpora: positive reviews were consistently somewhat longer or harder to read and less lexically varied than negative reviews. Practically, this means readability-based heuristics (e.g., "clearer reviews are more positive") would misclassify valence in both domains tested here. Third, exclamation-mark frequency did not generalize: it favored negative reviews in film criticism but positive reviews in product reviews, consistent with punctuation intensifying whichever affect a category's reviews conventionally carry rather than signaling positivity or negativity in a fixed way (Biber, 1988).

The multivariate models add an operational nuance: in Corpus 1, five features carried independent predictive value beyond sentiment alone, meaning a sentiment-only classifier would leave usable signal on the table. In Corpus 2, only sentiment intensity remained independently predictive once all features were considered jointly, with the univariate effects for the other features becoming statistically redundant — most plausibly a function of Corpus 2's smaller sample and lower power rather than a

genuine difference in the underlying relationships, and an open question for future work with a larger second category.

5.1 Managerial implications

Review-triage and moderation systems. Because sentiment lexicons alone missed independently predictive signal in the film-review domain (lexical diversity, word length, length, and punctuation all retained unique predictive value net of sentiment), platforms that rely solely on off-the-shelf sentiment scores to triage, flag, or route reviews may be discarding usable structural signal — particularly in longer-form review categories. A practical recommendation is to treat sentiment score as one input among several rather than a sufficient statistic for valence.

Category-specific tooling rather than one-size-fits-all rules. The reversal of the exclamation-mark effect between film and product reviews is the clearest actionable finding here: a rule calibrated on one review category (e.g., "more exclamation marks → more positive," true for product reviews) would misclassify in another (film reviews, where the opposite held). Firms operating review systems across multiple product categories should validate linguistic-valence rules within each category before deploying them platform-wide, rather than assuming a rule that works for one vertical transfers to another.

Customer-service prioritization. Because negative reviews in the film-review corpus were reliably shorter on some structural dimensions but more lexically differentiated and more exclamation-heavy, teams building automated escalation rules for negative feedback may find punctuation and lexical-diversity signals useful supplements to sentiment score for early identification — again, subject to category-specific validation given the Corpus 2 divergence.

Caution on generalizability. The modest pseudo-R^2 values (.12, .26) indicate that most variance in valence is not explained by text-surface features alone; managers should treat these features as useful supplementary signals for triage and monitoring, not as a substitute for direct rating data where it is available.

6. Limitations

Several limitations qualify these conclusions for a managerial audience. Both corpora are restricted to English and to a binary operationalization of valence; reviews in other languages, or on other platforms with different review-writing conventions (e.g., short-form mobile reviews, star-only ratings with no text), may show different patterns, and the present findings should not be assumed to transfer without validation. Corpus 2 is small (N = 294) and imbalanced (64% positive), its document-level valence label was derived from aggregated sentence-level feature-sentiment annotations rather than a reviewer-reported overall rating, and its text was reconstructed from pre-tokenized sentences rather than original raw text, all of which make it a directional generalization check rather than a precise effect-size replication, and firms should validate these relationships on their own review data before operationalizing them. The design is correlational and cross-sectional, so no causal claim can be made

about whether linguistic features cause valence judgments or vice versa; for firms designing review-prompt interfaces or moderation nudges, this means the findings support monitoring and classification use cases more directly than intervention design. Finally, sentiment and readability scores were computed with general-purpose, non-domain-tuned tools (Flesch, 1948; Hutto & Gilbert, 2014), and firms deploying similar tooling on branded review data should expect some calibration loss relative to domain-tuned alternatives. The study reuses fully public, de-identified, previously published secondary corpora and involved no new data collection from human participants.

7. Conclusion

Review valence leaves a measurable linguistic signature beyond sentiment words, and part of that signature holds across product category. Sentiment intensity, readability, and lexical diversity distinguished positive from negative reviews in both a film-review and a product-review corpus, in some cases running counter to an intuitive processing-fluency intuition, while affective punctuation proved category-specific rather than universal. For firms and platforms building or buying review-analytics tooling, these findings argue against relying on sentiment score alone and against assuming that a linguistic-valence rule validated in one product category transfers to another. Larger-scale, multi-category replication and validation against firm-specific review data and business outcomes such as helpfulness votes or downstream sales — is the natural next step toward operational tools that use review text as a reliable input to business decisions.

References

1. Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. Review of General Psychology, 5(4), 323–370.
2. Berger, J., & Milkman, K. L. (2012). What makes online content viral? Journal of Marketing Research, 49(2), 192–205.
3. Biber, D. (1988). Variation across speech and writing. Cambridge University Press.
4. Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. Journal of Marketing Research, 43(3), 345–354.
5. Fish, S. (1980). Is there a text in this class? The authority of interpretive communities. Harvard University Press.
6. Flesch, R. (1948). A new readability yardstick. Journal of Applied Psychology, 32(3), 221–233.
7. Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 168–177). Association for Computing Machinery.
8. Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (pp. 216–225).

9. Iser, W. (1978). The act of reading: A theory of aesthetic response. Johns Hopkins University Press.
10. Mudambi, S. M., & Schuff, D. (2010). What makes a helpful online review? A study of customer reviews on Amazon.com. MIS Quarterly, 34(1), 185–200.
11. Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (pp. 271–278).
12. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1–2), 1–135.
13. Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. Advances in Experimental Social Psychology, 19, 123–205.
14. Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. Personality and Social Psychology Review, 5(4), 296–320.