applied sciences

MDPI

Article

Abstract-Level Reporting Practices and Methodological Composition in 245 Studies Using User-Generated Video Game Reviews

Tibor Guzsvinecz *, Krisztian Nemeth and Judit Szűcs

Department of Information Technology and Its Applications, University of Pannonia, Gasparich M. Utca 18/A, H-8900 Zalaegerszeg, Hungary; nemeth.krisztian@zek.uni-pannon.hu (K.N.); szucs.judit@zek.uni-pannon.hu (J.S.)
* Correspondence: guzsvinecz.tibor@zek.uni-pannon.hu

Abstract

Studies of user-generated video game reviews make sampling decisions that limit how far their findings can be generalized. How often these decisions are reported has not been measured. This abstract-level systematic mapping study screened 868 records from Scopus and Web of Science and coded the 245 eligible studies on eight dimensions from titles, abstracts and author keywords. Agreement with independent manual coding gave a median per-column Cohen's kappa of 0.92 and a screening kappa of 0.93, and a second coder agreed with the final coding at a median of 0.89. Language scope was absent from 93.1% of abstracts, corpus size from 49.8% and the review platform from 41.2%; only 4.1% reported all three. Platform disclosure improved over time; language disclosure did not. In the 56 accessible full texts of a random sample of 75, nearly all missing platforms and corpus sizes were stated, but language scope remained unstated in 36. Among studies naming a source, 72.2% used Steam. Transformer-based methods rose (odds ratio 1.56 per year, $p < 0.001$), while dictionary-based sentiment analysis did not decline in absolute use. Syntheses in this area need full-text extraction, review-based findings largely describe one storefront, and language scope is underreported in abstracts and full texts alike.

Keywords: abstract-level coding; content analysis; methodological review; reporting practices; research synthesis; Steam; systematic mapping; text mining; user-generated reviews; video games

Check for updates

Academic Editor: Douglas O'Shaughnessy

Received: 7 September 2026
Revised: 27 October 2026
Accepted: 29 October 2026
Published: 31 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Digital distribution platforms collect written reviews from the people who buy and play video games, and those reviews accumulate at a scale that no laboratory study can match. A single storefront can hold millions of reviews attached to identifiable products, each carrying a rating, a timestamp, and in many cases a record of how long the reviewer had played before writing [1]. Work on game reviews grew out of a longer line of app store analysis, in which user feedback is treated as a repository of requirements, defect reports and quality signals [2–4], and the same techniques were carried over to games once the volume of available review text made it worthwhile [1,5,6]. This article is about that body of research, not about the reviews themselves. Its unit of analysis is a published study that analyzes game reviews. Its question, which has not yet been measured, is how completely such studies describe the review data they use.

Reviews are unsolicited, they are written after use rather than during a scheduled session, and they are already labeled by the reviewer as favorable or unfavorable. Studies in marketing and information systems established early that such text carries measurable consequences for sales and for how other consumers behave [7,8], and that the properties of an individual review predict how useful other readers find it [9–11]. The same relationships have been observed for video games, where review sentiment and volume shape purchase recommendations [12] and where reading reviews before play measurably alters the experience that follows [13,14].

Every study that analyzes this material makes sampling decisions: which storefront the reviews come from, how many are collected, which languages are kept, and whether the text was written by ordinary users or by professional critics. Those decisions determine what the results can be generalized to, and they determine whether a later study can compare its results against an earlier one. The decisions matter more than they might appear. Language restriction is the clearest case. Filtering a review corpus to English is defensible for many purposes, but it is a choice about which players are represented, and the consequences of concentrating computational work on a small number of well-resourced languages have been documented at length in natural language processing [15]. Platform choice carries similar weight, since storefronts differ in their review interfaces, their voting mechanisms, their moderation and their user populations [1]. Corpus size determines what a statistical result can support, and without it a reported effect cannot be weighed against a competing one. Similar concerns have been raised in other fields (Section 2.4). For video game review research, a more basic question has not been answered: what do these studies actually report about the data they analyze?

The answer matters to three groups of readers. Researchers who synthesize this literature need to know whether two findings rest on comparable data before combining them, and they can only know what primary studies disclose. Developers, platform operators and analysts who act on review-based findings need to know which players and which storefront a finding describes before applying it to their own product. Authors of new studies need a baseline against which their own reporting can be judged. The abstract is where this is best measured, because it is what database searches index, what screening decisions are made on, and the only part of a study that every reader can reach regardless of institutional subscriptions.

This study answers that question at the level of the abstract, by coding the literature rather than summarizing it. We retrieved records from Scopus (Elsevier B.V., Amsterdam, The Netherlands; https://www.scopus.com) and the Web of Science Core Collection (Clarivate, London, UK; https://www.webofscience.com), screened 868 of them against explicit eligibility criteria, and coded the 245 eligible studies on a fixed set of dimensions covering the platform, the corpus size, the language scope, the analytical methods, the constructs examined, and the type of contribution. Coding was performed on titles, abstracts and author keywords, the reliability of every coded dimension was assessed against independent manual coding, and a random sample of full texts was audited to measure how much of what abstracts omit is stated elsewhere in the paper.

Two aspects of the design deserve mention at the outset. First, the coding was assisted by a large language model. Automated annotation of this kind can reproduce human judgment well on some tasks and poorly on others, and performance cannot be assumed without measurement [16,17], so every dimension was validated on a stratified sample and the coefficients are reported per dimension rather than pooled. Second, the screening decision was validated separately, on a sample drawn from the excluded records. A sample of included records alone cannot detect the most serious error in a synthesis: a record that was wrongly left out.

The study addresses the following four research questions (RQs):

RQ1. *To what extent do studies that analyze user-generated video game reviews report the platform, the corpus size, and the language scope of their data?*

RQ2. *How has the composition of analytical methods in this literature changed over time, and in particular, has the adoption of transformer-based methods coincided with a decline in dictionary-based sentiment analysis?*

RQ3. *Among studies that identify the source of their review data, how concentrated is that source across platforms?*

RQ4. *Which constructs are examined with which families of analytical methods?*

The contribution is threefold. The substantive contribution is a measured account of reporting practice and methodological composition in a literature that is frequently drawn on but has not been described in these terms. The methodological contribution is the validation design: a two-sided reliability assessment for a screening decision, with the resulting error rates projected onto the corpus, which gives an interval for the number of eligible studies rather than a single count. The practical contribution is a short set of reporting items, each tied to the share of abstracts that currently omit it (Section 5.1), which authors and journals can adopt directly.

The remainder of the paper is organized as follows. Section 2 reviews the literature that this corpus is drawn from, along three lines: what reviews are taken to provide evidence about, which analytical procedures are applied to them, and which constructs are examined. Section 3 describes the corpus, the coding scheme, the validation design and the full-text audit. Section 4 reports the results. Section 5 discusses them, and Section 6 concludes.

2. Related Work

2.1. Reviews as Evidence About Play

The premise behind this literature is that written reviews contain information about the experience of playing that other instruments do not capture. Kang et al. argued that review text reflects aspects of player experience that gameplay logs cannot record [18], and Wang and Goh identified a set of recurring dimensions in review content, including narrative, achievement, social interaction and general experience [19]. Hedegaard and Simonsen showed that usability and user experience information can be extracted from review text at a level of detail comparable to what dedicated instruments produce [20]. Zhu and Fang applied a lexical approach to a large collection of reviews and recovered psychologically grounded dimensions such as playability, usability and creativity [21]. Busurkina et al. used review text specifically as an instrument for game experience evaluation [22].

The construct these studies address has a longer history. Player experience is generally taken to refer to the subjective and affective responses that arise during play [23,24], and the recognition that games evoke a wide emotional range predates the availability of large review corpora. What changed is the instrument. Early large-scale measurement of online game populations relied on behavioral traces rather than text [25], while the review corpus offers self-reported accounts at comparable scale.

The developer-facing value of this material was established first in software engineering. App store reviews were shown to contain actionable feature requests, defect reports and fine-grained sentiment about individual features [4,26], and the specific complaints that users raise were catalogued systematically [3]. Lin et al. then asked whether game reviews share the characteristics of app reviews, and found enough similarity to justify carrying the tooling across [1]. Adjacent work on the same platform examined update behavior and early access dynamics [5,6].

2.2. Analytical Procedures Applied to Review Text

Dictionary-based sentiment analysis assigns polarity or discrete emotion categories to words using a fixed lexicon. VADER and the NRC Emotion Lexicon are the two most common instruments [27,28], with related resources developed for aspect-level and tweet-level work [29,30]. The general approach is well surveyed [31–33], and it has been applied directly to game review corpora at scale [34]. Its appeal is transparency and cost; its limitations with sarcasm, domain-specific slang and context sensitivity are acknowledged throughout, and the causes of misclassification have been examined specifically for game reviews [35].

Probabilistic topic modeling describes what a large collection of documents is about without requiring the documents to be read [36]. Applied to reviews, it has been combined with classification steps to score the playability of a single title along functionality, gameplay and usability [37]. Embedding-based approaches represent documents in a vector space before clustering them, which allows topics to be recovered from semantic similarity rather than word co-occurrence [38,39].

Transformer classifiers apply pretrained architectures to labeling tasks. RoBERTa and models built on it [40] have been packaged into ready-to-use classifiers [41] that make it practical to score tens of millions of reviews for properties that would once have required manual coding [42]. Classical supervised learning remains in use alongside these [43], as does aspect-based sentiment analysis, which attributes sentiment to specific game features rather than to the review as a whole [44,45].

Work outside the video games domain has repeatedly shown that the choice among these procedures is consequential. Annotation quality varies with language and annotator [46], emotion recognition transfers unevenly across languages [47], and feature-level sentiment behaves differently from document-level sentiment [48]. Similar procedures have been applied to other kinds of user-generated developer-facing text [49].

2.3. What Is Examined Through Reviews

First, player experience and genre. Differences in experience across genres have been examined both through surveys and through review text [50], as have the roles of difficulty [51], narrative design [52], moral choice mechanics [53] and deliberately uncomfortable experiences [54]. Reviews have also been used to study genre-specific reception directly [34].

Second, design feedback. A parallel line treats reviews as a source of concrete design lessons, covering control feel and responsiveness [55], difficulty pacing [56], level design metrics [57] and input configuration [58,59].

Third, purchase and market behavior. Reviews have been used to study marketing of novel products [60] and purchase intention [61]. The structure of digital distribution has been described independently of review data [62].

Fourth, social and multiplayer experience. Multiplayer contexts introduce performance pressure, matchmaking and team dynamics as sources of affect [63–65], and experience management in multiplayer settings has been treated as a design problem in its own right [66]. Emotionally expressive characters occupy an intermediate position between these and single-player narrative work [67].

Fifth, health and discomfort. Immersive settings introduce physical symptoms as a reported experience, and review corpora have been used to examine both the syntactic properties of virtual reality reviews [68] and the ways players describe and mitigate discomfort [69].

Sixth, hostile language. Toxicity in player communities has become a research object in its own right. Prevalence and psychological consequences have been documented through

surveys and community studies [70–72], alongside work on normalization [73], coping strategies [74,75], griefing [76], anonymity and disinhibition [77], and moderation and extremism [78,79]. Review text is one of the corpora on which this is now measured [42].

Seventh, authenticity and manipulation. A separate line treats the review corpus itself as contested ground. Coordinated rating campaigns can produce large surges of reviews that respond to something other than the product, and analysis of one prominent case shows that such campaigns are better described as ideological polarization than as straightforward misinformation, with a grassroots counter-response following the initial surge [80]. Reviews have also been used as an input to recommendation instead of as an object of study. There the text is tested alongside behavioral signals such as playtime in predicting what a player will want next, and the sentiment carried by the text contributed less than the behavioral signal alone [81].

Lastly, reception of specific design features. Reviews have been used to study how players respond to particular properties of a game rather than to the game as a whole, such as how a title was localized into several languages [82] and which monetization mechanics players experience as problematic [83].

2.4. Reporting and the Limits of Secondary Use

The breadth described above is an asset for the field and a problem for anyone attempting to combine its findings. Results drawn from different platforms, different corpus sizes and different language subsets are not automatically comparable, and whether they can be compared at all depends on what each study discloses.

This concern is not specific to video games. Systematic reviews and surveys of app store research have documented the difficulty of synthesizing primary studies whose data descriptions are underspecified [84–86], and the classification schemes such syntheses rely on assume a level of detail that primary studies do not always provide [87]. Comparable concerns about concentration on a small number of languages have been raised across natural language processing [15]. Reporting guidelines address the conduct of syntheses [88] rather than the primary studies they draw on, and the level of detail those guidelines presuppose is exactly what the present study sets out to measure.

3. Materials and Methods

3.1. Design and Scope

This is a systematic mapping study [89,90]. Each eligible record was coded on a fixed set of substantive dimensions, and the reported results are distributions and trends over those codes. The design follows the classification procedure that systematic mapping guidelines describe, in which reviewers read abstracts to build a classification scheme and sort the studies into it [89]. Two elements are added to that procedure: a reliability protocol for every coded dimension and a full-text audit. Standard bibliometric indicators such as source concentration, author productivity and citation or collaboration networks were not computed. They describe the structure of a field rather than the reporting behavior of the studies within it, and the research questions here concern the latter.

Systematic mapping differs from a systematic literature review in purpose. A review establishes what the evidence says about a question and appraises the quality of the primary studies. A map classifies a body of research to show what has been studied and how, and does not appraise quality [89]. The present study asks the second kind of question, about the reporting and the methods of a literature rather than its findings, so no quality appraisal of the included studies was performed.

The mapping process has five steps [89], and each corresponds to a part of this study:

- Definition of the research questions (Section 1);

- Search for primary studies, here in two databases with a saturation test and precision checks (Section 3.2);
- Screening of the records against explicit criteria (Section 3.3);
- Development of a classification scheme from the abstracts, here through a calibration batch and six recorded revisions (Sections 3.3 and 3.4);
- Classification of the included studies, reported as frequencies and trends (Section 4).

Consistent with the updated guidelines [90], every coded dimension was checked against independent manual coding (Section 3.5).

Coding was performed on titles, abstracts and author keywords, and full texts were not used for coding. The decision is deliberate, for three reasons. First, it holds the coding surface constant across every record. Second, it matches the surface on which database searches and first-pass screening actually operate. Third, it is the one surface available to every reader regardless of access rights, since a large share of the scholarly literature remains behind subscription access [91]; in the full-text audit of this study, 44.0% of the sampled studies were not openly available at the publisher (Section 4.8). The consequence is that every proportion reported here describes disclosure in the abstract, not practice in the full study. Section 3.6 describes an audit that estimates the distance between the two, and Section 5.2 discusses what remains.

3.2. Search Strategy and Corpus Assembly

Two multidisciplinary databases were searched on 1 September 2026: Scopus (Elsevier B.V., Amsterdam, The Netherlands; https://www.scopus.com) and the Web of Science Core Collection (Clarivate, London, UK; https://www.webofscience.com). Both queries combined two blocks with AND. The first block captured the object of study, using the phrases video game, digital game, computer game, online game, mobile game, game platform, gaming platform and gaming, in singular and plural forms. The second block captured the data type, using user review, player review, game review, user-generated review, online review, customer review, consumer review, user-generated content, player feedback, user feedback and review text, again in both forms. The token Steam was never used alone, since the unbound form retrieves a large volume of steam engineering literature. It appeared only in the bound phrases Steam platform, Steam reviews and Steam store. Two diagnostic queries checked the remaining sources of noise. The bound Steam phrases and steam engineering vocabulary had no overlap in the retrieved set. The token gaming overlapped with gambling terminology in 2 records.

In Scopus, the query used the TITLE-ABS-KEY field operator and returned 801 records. Restricting to English, to articles, conference papers and reviews, and to the subject areas Computer Science, Social Sciences, Psychology, Business, Engineering, Arts and Humanities, and Decision Sciences left 630 records. Query saturation was tested by extending the second block in three directions: app store terminology, rating-as-data terminology, and community text terminology. The three extensions added 1, 2 and 0 records, so the query was treated as saturated. In Web of Science, the same blocks were expressed with the TS field tag, which additionally covers Keywords Plus and is therefore slightly broader than TITLE-ABS-KEY. The query returned 581 records, 570 after the language and document type filters, and 428 after research area filters chosen to mirror the Scopus subject areas.

IEEE Xplore (IEEE, Piscataway, NJ, USA; https://ieeexplore.ieee.org) was evaluated as a third source and was not included. The same conceptual query, expressed with the All Metadata scope, returned 616 records, of which 493 carried DOIs absent from the Scopus set. A random sample of 50 of those, screened against the eligibility criteria, yields a precision that projects roughly 30 additional eligible records (Appendix A). The exclusion does not rest on that yield. A single-publisher database would shift the distribution of venues and

analytical methods towards engineering conference literature, and those distributions are the object of study here. Web of Science was retained on the same reasoning applied in the other direction. It is multidisciplinary and broadly symmetrical with Scopus, so it does not introduce a directional bias in the quantities being measured. A parallel precision check on records unique to Web of Science is reported in Appendix A. Appendix A also reports what the inclusion of IEEE Xplore would most plausibly have changed (Table A1).

Figure 1 shows the full sequence of retrieval, filtering, deduplication and screening for both databases.

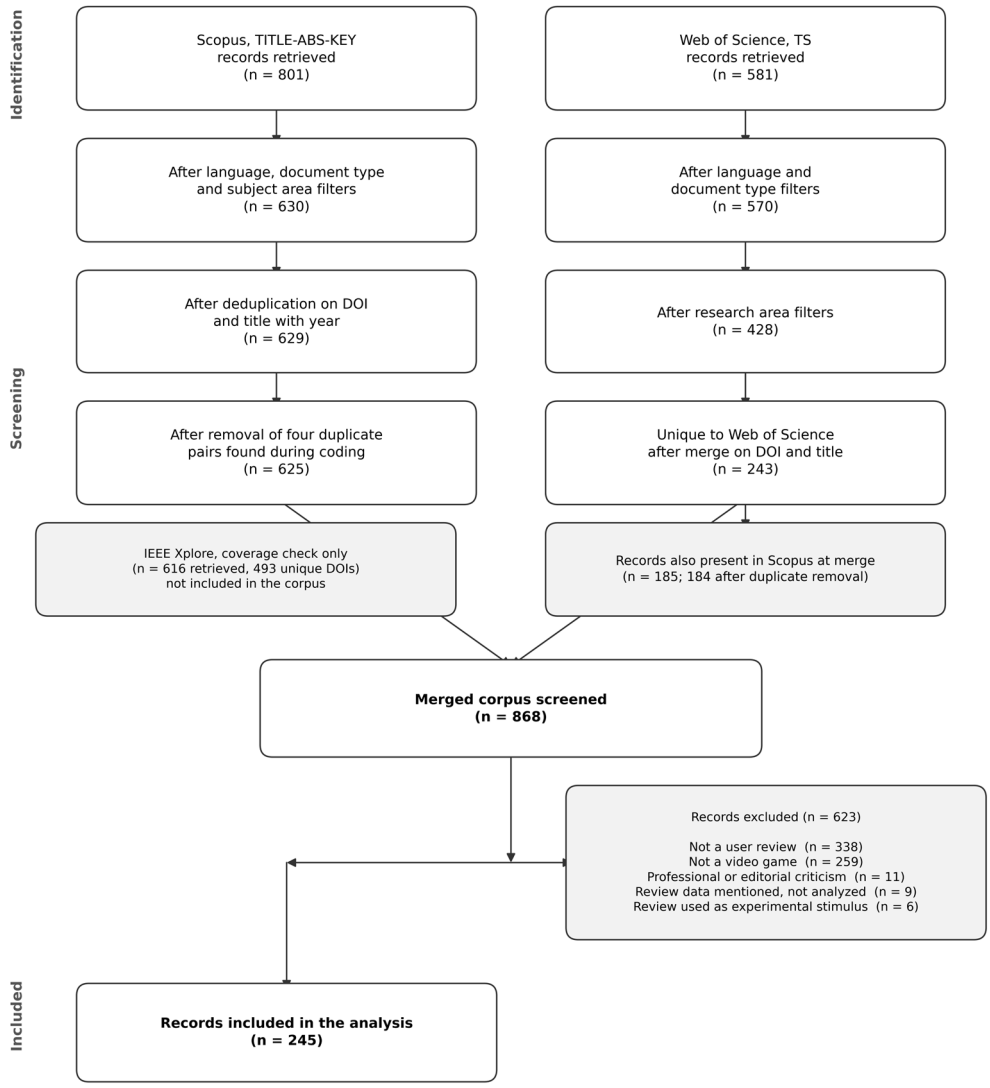


Figure 1. Flow of records through retrieval, filtering, deduplication and screening. Counts for the Scopus and Web of Science branches are shown separately up to the merge. IEEE Xplore was evaluated as a coverage check and is shown outside the main flow, since none of its records entered the corpus.

Within the Scopus set, records were deduplicated on a normalized DOI, falling back to a normalized title and publication year where no DOI was present, which reduced 630 records to 629. Four further duplicate pairs were identified during coding, none of which either key had caught. In two of them, a conference and a journal version of the same work were both present and the journal version was retained. In the third, neither version was a journal article and the one carrying a DOI was retained. In the fourth, the conference version was retained. The Scopus block therefore holds 625 records. The Web of Science set was merged against that block on the same two keys, since conference records in Web

of Science frequently lack DOIs. Of the 428 subject-filtered records, 185 matched a Scopus record at the point of merging and 243 did not. One of those 185 was subsequently removed as a duplicate, so 184 of the retained Scopus records are also present in Web of Science. The 243 unique records were appended as a separate block so that the existing record order and coding sequence remained unchanged. The final corpus holds 868 records: 441 from Scopus only, 184 from both databases, and 243 from Web of Science only.

3.3. Eligibility Criteria and Coding Scheme

A record was eligible if three conditions held. It had to concern video games, excluding gambling, sports betting and gamification outside a game context. It had to use naturally occurring user reviews or review text as analyzed data rather than mention them. It had to make an empirical or methodological contribution. Exclusion reasons were recorded from a closed list: not a video game, not a user review, professional or editorial criticism, review used as experimental stimulus, review data mentioned but not analyzed, and other.

The object of the review had to be a video game. Records analyzing reviews of game hardware, physical toys, venues, gamified non-game products, or mixed e-commerce corpora were excluded even when the study was set in a gaming context.

Two asymmetries in the criteria are deliberate. Uncertainty about the object of the review leads to exclusion, since out-of-scope records would contaminate the corpus. Uncertainty about the authorship of the reviewed text leads to inclusion with a flag, since excluding on an unstated property would introduce a bias that cannot be inspected afterwards.

Appendix B gives worked examples. Table A2 shows, for each exclusion reason and for each boundary rule above, a record from the corpus, what its abstract states and the resulting decision. Table A3 shows the full set of codes received by three eligible records that differ in how much they disclose.

Eligible records were coded on eight dimensions. Platform records the named source of the review data and is multi-valued. Corpus size records the reported figure together with its unit, distinguishing reviews, games, users and other units, and its reporting status. Language scope records the language of the analyzed reviews. Method family is multi-valued across eleven families (dictionary-based sentiment analysis, probabilistic topic models, embedding-based representation, transformer-based models, classical supervised learning, other deep neural architectures, network analysis, psychometric and lexical factor analysis, inferential statistical modeling, qualitative coding, and descriptive statistics) together with an explicit flag for records naming no analytical procedure at all. Construct is multi-valued across twelve constructs, with a residual category and an unclear value alongside them. Study type distinguishes primary, secondary, methodological and position contributions. Database provenance records the source database. This value is taken from the export records rather than judged from the abstract, so no agreement coefficient is reported for it in Appendix C. A derived measure counts how many of the three basic sampling attributes, namely platform, corpus size and language, are missing from the abstract, and therefore ranges from zero to three. A separate flag records whether the authorship of the analyzed text is identifiable.

A further dimension, validation practice (D6), was defined initially and removed after calibration; the column labels in Appendix C retain that gap. Of the 18 eligible records in the calibration batch, 15 could not be resolved, since methodological validation is almost never described in an abstract.

Two conventions govern the coding. A value of not reported means the abstract does not state the attribute, and is a substantive observation. A value of unclear means the coding scheme cannot accommodate what the abstract states, and is an instrument failure.

A threshold of 5% unclear per dimension was set in advance, above which the scheme was to be revised rather than individual records re-judged.

3.4. Coding Procedure

Coding was assisted by a large language model (Claude Opus 5; Anthropic; San Francisco, CA, USA) working from the coding scheme, in batches of 25 records, with per-batch logging of the eligibility rate and publication year distribution. Because the performance of such annotation varies with the task [16,17], every dimension was validated against human labels (Section 3.5).

A calibration batch of 50 randomly selected records was coded first, reviewed manually, and used to revise the coding scheme before the remaining records were processed. Batches after calibration followed publication year order, so per-batch eligibility rates are not comparable across batches, and the random calibration batch provides the only unbiased estimate of the eligibility rate obtained during coding.

The scheme passed through six revisions, each triggered either by calibration or by a conflict detected during coding, and each recorded with the records affected. Where a revision changed the definition of a column, that column was recoded across all eligible records rather than applied only to later batches. Two columns were treated this way: the flag for records naming no analytical procedure, and the flag distinguishing objects of review that are not video games. Ten of the eligible records are publications by the authors of this study. They entered the corpus through the same queries as every other record and were screened and coded under the same scheme, with no separate handling at any stage.

3.5. Reliability and Analysis

Reliability was assessed against independent manual coding by the first author, who did not see the machine coding before completing each sample.

Two samples were drawn, because one sample cannot validate both parts of the decision. The first covered the coded dimensions: 148 eligible records, stratified by publication year and study type. The sample was drawn at 150 records from the eligible set as it stood at the time; two records were later reclassified as ineligible under a codebook revision and were removed from the sample rather than replaced. The stratum proportions therefore depart from those of the final eligible set by a fraction of a percentage point, which does not affect the coefficients. The second covered the screening decision: 122 excluded records, stratified by exclusion reason, with the three rare strata sampled in full rather than proportionally, since proportional allocation would have left one or two records in each.

A second author then independently coded 60 records drawn from the first sample, under the final version of the coding scheme and without access to either earlier coding. The sample was enriched with every record that the final coding marks as naming no analytical procedure, as using descriptive statistics only or as addressing a residual construct (28 records), and completed at random. Agreement was computed both with the final dataset, which reassesses the coding after the rule revisions, and with the first manual coding, which gives an agreement between two human coders. Records that the second coder excluded were adjudicated under the same rules as the earlier disagreements.

Cohen's kappa [92] was computed separately for every codable column, using the asymptotic variance estimator of Fleiss, Cohen and Everitt [93]. The coded dimensions can only be compared on records that both coders judged eligible, so the per-dimension coefficients rest on the 144 records of that sample which both retained, and on 143 for the unit of the corpus size figure, which one record could not be compared on. Confidence intervals for kappa follow the normal approximation and are truncated to the interval from minus one to one, since that approximation can otherwise place a bound outside the

range of the coefficient. Confidence intervals for proportions are Wilson intervals [94]. Conventional thresholds for interpreting kappa magnitudes follow [95]. Two properties of the coefficient constrain what can be reported for some columns. Where a column takes a single value across both coders, kappa is undefined and raw agreement is reported in its place. Where the two codings agree completely, the variance estimate is not positive and no interval can be formed, so the point estimate is reported without one. No coefficient is pooled across dimensions, since the dimensions differ in difficulty and a single figure would obscure exactly the variation that matters here. Within the three multi-valued dimensions, the unweighted mean of the component columns is reported alongside the individual coefficients as a summary of that dimension, and it is not used for any inference.

The screening decision required an additional step. Within each sample, the machine decision is the same for every record, so kappa is zero regardless of how well the coders agree and says nothing about the screening decision. The two samples were therefore combined into a single set of 270 records in which both outcomes occur. The prevalence of that combined set is artificial, so the coefficient is reported alongside the two directional error rates, which do not depend on prevalence. Because the excluded sample was stratified, with the three rare strata sampled in full, the false-exclusion rate was projected onto the corpus stratum by stratum, weighting each stratum by its size. The eligible sample was allocated proportionally and is self-weighting. Intervals for the projected counts, and for the estimated number of truly eligible records, were obtained jointly by simulation: each stratum's error rate was drawn from its Jeffreys distribution and the errors among unsampled records from the corresponding binomial distribution (200,000 draws).

Disagreements were resolved by general rules rather than case by case, so that another coder could apply each rule independently. The reliability coefficients were computed before these resolutions and are therefore conservative. They were not recomputed afterwards, since a coefficient computed after resolution would no longer measure independent agreement.

All results are reported on the eligible set in two versions: the full corpus, and the Scopus core, defined as records present in Scopus regardless of Web of Science presence. The second serves as a sensitivity check on the database decision described in Section 3.2. Proportions are reported with their denominators, since several quantities are conditional. Platform distribution in particular is computed among records that name a source. Corpus size was summarized only on records whose reported figure refers to reviews, as figures given in other units are not comparable.

Four trend tests were specified before analysis, one per column, using logistic regression of the coded value on publication year: dictionary-based methods, transformer-based methods, probabilistic topic models, and the flag for records naming no analytical procedure. No further columns were tested and no correction for multiple comparisons was applied, since the set was fixed in advance. A sensitivity analysis repeated all four tests with the 2027 records removed, as those sit at the edge of the year range where leverage is highest. The same model was fitted to the three reporting attributes, with non-reporting of the platform, the corpus size and the language scope as outcomes. These three tests were not specified in advance, are reported as exploratory, and are kept separate from the four prespecified tests (Appendix D).

No chi-square test was computed for the cross-tabulation of method family against construct, because the table is sparse. Associations there are reported as conditional proportions with intervals.

The search was run on 1 September 2026, so the 2026 publication year is incomplete, and two records carry a 2027 year because they appeared online ahead of issue. Both years are retained in all tables and figures and are marked as partial. Figures showing annual

proportions include only years with at least five eligible records, since proportions computed on one to four records would suggest trends the data do not support; the underlying tables contain every year.

Corpus assembly and deduplication were carried out in R 4.3.2 (R Foundation for Statistical Computing, Vienna, Austria) using dplyr 1.1.4, stringr 1.5.1, readr 2.1.5 and tidyr 1.3.0. Agreement coefficients were computed in R using base functions only, with no inter-rater agreement package; the coefficients for the second coder were computed in Python 3.12.4 with the same estimator and checked against the R implementation. Descriptive statistics, confidence intervals for proportions, the logistic regression models and all figures were produced in Python 3.12.4 (Python Software Foundation, Wilmington, DE, USA) with pandas 2.2.2, numpy 1.26.4, scipy 1.12.0, statsmodels 0.14.5 and matplotlib 3.10.1.

3.6. Full-Text Audit

To estimate how much of what abstracts omit is stated in the full text, a simple random sample of eligible records was drawn (seed 42). The audit was carried out on 75 records (30.6% of the eligible set). For each record, the first author read the full text and recorded the platform, the corpus size with its unit and the language scope, together with the section in which each was stated. Records were processed in the order in which they were drawn, so that any initial segment of the sample is itself a simple random sample. For every record, the audit also recorded whether the full text was accessible to the authors and by which route: open access at the publisher, publisher access through institutional subscription, a public repository, or a copy posted by the authors. Three of the sampled records are publications by the authors of this study. They were retained, since replacing them would have made the sample non-random. The audit measures the distance between abstract and full text. It was not used to recode any record, and all other results remain abstract-level. To check the consistency of extraction, a second author independently extracted the same three attributes from the accessible full texts of the first 30 records, without access to the first extraction.

4. Results

4.1. Corpus Composition

Of the 868 screened records, 245 met the eligibility criteria and 623 did not. No record remained unresolved. Among the exclusions, 338 records did not concern user reviews, 259 did not concern video games, 11 analyzed professional or editorial criticism, 9 mentioned review data without analyzing it, and 6 used reviews as an experimental stimulus. The eligible records span 2006 to 2027. Of these, 141 were retrieved from Scopus only, 86 from both databases, and 18 from Web of Science only. Primary studies account for 165 records and methodological contributions for 80. No record was coded as a secondary study or as a position paper. Annual output is concentrated in recent years: 201 of the 245 eligible records appeared from 2019 onward. This is shown in detail in Figure 2.

4.2. Coding Reliability

Across the 37 codable columns, the median Cohen's kappa was 0.92. Ten of them showed complete agreement between the two coders, so the variance estimate is not positive and the coefficient is reported without an interval. Two further columns were constant across both coders and are therefore not among the 37: for those, kappa is not defined and raw agreement is reported in its place. Three columns fell below 0.70: the descriptive statistics flag (0.432), the residual construct category (0.475), and the flag for records naming no analytical procedure (0.655). All three are addressed in Section 5.2. Table A4 in Appendix C reports the coefficient, its confidence interval and the raw agreement for every column.

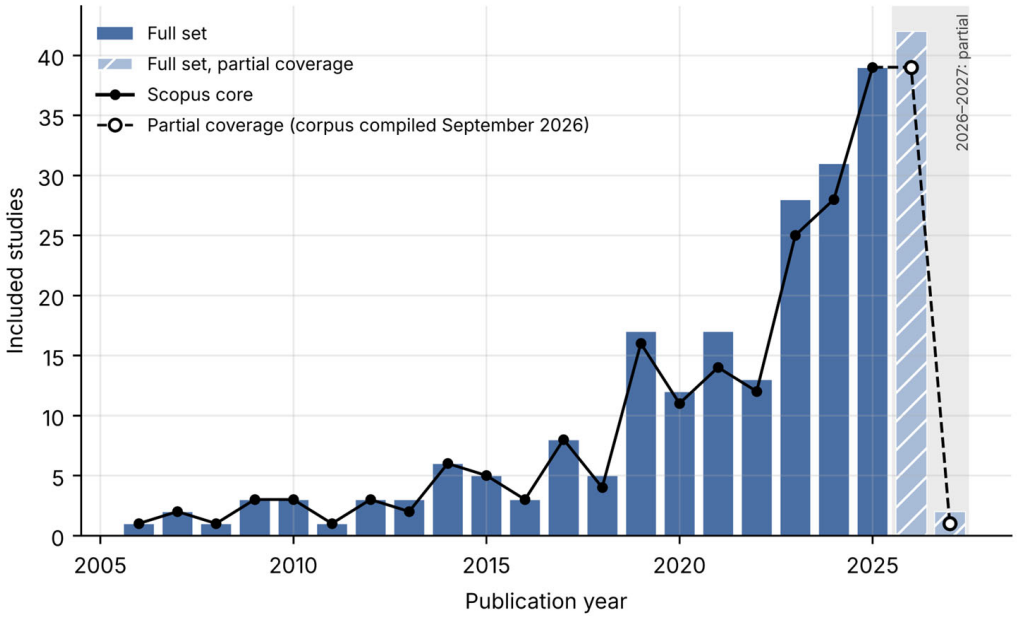


Figure 2. Eligible records per publication year, full corpus and Scopus core. The 2026 and 2027 columns are marked as partial, since the search was run on 1 September 2026 and two records carry a 2027 publication year from advance online publication.

For the screening decision, on the combined 270-record sample (Section 3.5), raw agreement was 0.967 and Cohen's kappa was 0.933 (95% CI [0.889, 0.976]). The two directional error rates are 5 of 122 in the excluded sample, or 4.1% (95% CI [1.8, 9.2]), and 4 of 148 in the eligible sample, or 2.7% (95% CI [1.1, 6.7]). Three of the five disagreements in the excluded sample fell in strata that were sampled in full (Table A5), so the unweighted rate overstates the error in the large strata. Weighted by stratum size, the projection corresponds to approximately 15 records (7 to 43) excluded that manual coding would have retained, and 7 records (4 to 12) retained that manual coding would have excluded, giving an estimated 254 truly eligible records (244 to 281). Agreement on the reason for exclusion, computed on the 117 records both coders excluded, was 0.987 (95% CI [0.962, 1.000]).

A second coder, working under the final scheme, agreed with the final dataset with a median per-column kappa of 0.885 across 36 informative columns, and with the first manual coder with a median of 0.816 (Table A6). The flag for records naming no analytical procedure, whose first coefficient predates the sharpening of its definition, reached 0.775 (95% CI [0.589, 0.962]) against the recoded column, and the residual construct category reached 0.936. The attributes behind the main results agreed closely: platform not reported 0.961, language scope 1.000, and the transformer-based, dictionary-based and topic model flags 1.000, 1.000 and 0.923. Agreement was lowest for the construct dimension (median 0.791), where six columns fell below 0.70. Of the 60 records, the second coder would have excluded five that both earlier codings had retained; adjudicated under the existing rules, two were confirmed as ineligible because their review data are mentioned but not analyzed, and three were retained. The two records remain in the analyzed set, consistent with the treatment of screening error in this section; removing them changes no reported proportion by more than 0.2 percentage points and no significance decision.

4.3. Reporting of Sampling Attributes

All proportions in this section describe abstracts. Language scope is absent from the abstracts of 228 of the 245 eligible records, or 93.1%. Corpus size is absent from 122 records and the platform from 101, or 49.8% and 41.2% (Table 1). Only 10 records (4.1%) report all three, and the mean number of missing attributes is 1.84. The authorship of the analyzed text cannot be determined from the abstract in 19 records, or 7.8% (95% CI [5.0, 11.8]). This flag is not part of the three-attribute measure.

Table 1. Reporting of the three sampling attributes in the abstracts of the 245 eligible records, with counts, proportions and 95% Wilson confidence intervals, and the distribution of the number of attributes missing per record.

Attribute	Not Reported	%	95% CI
Language scope	228	93.1	89.2–95.6
Corpus size	122	49.8	43.6–56.0
Platform	101	41.2	35.2–47.5
Attributes Missing Per Record		**Records**	**%**
0		10	4.1
1		66	26.9
2		122	49.8
3		47	19.2
Mean		1.84	–

Across years with at least five eligible records, the mean number of missing attributes ranges from 1.58 to 2.40 with no monotonic movement. The three components behave differently over time (Figure 3). In exploratory logistic regressions on publication year, the odds of an abstract not naming the platform fell by about 14% per year (odds ratio 0.856, 95% CI [0.801, 0.914], $p < 0.001$), while neither corpus size (0.962, [0.909, 1.018], $p = 0.178$) nor language scope (1.000, [0.896, 1.116], $p = 0.997$) showed a trend. The same pattern holds in the Scopus core and without the 2027 records (Table A7 in Appendix D).

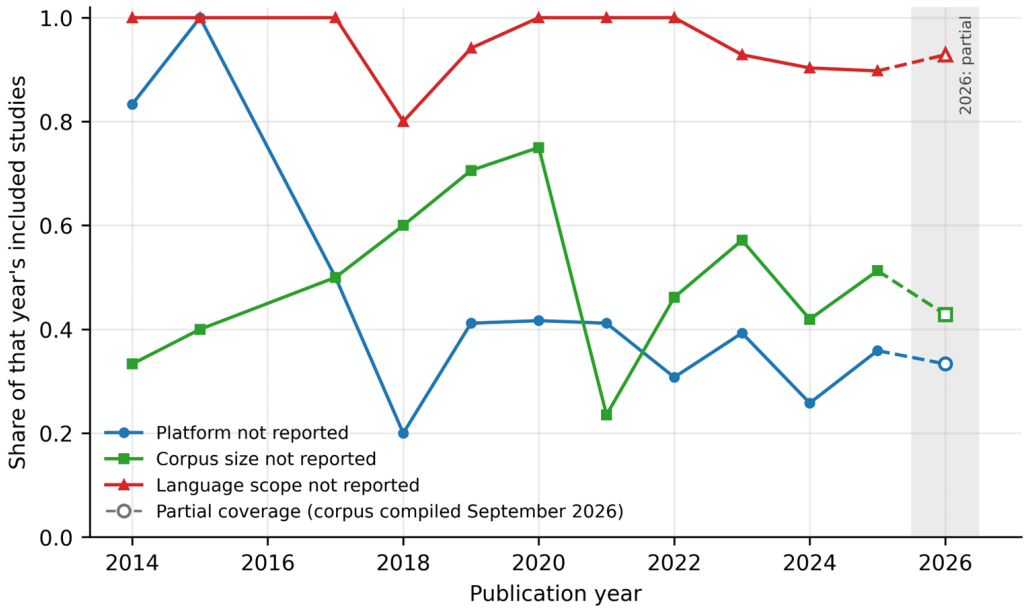


Figure 3. Proportion of eligible records whose abstract does not report platform, corpus size or language scope, by publication year, for years with at least five eligible records. Partial years are marked.

4.4. Method Families and Change over Time

Inferential statistical modeling appears in 59 eligible records (24.1%) and qualitative coding in 57 (23.3%). Probabilistic topic models appear in 37 (15.1%) and transformer-based models in 34 (13.9%). A further 33 records (13.5%) name no analytical procedure at all. No other family reaches 30 records. The dimension is multi-valued, so the proportions sum to more than one. Table 2 gives every family with its interval.

Table 2. Frequency of each method family among the 245 eligible records, with counts, proportions and 95% Wilson confidence intervals. The dimension is multi-valued. The descriptive statistics row is marked as low reliability and is not interpreted as a substantive category. The flag for records naming no analytical procedure also falls below 0.70 but is not marked, since its coefficient predates the definition change described in Section 5.2. * denotes low reliability (kappa = 0.432).

Method Family	*n*	%	95% CI
Inferential statistical modeling	59	24.1	19.2–29.8
Qualitative coding	57	23.3	18.4–28.9
Probabilistic topic models	37	15.1	11.2–20.1
Transformer-based models	34	13.9	10.1–18.8
No analytical procedure named	33	13.5	9.8–18.3
Classical supervised learning	29	11.8	8.4–16.5
Dictionary-based sentiment analysis	19	7.8	5.0–11.8
Embedding-based representation	17	6.9	4.4–10.8
Other deep neural architectures	14	5.7	3.4–9.4
Psychometric or lexical factor analysis	11	4.5	2.5–7.9
Network analysis	7	2.9	1.4–5.8
Descriptive statistics *	3	1.2	0.4–3.5
Method unclear	0	0.0	0.0–1.5

Of the four prespecified trend tests, transformer-based models and probabilistic topic models rise with publication year (odds ratios 1.556 and 1.358 per year, both $p < 0.001$), while dictionary-based sentiment analysis and the flag for records naming no analytical procedure show no significant trend ($p = 0.237$ and $p = 0.417$). Table 3 reports all four with their intervals, both with and without the two 2027 records. No significance decision changes, and no odds ratio moves by more than 0.02. Figure 4 traces the two series that carry the comparison. Table A8 in Appendix D gives the underlying counts by period, both as a share of all eligible records and as a share of records that use a computational text method. No eligible record uses both families.

Table 3. Logistic regression of four prespecified method indicators on publication year, for the full corpus ($n = 245$) and the Scopus core ($n = 227$), and repeated with the two 2027 records removed ($n = 243$ and $n = 226$). The year is centered at zero on the minimum of each set. No correction for multiple comparisons was applied, since the set of tests was fixed in advance.

Set	Indicator	Positive	OR Per Year	95% CI	*p*	OR Without 2027	*p*
Full	Dictionary-based sentiment	19	1.082	0.950–1.232	0.237	1.066	0.330
Full	Transformer-based models	34	1.556	1.246–1.942	<0.001	1.548	<0.001
Full	Probabilistic topic models	37	1.358	1.149–1.605	<0.001	1.350	<0.001
Full	No analytical procedure named	33	1.038	0.949–1.135	0.417	1.041	0.383
Scopus core	Dictionary-based sentiment	18	1.098	0.956–1.260	0.186	1.078	0.273
Scopus core	Transformer-based models	31	1.498	1.201–1.869	<0.001	1.518	<0.001
Scopus core	Probabilistic topic models	32	1.307	1.107–1.543	0.002	1.288	0.002
Scopus core	No analytical procedure named	30	1.035	0.944–1.134	0.466	1.036	0.447

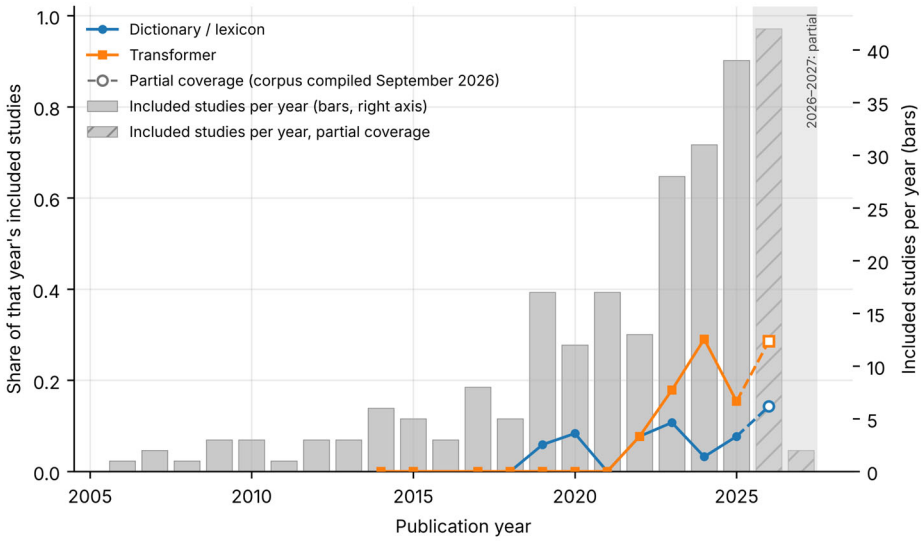


Figure 4. Annual proportion of eligible records using dictionary-based sentiment analysis and transformer-based models. The lines cover years with at least five eligible records and refer to the left axis. The bars behind them give the annual record count for every year and refer to the right axis; hatched bars mark the partial years 2026 and 2027. Partial years are also marked by the shaded band, open markers and dashed line segments.

4.5. Constructs and Their Association with Method Families

Topical content is the most frequent construct, appearing in 123 eligible records (50.2%), followed by sentiment in 118 (48.2%) and player experience in 95 (38.8%). Design feedback appears in 59 records (24.1%) and purchase behavior in 43 (17.6%). The remaining constructs each appear in fewer than 30 records, with toxicity the rarest substantive construct. Table 4 gives every construct with its interval.

Table 4. Frequency of each construct among the 245 eligible records, with counts, proportions and 95% Wilson confidence intervals. The dimension is multi-valued. The residual construct row is marked as low reliability and is not interpreted as a substantive category. * denotes low reliability (kappa = 0.475).

Construct	*n*	%	95% CI
Topical content	123	50.2	44.0–56.4
Sentiment	118	48.2	42.0–54.4
Player experience	95	38.8	32.9–45.0
Design feedback	59	24.1	19.2–29.8
Purchase behavior	43	17.6	13.3–22.8
Discrete emotion	29	11.8	8.4–16.5
Reviewer behavior	23	9.4	6.3–13.7
Other construct *	20	8.2	5.4–12.3
Review helpfulness	15	6.1	3.8–9.9
Recommendation	14	5.7	3.4–9.4
Health and safety	10	4.1	2.2–7.4
Fake review detection	7	2.9	1.4–5.8
Toxicity	5	2.0	0.9–4.7
Construct unclear	2	0.8	0.2–2.9

The cross-tabulation of method family against construct is sparse: 121 of its 168 cells contain fewer than five records (Table A9 in Appendix E). Among method families with at least ten records, the strongest conditional proportions are topic models with topical

content (33 of 37, 89.2%, lift 1.78 over the overall rate) and transformer-based models with sentiment (28 of 34, 82.4%, lift 1.71). Table 5 lists every pair above 0.60 within its method family.

Table 5. Conditional proportions for the method and construct pairs with the highest within-method share, restricted to method families with at least ten eligible records. The denominator is the size of the method family, not the eligible set. Lift compares the conditional proportion with the overall rate of that construct. No test of independence was computed.

Method Family	Construct	*n*/*n* Method	%	95% CI	Overall Rate	Lift
Probabilistic topic models	Topical content	33/37	89.2	75.3–95.7	50.2	1.78
Transformer-based models	Sentiment	28/34	82.4	66.5–91.7	48.2	1.71
Psychometric/lexical factor analysis	Topical content	9/11	81.8	52.3–94.9	50.2	1.63
Classical supervised learning	Sentiment	23/29	79.3	61.6–90.2	48.2	1.65
Qualitative coding	Topical content	44/57	77.2	64.8–86.2	50.2	1.54
Dictionary-based sentiment	Sentiment	14/19	73.7	51.2–88.2	48.2	1.53
Inferential statistical modeling	Sentiment	40/59	67.8	55.1–78.3	48.2	1.41
Embedding-based representation	Topical content	11/17	64.7	41.3–82.7	50.2	1.29
Other deep neural architectures	Sentiment	9/14	64.3	38.8–83.7	48.2	1.34
Psychometric/lexical factor analysis	Player experience	7/11	63.6	35.4–84.8	38.8	1.64
No analytical procedure named	Topical content	21/33	63.6	46.6–77.8	50.2	1.27
Probabilistic topic models	Player experience	23/37	62.2	46.1–75.9	38.8	1.60

4.6. Platforms and Corpus Size

Among the 144 eligible records that name a source, 104 use Steam, or 72.2%. No other platform reaches 20 records. No eligible record draws on a console storefront. Corpus size can be compared only on the 74 records whose reported figure refers to reviews. Across those, the median is 45,142 reviews, and the spread is wide. Table 6 gives the platform distribution, the corpus size statistics and the distribution of reporting units.

Table 6. Platform distribution among the 144 eligible records that name a source, with 95% Wilson confidence intervals, and descriptive statistics for corpus size among the 74 records reporting a figure in reviews. The platform dimension is multi-valued.

Platform	*n*	% of 144	95% CI
Steam	104	72.2	64.4–78.9
Other	18	12.5	8.1–18.9
Google Play	15	10.4	6.4–16.5
Metacritic	8	5.6	2.8–10.6
Apple App Store	5	3.5	1.5–7.9
Console storefront	0	0.0	0.0–2.6
Corpus Size in Reviews (n = 74)		**Value**	
Minimum		85	
First quartile		3232	
Median		45,142	
Third quartile		401,720	
Maximum		35,983,481	
Mean		888,245	
Unit of the Reported Figure (n = 245)		**Records**	**%**
Reviews		74	30.2
Games		44	18.0
Other unit		5	2.0
Not reported		122	49.8

4.7. Sensitivity to Database Selection

Restricting the analysis to the 227 records present in Scopus leaves every main quantity close to its value on the full corpus. Language scope is absent from 92.5% of records rather than 93.1%, corpus size from 51.1% rather than 49.8%, and platform from 41.0% rather than 41.2%. Among records naming a source, 75.4% use Steam rather than 72.2%. All four trend tests retain the same direction and the same significance decisions, and the exploratory platform trend is of similar size (odds ratio 0.849, 95% CI [0.793, 0.909]).

4.8. Results of the Full-Text Audit

Of the 75 audited records, 56 (74.7%, 95% CI [63.8, 83.1]) were accessible to the authors, but only 42 (56.0%, [44.7, 66.7]) were openly available at the publisher. Of the 33 behind a publisher paywall, 13 could be read only because the authors had posted a copy (11) or deposited one in a repository (2), one through an institutional subscription, and 19 (25.3% of the sample) could not be obtained at all. Inaccessible full texts were concentrated in IEEE publications: 7 of the 14 IEEE records in the sample could not be obtained, against 12 of the 61 others.

On the 56 accessible records, the full text supplied almost everything the abstract left out for platform and corpus size, but not for language scope (Table 7). The platform was missing from 26 of the abstracts and was stated in 25 of the corresponding full texts. Corpus size was missing from 26 abstracts and was stated in 24 of the full texts. Language scope was missing from 54 of the 56 abstracts and was stated in 18 of those full texts, so 36 of the 56 full texts (64.3%, 95% CI [51.2, 75.5]) do not state the language of the reviews they analyze. No audited record stated an attribute in its abstract that its full text omitted. In the audited sample, the abstract-level shares of missing platform (46.4%), corpus size (46.4%) and language scope (96.4%) are close to those of the full eligible set (41.2%, 49.8% and 93.1%). The two extractions of the first 30 records agreed on whether the platform was stated in 27 of 27 cases, on corpus size in 26 of 27 and on language scope in 25 of 27 (Cohen's kappa 0.834, 95% CI [0.617, 1.000]); where both found a value, the corpus sizes and languages were identical in every case and the platforms in 23 of 26.

Table 7. Full-text audit of the 56 accessible records among the 75 drawn at random from the eligible set: sampling attributes missing from the abstract, how many of those the full text states, and how many remain missing in the full text, with 95% Wilson confidence intervals.

Attribute	Missing in Abstract	Stated in Full Text	Missing in Full Text (%)	95% CI
Platform	26	25 of 26	1 (1.8%)	0.3–9.5
Corpus size	26	24 of 26	2 (3.6%)	1.0–12.1
Language scope	54	18 of 54	36 (64.3%)	51.2–75.5

5. Discussion

RQ1 asked how far studies report the platform, the corpus size and the language scope of the review data they analyze. The three attributes behave differently, and the difference is the substantive result. Language scope is almost never stated in abstracts. It is missing from 93.1% of the eligible records, while corpus size and platform are missing about half the time (Table 1). Corpus size and platform are the kind of details that find their way into an abstract because they convey scale and provenance, both of which make a study sound substantial. Language scope conveys neither. It is a restriction rather than an asset, and restrictions are what abstracts leave out. As a result, the attribute that most directly determines who is represented in the data is the one a reader is least likely to find in the abstract. This matters because language restriction is not a neutral technical choice. The

concentration of computational resources on a small number of well-resourced languages is well documented [15], and a review corpus filtered to English is a corpus of the players who write in English. Whether that filter was applied at all, let alone why, cannot be determined from the abstract in the overwhelming majority of cases here. Nor is this gap narrowing. Naming the platform has become more common over the period covered, while stating the language scope shows no trend (odds ratio 1.000, 95% CI [0.896, 1.116]; Section 4.3). The attribute that describes provenance has improved; the one that describes representation has not. The full-text audit qualifies this. Platform and corpus size that abstracts omit are almost always stated further in, so for those two attributes the gap is largely one of placement. Language scope was not stated in 36 of the 56 accessible full texts (64.3%; Table 7), so for language the gap extends beyond the abstract and into the reporting of the studies themselves. A related gap concerns who wrote the analyzed text. In 7.8% of records the abstract does not make clear whether the material is user-generated or professional criticism. That share is small, but it bears on a distinction the literature otherwise treats as fundamental.

RQ2 asked how the composition of analytical methods has changed, and specifically whether transformer-based methods have displaced dictionary-based sentiment analysis. The answer depends on what displacement is measured against. Transformer use rises steeply with publication year, as does probabilistic topic modeling. Dictionary-based methods did not decline in absolute use: they appear in 2 records before 2019, in 3 between 2019 and 2022, and in 13 between 2023 and 2026, which is 4.5%, 5.1% and 9.3% of the eligible records in those periods (Table A8). A non-significant trend is not evidence of no change, but the interval for the dictionary trend (odds ratio 1.082 per year, 95% CI [0.950, 1.232]) excludes a decline in the odds of use steeper than about 5% per year. Relative to other computational text methods, the picture is less clear. Among records that use any such method, the dictionary share is lower in recent years (2 of 6 before 2019, 15.3% from 2023), and the estimated trend points downward (odds ratio 0.908 per year, 95% CI [0.778, 1.060]), but the interval is compatible both with no change and with a substantial relative decline. No eligible record uses both families, so the two are applied in different studies rather than combined or compared within one. The supported reading is therefore that transformer-based methods were added without reducing the use of dictionary-based ones, while their effect on the relative position of dictionary methods cannot be determined from this corpus. Lexicon-based instruments remain in use because they are transparent, cheap and interpretable [27,28], and surveys of the area anticipated coexistence rather than succession [31–33]. Inferential statistical modeling and qualitative coding are the two most frequent families, ahead of every computational text method. This is not a purely computational literature, and treating it as one would misdescribe it. The other feature is less comfortable. In 33 records (13.5%), the abstract names no analytical procedure at all, and that share shows no trend (Table 3). Methodological reporting in abstracts has not improved alongside methodological sophistication.

RQ3 asked how concentrated the data sources are among studies that name one. Among records that identify a platform, 72.2% use Steam. Not one eligible record draws on a console storefront. Across the full eligible set, the Steam share is 104 of 245 (42.4%), so the dominance holds among studies that disclose their source rather than across the literature as a whole. Whether it supplies three-quarters of all review research is not something these data can say. Even with that qualification, the concentration has consequences. Steam's review system has specific affordances: a binary recommendation rather than a numeric rating, a public playtime figure attached to each review, and a helpfulness voting mechanism. Findings that depend on any of these are findings about Steam. The playtime field in particular has no equivalent on most other storefronts, and it underpins a substantial body

of work relating review timing to review content [1,34]. The absence of console storefronts is a coverage gap rather than a finding about console players, since console review systems are largely inaccessible to automated collection.

RQ4 asked which constructs are examined with which method families. Topical content, sentiment and player experience account for most of the coded constructs. The method pairings that stand out (Table 5) reach lifts between roughly 1.5 and 1.8, which indicates a tendency and not a rule. What is more notable is the scarcity at the edges. Toxicity is an active research area with a substantial literature of its own [70–79], yet it appears in five of the eligible records. The same holds for fake and coordinated reviews [80], coded as fake review detection, which appear in seven. Review corpora are evidently not the primary instrument through which these questions are being pursued, even though they are one of the few sources that capture hostile language at scale in a publicly archived form.

5.1. Implications

The findings have four practical consequences, which differ in who can act on them.

For researchers synthesizing this literature, abstract-level screening cannot establish whether two studies rest on comparable data. The full-text audit shows that platform and corpus size can usually be recovered from the full text, so a systematic review or meta-analysis of game review research should plan and budget for full-text extraction of those attributes instead of expecting to filter on them at the screening stage. Automated or language-model-assisted screening faces the same limit, since it works on the layer measured here. Language scope is different: in the audited sample it could not be recovered from 36 of the 56 accessible full texts, so a synthesis that needs it will often have to treat it as unknown or contact the authors.

For authors and journals, most of the gap is cheap to close. Table 8 lists five reporting items, each with the share of eligible abstracts that currently omit it. For platform and corpus size the information is usually already in the Section 3, and one sentence in the abstract would move it to the layer where studies are found, screened and compared. For language scope, the information is usually absent altogether, and it needs to be stated in the methods first: which languages were retained, and how language was identified. Journals in this area could ask for these items as readily as they ask for a data availability statement.

Table 8. Reporting items for studies that analyze user-generated game reviews, with the share of the 245 eligible abstracts that did not state each item.

Item	What to State	Not Stated in the Abstract
Platform	The storefront or site from which the reviews were collected	101 (41.2%)
Corpus size	The number analyzed, with the unit counted (reviews, games, users)	122 (49.8%)
Language scope	The languages retained and how language was identified	228 (93.1%)
Authorship	Whether the text was written by users or by professional critics	19 (7.8%)
Analytical procedure	The family of method applied to the text	33 (13.5%)

For readers applying review-based findings, the platform concentration sets the scope of what is known. Among studies that name a source, 72.2% use Steam, so conclusions about how players write, rate and vote are largely conclusions about one storefront's review design, with its binary recommendation, visible playtime and helpfulness voting. A developer or analyst who applies such findings to a console or mobile title is extrapolating, and should weigh the evidence accordingly.

For accounts of methodological change, a rising method is not evidence of a declining one. Transformer-based models rose, while dictionary-based sentiment analysis did not

decline in absolute use, so a study that presents the older method as abandoned is describing a change that this corpus does not show. Whether it has lost ground relative to newer methods is a separate question, which these data cannot settle.

5.2. Limitations

First, the coding surface is the main limitation. Titles, abstracts and author keywords were used throughout. Classifying studies from their abstracts is the standard procedure of systematic mapping [89], but here the classified attributes are themselves reporting attributes, so every reported proportion describes disclosure in the abstract, and that is narrower than practice. A study that filtered to English and said so only in its Section 3 is counted here as not reporting language scope, which is the correct coding under the scheme and a genuine limit on interpretation. The full-text audit (Section 4.8) shows that the abstract-level proportions are upper bounds on non-reporting in the full text. For language scope the bound is close to the full-text level; for platform and corpus size it is far above it. They are also direct measures of what a reader or a screening tool can see before opening the paper. Nineteen sampled full texts could not be obtained. One of them states its language scope in the abstract, so its full text does as well; across all 75 sampled records, language scope is therefore unstated in between 36 (48.0%, 95% CI [37.1, 59.1]) and 54 (72.0%, [61.0, 80.9]) full texts, depending on the 18 that could not be read.

Second, one dimension was lost to this constraint entirely. Validation practice, meaning whether automated classifications were checked against human judgment and whether an agreement statistic was reported, could not be resolved from abstracts and was removed from the scheme (Section 3.3). Given how much of this literature now rests on automated classification, this is the most valuable question that a full-text follow-up could address.

Third, agreement was not uniform across columns. Two residual categories, the descriptive statistics flag and the residual construct category, fell below conventional thresholds in the first validation and are reported without interpretation. The flag for records naming no analytical procedure also fell below 0.70 in the first validation, but that coefficient predates the sharpening of its definition; measured by a second coder against the recoded column, it reached 0.775 (Section 4.2). The second coding also showed that the construct dimension is the least reliable part of the scheme, with six construct columns below 0.70, so the associations between methods and constructs (RQ4) should be read as tendencies rather than precise rates. The language scope column reached 0.720 (95% CI [0.419, 1.000]) in the first validation and complete agreement with the final dataset in the second, but both rest on few positive cases: only 17 of the eligible records state a language scope at all. The headline proportion is therefore measured on a column whose agreement is well established in the negative direction and more thinly in the positive one.

Fourth, the screening decision carries a quantified error. As reported in Section 4.2, the stratum-weighted projection gives an interval of 244 to 281 truly eligible records against the 245 analyzed. The interval treats the manual coding as the standard, and it rests on the disagreements observed before they were resolved; all five disputed exclusions in the excluded sample (Table A5) were confirmed as exclusions when the rules were applied. Reporting it is preferable to presenting 245 as exact, but it should not be read as a correction to the counts.

Fifth, coverage is bounded by two databases. IEEE Xplore was evaluated and excluded on the grounds that a single-publisher source would bias the venue and method distributions that this study measures; the measured precision of its unique records was 0.06. Simulating its inclusion from the profile of IEEE-published records already in the corpus shifts the method shares in the expected direction but leaves the conclusions of the prespecified trend tests unchanged at the projected yield (Appendix A, Table A1). The Scopus core analysis,

covering 227 of the 245 records, reproduces every main quantity within 3.2 percentage points, which indicates that the Web of Science addition did not drive the results.

Finally, the search was run on 1 September 2026. The 2026 publication year is incomplete and two records carry a 2027 year from advance online publication. Both are marked as partial throughout, and the trend tests were repeated without the 2027 records with no change to any significance decision.

6. Conclusions

Studies that analyze user-generated video game reviews rarely state, at the level where they are found and screened, what data they analyzed. We coded 245 eligible studies from 868 screened records on an abstract-level scheme, validated every coded dimension and the screening decision against independent manual coding, and audited a random sample of full texts.

The four research questions have direct answers.

- RQ1: Language scope is missing from 93.1% of abstracts, corpus size from 49.8% and the platform from 41.2%, and only 4.1% report all three. Language disclosure has not improved over the period covered, and in a random audit language scope remained unstated in 36 of 56 accessible full texts, whereas the full texts almost always state the platform and the corpus size.
- RQ2: Transformer-based methods were added without a decline in the use of dictionary-based ones, although whether dictionary-based methods lost ground relative to other computational methods could not be determined, and inferential statistics and qualitative coding remain the most frequent method families, so this is not a purely computational literature.
- RQ3: Among studies that name a source, 72.2% use Steam and none uses a console storefront, so much of what the field knows about player reviews is knowledge about one storefront.
- RQ4: Methods and constructs pair by tendency rather than by rule, with topic models most often applied to topical content and transformer-based models to sentiment, while toxicity and fake review detection are rarely studied through reviews.

The practical consequences follow directly. Syntheses in this area need full-text extraction of sampling attributes, and should expect language scope to be unrecoverable in many studies. Authors and journals can close most of the gap with the items in Table 8, and for language scope the first step is to state it in the methods at all. Findings based on Steam reviews should be read as findings about Steam until other storefronts are studied at comparable depth.

The validation design is reusable beyond this corpus: per-dimension agreement for language-model-assisted coding, a screening check drawn from the excluded records, and stratum-weighted error rates projected onto the corpus as an interval of 244 to 281 truly eligible studies. The same scheme can be applied to app store, e-commerce and travel review research to test whether these reporting gaps are specific to game studies. Whether automated classifications in this literature are validated against human judgment remains open, since abstracts do not carry that information.

Author Contributions: Conceptualization, T.G.; methodology, T.G.; software, T.G.; validation, T.G. and J.S.; formal analysis, T.G., K.N. and J.S.; investigation, T.G. and J.S.; resources, T.G.; data curation, T.G.; writing (original draft preparation), T.G., K.N. and J.S.; writing (review and editing), T.G., K.N. and J.S.; visualization, T.G.; supervision, T.G.; project administration, T.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The coded dataset, the validation samples, the second coding, the full-text audit with its second extraction, and the analysis scripts are openly available on Zenodo at https://doi.org/10.5281/zenodo.23042331.

Acknowledgments: During the preparation of this study, the authors used Claude Opus 5 (Anthropic) to apply the coding scheme to the titles, abstracts and author keywords of the screened records, as described in Section 3.4. All coded dimensions and the screening decision were validated against independent manual coding by the first author, a subsample was coded independently by a second author, and the agreement statistics are reported in Section 4.2 and Appendix C. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A

The IEEE Xplore query returned 616 records. Of these, 93 shared a DOI with the Scopus set, 493 carried DOIs absent from it, and 30 carried no DOI. A random sample of 50 of the 493 was screened against the eligibility criteria and yielded 3 eligible records, a precision of 0.06 (95% CI [0.016, 0.175]), projecting to roughly 30 additional eligible records (8 to 86). None of the 30 records without DOIs matched a Scopus title. A parallel check was run for transparency: a random sample of 50 records from the unfiltered Web of Science set yielded 5 eligible records; 38 of the 50 survived the subject filter, and all 5 eligible ones were among them. The subject filter removed 12 sampled records and no eligible ones.

To assess what the excluded records could have changed, the 54 eligible records already in the corpus that were published by IEEE (DOI prefix 10.1109) were used as a profile of IEEE Xplore content. They differ from the rest of the corpus in the direction the exclusion anticipated: 29.6% use transformer-based models and 20.4% classical supervised learning, against 13.9% and 11.8% in the full eligible set, while 7.4% use qualitative coding, against 23.3%. Records drawn with replacement from this profile were added to the corpus in the numbers given by the precision check (8, 30 and 86), and every analysis was repeated 1000 times (Table A1). At the projected 30 records, no significance decision of the four prespecified trend tests changed in any replicate, the platform and language trends were unchanged, and only the exploratory corpus size trend changed its decision, in 2.9% of the replicates (40.8% at 86 records). The shares of methods moved towards engineering practice (transformer-based models from 13.9% to 15.6%, qualitative coding from 23.3% to 21.5%). Because the added records are resampled copies, the intervals of the augmented analyses are somewhat narrower than real additional records would give.

Table A1. Sensitivity of the main estimates to adding records with the profile of IEEE-published eligible studies (1000 replicates per scenario). Values are medians with 2.5th and 97.5th percentiles; the last column gives the share of replicates in which the significance decision at 0.05 differs from the original analysis.

Estimate	Original	+8 Records	+30 Records	+86 Records	Decision Changed (+8/+30/+86)
OR per year, dictionary-based	1.082	1.079 [1.033, 1.092]	1.061 [0.998, 1.094]	1.011 [0.952, 1.081]	0.0%/0.0%/0.1%
OR per year, transformer-based	1.556	1.567 [1.497, 1.614]	1.582 [1.482, 1.688]	1.625 [1.497, 1.800]	0.0%/0.0%/0.0%

Table A1. *Cont.*

Estimate	Original	+8 Records	+30 Records	+86 Records	Decision Changed (+8/+30/+86)
OR per year, topic models	1.358	1.357 [1.322, 1.379]	1.345 [1.291, 1.394]	1.324 [1.259, 1.401]	0.0%/0.0%/0.0%
OR per year, no procedure named	1.038	1.037 [1.030, 1.044]	1.034 [1.022, 1.048]	1.028 [1.009, 1.047]	0.0%/0.0%/0.0%
OR per year, platform not reported	0.856	0.855 [0.849, 0.862]	0.853 [0.842, 0.864]	0.850 [0.832, 0.869]	0.0%/0.0%/0.0%
OR per year, corpus size not reported	0.962	0.961 [0.953, 0.967]	0.958 [0.946, 0.970]	0.951 [0.933, 0.968]	0.0%/2.9%/40.8%
OR per year, language scope not reported	1.000	1.002 [0.979, 1.031]	1.005 [0.968, 1.054]	1.022 [0.957, 1.080]	0.0%/0.0%/2.6%
Share using transformer-based models (%)	13.9	14.2 [13.4, 15.4]	15.6 [13.8, 17.5]	17.8 [15.7, 20.5]	–
Share using qualitative coding (%)	23.3	22.5 [22.5, 23.3]	21.5 [20.7, 22.5]	19.0 [17.8, 20.5]	–
Steam share among records naming a source (%)	72.2	72.5 [70.9, 73.5]	73.1 [70.5, 75.3]	73.9 [70.6, 77.3]	–

Appendix B

Table A2. Worked examples of screening decisions, one per exclusion reason and per boundary rule of Section 3.3. The index refers to the record identifier in the openly available coded dataset (see Data Availability Statement).

Index	Year	Record (Title, Shortened)	What the Abstract States	Decision	Rule Applied
219	2017	Pokémon Go: Impact on Yelp restaurant reviews	Yelp reviews of restaurants, analyzed for the effect of the game	Not a video game	The reviewed object must be a video game; here it is a restaurant
454	2024	Evaluation of virtual reality experiential dimensions using sentiment analysis	User-generated content about virtual reality gaming centres in seven cities	Not a video game	The reviewed object is a venue
179	2015	User reviews of gamepad controllers	User reviews of the gamepads of four consoles	Not a video game	Game hardware is excluded
285	2020	Cross-domain helpfulness prediction of online consumer reviews	Amazon reviews from five product categories, one of them video games, with no game-specific result	Not a video game	Mixed e-commerce corpora are excluded
78	2009	Characterizing and understanding game reviews	Reviews from popular websites, which the abstract itself frames as videogame journalism	Professional or editorial criticism	Authorship is stated in the abstract
254	2019	Digital empathic games and their relation with mortality	Player opinions collected from discussion forums, although the abstract also calls them user reviews	Not a user review	Forum text is not review text
7	2015	Are warnings from online users effective?	Participants rate researcher-selected game reviews in an experiment	Review used as experimental stimulus	Reviews must be naturally occurring data

Table A2. *Cont.*

Index	Year	Record (Title, Shortened)	What the Abstract States	Decision	Rule Applied
321	2021	Improving quality of service industries using online customer reviews	Analysis of published studies across six service sectors, one of them online games	Review data mentioned, not analyzed	Reviews serve only as background
361	2022	To pay or not to pay	4092 comments from app review sites and social media about in-app purchases	Eligible	Reviews are one of the analyzed sources; the unit is recorded as other
636	2023	Impact of seller- and buyer-created content on product sales	Amazon review content and sales; the product category appears only in the author keywords	Eligible	The coding surface includes author keywords

Table A3. Complete coding of three eligible records that differ in how much they disclose, as assigned from their titles, abstracts and author keywords. In the first record the abstract also names sentiment analysis and two classification steps without naming a method family, so only the topic model is coded.

Dimension	Li et al. [37]	Hsu and O'Hagan [82]	Petrovskaya et al. [83]
Platform	Steam	Steam	Not reported
Corpus size	99,993 reviews	1 game	801 reviews
Language scope	Not reported	Multilingual	Not reported
Attributes missing (0–3)	1	0	2
Method family	Probabilistic topic models	Qualitative coding	Qualitative coding
Construct	Sentiment; topical content; player experience; design feedback	Topical content; player experience; design feedback	Topical content; purchase behavior; player experience
Study type	Methodological	Primary	Primary

Appendix C

Table A4. Inter-coder agreement for every codable column, on the 144 records of the reliability sample that both coders judged eligible. Kappa is not defined (n.d.) where a column takes a single value across both coders, and no interval can be formed where the two codings agree completely. Columns below 0.70 are marked with an asterisk.

Dimension	Column	*n*	Raw Agr.	Kappa	95% CI
Authorship flag	Authorship not identifiable	144	0.958	0.801	0.648–0.954
D1 Platform	Steam	144	0.972	0.942	0.886–0.998
D1 Platform	Apple App Store	144	1.000	1.000	–
D1 Platform	Google Play	144	0.993	0.930	0.793–1.000
D1 Platform	Metacritic	144	1.000	1.000	–
D1 Platform	Console storefront	144	1.000	n.d.	–
D1 Platform	Other platform	144	0.993	0.944	0.834–1.000
D1 Platform	Platform not reported	144	0.972	0.944	0.889–0.998
D1 Platform	mean over 6 columns	144	0.988	0.960	–
D2 Corpus size	Corpus size unit	143	0.951	0.920	0.862–0.977
D2 Corpus size	Corpus size reporting status	144	0.951	0.903	0.833–0.973
D3 Language scope	Language scope	144	0.979	0.720	0.419–1.000
D4 Method family	Dictionary-based sentiment	144	0.993	0.949	0.849–1.000
D4 Method family	Probabilistic topic models	144	1.000	1.000	–
D4 Method family	Embedding-based representation	144	1.000	1.000	–
D4 Method family	Transformer-based models	144	1.000	1.000	–
D4 Method family	Classical supervised learning	144	0.993	0.968	0.904–1.000

Table A4. *Cont.*

Dimension	Column	*n*	Raw Agr.	Kappa	95% CI
D4 Method family	Other deep neural architectures	144	1.000	1.000	–
D4 Method family	Network analysis	144	1.000	1.000	–
D4 Method family	Psychometric/lexical factor analysis	144	0.993	0.919	0.763–1.000
D4 Method family	Inferential statistical modeling	144	0.938	0.822	0.710–0.933
D4 Method family	Qualitative coding	144	0.951	0.869	0.775–0.963
D4 Method family	Descriptive statistics *	144	0.965	0.432	0.030–0.835
D4 Method family	No analytical procedure named *	144	0.917	0.655	0.478–0.832
D4 Method family	Method unclear	144	1.000	n.d.	–
D4 Method family	mean over 12 columns	144	0.979	0.884	–
D5 Construct	Sentiment	144	0.903	0.801	0.703–0.899
D5 Construct	Discrete emotion	144	0.972	0.893	0.790–0.996
D5 Construct	Toxicity	144	0.993	0.797	0.408–1.000
D5 Construct	Review helpfulness	144	0.993	0.944	0.834–1.000
D5 Construct	Topical content	144	0.868	0.737	0.631–0.843
D5 Construct	Purchase behavior	144	0.965	0.863	0.747–0.980
D5 Construct	Player experience	144	0.889	0.765	0.656–0.873
D5 Construct	Health and safety	144	0.993	0.930	0.793–1.000
D5 Construct	Design feedback	144	0.896	0.740	0.619–0.862
D5 Construct	Fake review detection	144	1.000	1.000	–
D5 Construct	Recommendation	144	1.000	1.000	–
D5 Construct	Reviewer behavior	144	0.938	0.708	0.528–0.887
D5 Construct	Other construct *	144	0.910	0.475	0.237–0.713
D5 Construct	Construct unclear	144	1.000	1.000	–
D5 Construct	mean over 14 columns	144	0.951	0.832	–
D7 Study type	Study type	144	0.917	0.818	0.720–0.916

Table A5. Screening validation in the excluded sample, by stratum. Disagreements are records that the machine coding excluded and the manual coding retained. Strata marked as census were sampled in full, so their count is the population value.

Stratum (Exclusion Reason)	Records in Corpus	Sampled	Disagreements	Projected
Not a user review	338	54	1	6.3
Not a video game	259	42	1	6.2
Professional or editorial criticism	11	11 (census)	2	2
Review data mentioned, not analyzed	9	9 (census)	1	1
Review used as experimental stimulus	6	6 (census)	0	0
Total	623	122	5	15.4 (7 to 43)

Table A6. Agreement of the second coder with the final dataset and with the first manual coder, on the 55 records of the second sample that the second coder judged eligible. The sample was enriched for the no-procedure flag, the descriptive statistics flag and the residual construct category, so prevalences differ from the eligible set. Kappa is not defined (n.d., raw agreement in parentheses) where a column takes a single value across both codings, and is not informative where one coding is constant. A dash in parentheses marks complete agreement, for which no interval can be formed.

Dimension	Column	*n*	Raw Agr. Versus Final	Kappa Versus Final [95% CI]	Kappa Versus First Coder [95% CI]
Authorship flag	Authorship not identifiable	55	0.927	0.737 [0.497–0.976]	0.756 [0.530–0.982]
D1 Platform	Steam	55	1.000	1.000 (–)	1.000 (–)
D1 Platform	Apple App Store	55	1.000	1.000 (–)	1.000 (–)
D1 Platform	Google Play	55	1.000	1.000 (–)	0.791 [0.394–1.000]
D1 Platform	Metacritic	55	0.982	0.658 [0.035–1.000]	0.658 [0.035–1.000]
D1 Platform	Console storefront	55	1.000	n.d. (1.000)	n.d. (1.000)
D1 Platform	Other platform	55	1.000	1.000 (–)	0.848 [0.555–1.000]
D1 Platform	Platform not reported	55	0.982	0.961 [0.886–1.000]	0.961 [0.886–1.000]

Table A6. *Cont.*

Dimension	Column	*n*	Raw Agr. Versus Final	Kappa Versus Final [95% CI]	Kappa Versus First Coder [95% CI]
D2 Corpus size	Corpus size unit	29	1.000	1.000 (–)	0.704 [0.489–0.920]
D2 Corpus size	Corpus size reporting status	55	0.873	0.738 [0.557–0.918]	0.816 [0.664–0.967]
D3 Language scope	Language scope	55	1.000	1.000 (–)	0.660 [0.042–1.000]
D4 Method family	Dictionary-based sentiment	55	1.000	1.000 (–)	1.000 (–)
D4 Method family	Probabilistic topic models	55	0.982	0.923 [0.773–1.000]	1.000 (–)
D4 Method family	Embedding-based representation	55	0.964	0.732 [0.380–1.000]	0.732 [0.380–1.000]
D4 Method family	Transformer-based models	55	1.000	1.000 (–)	1.000 (–)
D4 Method family	Classical supervised learning	55	1.000	1.000 (–)	0.879 [0.646–1.000]
D4 Method family	Other deep neural architectures	55	1.000	1.000 (–)	1.000 (–)
D4 Method family	Network analysis	55	0.982	0.848 [0.555–1.000]	0.848 [0.555–1.000]
D4 Method family	Psychometric/lexical factor analysis	55	1.000	1.000 (–)	1.000 (–)
D4 Method family	Inferential statistical modeling	55	0.909	0.652 [0.371–0.933]	0.562 [0.252–0.873]
D4 Method family	Qualitative coding	55	0.927	0.758 [0.536–0.980]	0.810 [0.605–1.000]
D4 Method family	Descriptive statistics	55	0.982	0.658 [0.035–1.000]	0.791 [0.394–1.000]
D4 Method family	No analytical procedure named	55	0.909	0.775 [0.589–0.962]	0.860 [0.708–1.000]
D4 Method family	Method unclear	55	0.964	not informative	not informative
D5 Construct	Sentiment	55	0.982	0.962 [0.889–1.000]	0.724 [0.538–0.911]
D5 Construct	Discrete emotion	55	0.946	0.809 [0.601–1.000]	0.824 [0.631–1.000]
D5 Construct	Toxicity	55	1.000	1.000 (–)	1.000 (–)
D5 Construct	Review helpfulness	55	0.982	0.658 [0.035–1.000]	1.000 (–)
D5 Construct	Topical content	55	0.836	0.680 [0.498–0.861]	0.707 [0.519–0.894]
D5 Construct	Purchase behavior	55	0.909	0.614 [0.307–0.922]	0.510 [0.171–0.849]
D5 Construct	Player experience	55	0.800	0.607 [0.416–0.798]	0.641 [0.448–0.834]
D5 Construct	Health and safety	55	1.000	1.000 (–)	1.000 (–)
D5 Construct	Design feedback	55	0.873	0.657 [0.423–0.890]	0.566 [0.334–0.798]
D5 Construct	Fake review detection	55	1.000	1.000 (–)	1.000 (–)
D5 Construct	Recommendation	55	0.982	0.791 [0.394–1.000]	0.791 [0.394–1.000]
D5 Construct	Reviewer behavior	55	0.927	0.562 [0.181–0.943]	0.813 [0.561–1.000]
D5 Construct	Other construct	55	0.982	0.936 [0.813–1.000]	0.588 [0.314–0.862]
D5 Construct	Construct unclear	55	1.000	n.d. (1.000)	n.d. (1.000)
D7 Study type	Study type	55	0.891	0.778 [0.611–0.946]	0.816 [0.662–0.970]

Appendix D

Table A7. Exploratory logistic regression of non-reporting of each sampling attribute on publication year, added after peer review and not part of the prespecified tests. The year is centered at zero on the minimum of each set.

Set	Attribute Not Reported	Positive	OR Per Year	95% CI	*p*	OR Without 2027	*p*
Full	Platform	101	0.856	0.801–0.914	<0.001	0.858	<0.001
Full	Corpus size	122	0.962	0.909–1.018	0.178	0.961	0.175
Full	Language scope	228	1.000	0.896–1.116	0.997	0.998	0.970
Scopus core	Platform	93	0.849	0.793–0.909	<0.001	0.850	<0.001
Scopus core	Corpus size	116	0.960	0.906–1.018	0.173	0.958	0.150
Scopus core	Language scope	210	0.998	0.895–1.113	0.970	0.997	0.952

No eligible record uses both families. Among the 111 records using a computational text method, logistic regression of the dictionary indicator on publication year gives an odds ratio of 0.908 per year (95% CI [0.778, 1.060], $p = 0.222$); without the 2027 records, 0.891 (95% CI [0.762, 1.042], $p = 0.149$). This test is exploratory.

Table A8. Records using dictionary-based sentiment analysis and transformer-based models by period, as a share of all eligible records and as a share of records using any computational text method (dictionary-based, topic model, embedding-based, transformer-based, classical supervised learning or other deep neural architecture), with 95% Wilson confidence intervals. The two 2027 records are not shown; one uses each family.

Period	All Eligible	Dictionary, *n* (%)	Transformer, *n* (%)	Computational	Dictionary, % [95% CI]	Transformer, % [95% CI]
2006–2018	44	2 (4.5)	0 (0.0)	6	33.3 [9.7, 70.0]	0.0 [0.0, 39.0]
2019–2022	59	3 (5.1)	1 (1.7)	18	16.7 [5.8, 39.2]	5.6 [1.0, 25.8]
2023–2026	140	13 (9.3)	32 (22.9)	85	15.3 [9.2, 24.4]	37.6 [28.1, 48.3]

Appendix E

Table A9. Full cross-tabulation of method family against construct, raw counts, on the 245 eligible records. Column headings abbreviate the constructs of Table 4: Sent, sentiment; Emo, discrete emotion; Tox, toxicity; Help, review helpfulness; Topic, topical content; Purch, purchase behavior; PlayX, player experience; Health, health and safety; Design, design feedback; Fake, fake review detection; Rec, recommendation; RevB, reviewer behavior; Other, other construct; Uncl, construct unclear. Row labels abbreviate the method families of Table 2. Both dimensions are multi-valued, so cell totals exceed the number of records. Of the 168 cells, 121 contain fewer than five records.

Method	Sent	Emo	Tox	Help	Topic	Purch	PlayX	Health	Design	Fake	Rec	RevB	Other	Uncl
Dictionary	14	4	0	2	7	1	10	1	3	0	0	2	2	0
Topic models	19	5	0	0	33	5	23	3	13	2	1	1	2	0
Embedding	7	2	0	1	11	1	4	0	5	1	5	0	2	0
Transformer	28	2	1	0	10	3	10	2	9	0	4	1	3	0
Classical ML	23	4	0	4	5	3	4	1	6	1	2	0	1	0
Deep neural	9	0	0	1	7	0	3	0	4	1	4	0	0	0
Network analysis	0	3	1	0	6	0	4	0	0	0	0	0	1	0
Psychometric	3	1	0	0	9	0	7	0	3	0	0	0	1	0
Statistical	40	8	1	9	10	26	11	2	6	0	1	13	4	0
Qualitative	9	6	2	1	44	8	33	3	23	0	0	6	6	0
Descriptive	2	1	0	0	1	0	1	1	1	0	0	0	0	0
None named	12	6	1	1	21	5	14	2	7	3	3	4	2	2

References

1. Lin, D.; Bezemer, C.-P.; Zou, Y.; Hassan, A.E. An empirical study of game reviews on the Steam platform. *Empir. Softw. Eng.* **2019**, *24*, 170–207. [CrossRef]
2. Martin, W.; Sarro, F.; Jia, Y.; Zhang, Y.; Harman, M. A survey of app store analysis for software engineering. *IEEE Trans. Softw. Eng.* **2017**, *43*, 817–847. [CrossRef]
3. Khalid, H.; Shihab, E.; Nagappan, M.; Hassan, A.E. What do mobile app users complain about? *IEEE Softw.* **2015**, *32*, 70–77. [CrossRef]
4. Pagano, D.; Maalej, W. User feedback in the appstore: An empirical study. In *Proceedings of the 21st IEEE International Requirements Engineering Conference (RE)*; IEEE: New York, NY, USA, 2013; pp. 125–134. [CrossRef]
5. Lin, D.; Bezemer, C.-P.; Hassan, A.E. Studying the urgent updates of popular games on the Steam platform. *Empir. Softw. Eng.* **2017**, *22*, 2095–2126. [CrossRef]
6. Lin, D.; Bezemer, C.-P.; Hassan, A.E. An empirical study of early access games on the Steam platform. *Empir. Softw. Eng.* **2018**, *23*, 771–799. [CrossRef]
7. Chevalier, J.A.; Mayzlin, D. The effect of word of mouth on sales: Online book reviews. *J. Mark. Res.* **2006**, *43*, 345–354. [CrossRef]
8. Dellarocas, C. The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Manag. Sci.* **2003**, *49*, 1407–1424. [CrossRef]
9. Mudambi, S.M.; Schuff, D. What makes a helpful online review? A study of customer reviews on Amazon.com. *MIS Q.* **2010**, *34*, 185–200. [CrossRef]
10. Cao, Q.; Duan, W.; Gan, Q. Exploring determinants of voting for the helpfulness of online user reviews: A text mining approach. *Decis. Support Syst.* **2011**, *50*, 511–521. [CrossRef]

11. Pan, Y.; Zhang, J.Q. Born unequal: A study of the helpfulness of user-generated product reviews. *J. Retail.* **2011**, *87*, 598–612. [CrossRef]
12. Andreev, G.; Saxena, D.; Verma, J.K. Impact of review sentiment and magnitude on customers' recommendations for video games. In *Proceedings of the 2021 International Conference on Computational Performance Evaluation (ComPE)*; IEEE: New York, NY, USA, 2021; pp. 992–995. [CrossRef]
13. Livingston, I.J.; Nacke, L.E.; Mandryk, R.L. The impact of negative game reviews and user comments on player experience. In *Proceedings of the 2011 ACM SIGGRAPH Symposium on Video Games*; ACM: New York, NY, USA, 2011; pp. 25–29. [CrossRef]
14. Livingston, I.J.; Nacke, L.E.; Mandryk, R.L. Influencing experience: The effects of reading game reviews on player experience. In *Entertainment Computing—ICEC 2011*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 89–100. [CrossRef]
15. Joshi, P.; Santy, S.; Budhiraja, A.; Bali, K.; Choudhury, M. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; ACL: Kerrville, TX, USA, 2020; pp. 6282–6293. [CrossRef]
16. Gilardi, F.; Alizadeh, M.; Kubli, M. ChatGPT outperforms crowd workers for text-annotation tasks. *Proc. Natl. Acad. Sci. USA* **2023**, *120*, e2305016120. [CrossRef] [PubMed]
17. Pangakis, N.; Wolken, S.; Fasching, N. Automated annotation with generative AI requires validation. *arXiv* **2023**, arXiv:2306.00176. [CrossRef]
18. Kang, H.-N.; Yong, H.-R.; Hwang, H.-S. A study of analyzing on online game reviews using a data mining approach: STEAM community data. *Int. J. Innov. Manag. Technol.* **2017**, *8*, 90–94. [CrossRef]
19. Wang, X.; Goh, D.H.-L. Components of game experience: An automatic text analysis of online reviews. *Entertain. Comput.* **2020**, *33*, 100338. [CrossRef]
20. Hedegaard, S.; Simonsen, J.G. Extracting usability and user experience information from online user reviews. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2013; pp. 2089–2098. [CrossRef]
21. Zhu, M.; Fang, X. A lexical approach to study computer games and game play experience via online reviews. *Int. J. Hum.-Comput. Interact.* **2015**, *31*, 413–426. [CrossRef]
22. Busurkina, I.; Karpenko, V.; Tulubenskaya, E.; Bulygin, D. Game experience evaluation: A study of game reviews on the Steam platform. In *Communications in Computer and Information Science*; Springer: Cham, Switzerland, 2020; pp. 117–127. [CrossRef]
23. Granic, I.; Lobel, A.; Engels, R.C.M.E. The benefits of playing video games. *Am. Psychol.* **2014**, *69*, 66–78. [CrossRef] [PubMed]
24. Bopp, J.A.; Mekler, E.D.; Opwis, K. Negative emotion, positive experience? Emotionally moving moments in digital games. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2016; pp. 2996–3006. [CrossRef]
25. Chambers, C.; Feng, W.-C.; Sahu, S.; Saha, D. Measurement-based characterization of a collection of on-line games. In *Proceedings of the 5th ACM SIGCOMM Conference on Internet Measurement (IMC '05)*; ACM: New York, NY, USA, 2005; pp. 1–14.
26. Guzman, E.; Maalej, W. How do users like this feature? A fine grained sentiment analysis of app reviews. In *Proceedings of the 22nd IEEE International Requirements Engineering Conference (RE)*; IEEE: New York, NY, USA, 2014; pp. 153–162. [CrossRef]
27. Hutto, C.; Gilbert, E. VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proc. Int. AAAI Conf. Web Soc. Media* **2014**, *8*, 216–225. [CrossRef]
28. Mohammad, S.; Turney, P. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*; ACL: Kerrville, TX, USA, 2010; pp. 26–34.
29. Kiritchenko, S.; Zhu, X.; Cherry, C.; Mohammad, S. NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*; ACL: Kerrville, TX, USA, 2014; pp. 437–442.
30. Bravo-Marquez, F.; Frank, E.; Pfahringer, B.; Mohammad, S.M. AffectiveTweets: A Weka package for analyzing affect in tweets. *J. Mach. Learn. Res.* **2019**, *20*, 1–6.
31. Liu, B. *Sentiment Analysis and Opinion Mining*; Synthesis Lectures on Human Language Technologies (SLHLT); Springer: Cham, Switzerland, 2012; Volume 5, pp. 1–167. [CrossRef]
32. Tang, H.; Tan, S.; Cheng, X. A survey on sentiment detection of reviews. *Expert Syst. Appl.* **2009**, *36*, 10760–10773. [CrossRef]
33. Zhang, H.; Gan, W.; Jiang, B. Machine learning and lexicon based methods for sentiment classification: A survey. In *Proceedings of the 11th Web Information System and Application Conference*; IEEE: New York, NY, USA, 2014; pp. 262–265. [CrossRef]
34. Guzsvinecz, T.; Szűcs, J. Length and sentiment analysis of reviews about top-level video game genres on the Steam platform. *Comput. Hum. Behav.* **2023**, *149*, 107955. [CrossRef]
35. Viggiato, M.; Lin, D.; Hindle, A.; Bezemer, C.-P. What causes wrong sentiment classifications of game reviews? *IEEE Trans. Games* **2022**, *14*, 350–363. [CrossRef]
36. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
37. Li, X.; Zhang, Z.; Stefanidis, K. A data-driven approach for video game playability analysis based on players' reviews. *Information* **2021**, *12*, 129. [CrossRef]

38. Grootendorst, M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv* **2022**, arXiv:2203.05794. [CrossRef]
39. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*; ACL: Kerrville, TX, USA, 2019; pp. 3982–3992. [CrossRef]
40. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv* **2019**, arXiv:1907.11692. [CrossRef]
41. Hanu, L. Detoxify. GitHub Repository. 2020. Available online: https://github.com/unitaryai/detoxify (accessed on 27 August 2026).
42. Guzsvinecz, T.; Szűcs, J. Genre matters: Analyzing toxicity patterns in Steam game reviews. In *Proceedings of the 2026 IEEE 24th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*; IEEE: New York, NY, USA, 2026; pp. 635–640. [CrossRef]
43. Purbolaksono, M.D. Sentiment analysis of game review in Steam platform using random forest. *Int. J. Inf. Commun. Technol.* **2024**, *10*, 161–169. [CrossRef]
44. Urriza, I.M.; Clarino, M.A.A. Aspect-based sentiment analysis of user created game reviews. In *Proceedings of the 24th Conference of the Oriental COCOSDA*; IEEE: New York, NY, USA, 2021; pp. 76–81. [CrossRef]
45. Yu, Y.; Dinh, D.-T.; Nguyen, B.-H.; Yu, F.; Huynh, V.-N. Mining insights from esports game reviews with an aspect-based sentiment analysis framework. *IEEE Access* **2023**, *11*, 61161–61172. [CrossRef]
46. Mozetič, I.; Grčar, M.; Smailović, J. Multilingual Twitter sentiment classification: The role of human annotators. *PLoS ONE* **2016**, *11*, e0155036. [CrossRef] [PubMed]
47. Ho, V.A.; Nguyen, D.H.-C.; Nguyen, D.H.; Pham, L.T.-V.; Nguyen, D.-V.; Van Nguyen, K.; Nguyen, N.L.-T. Emotion recognition for Vietnamese social media text. In *Communications in Computer and Information Science*; Springer: Singapore, 2020; pp. 319–333. [CrossRef]
48. Cataldi, M.; Ballatore, A.; Tiddi, I.; Aufaure, M.-A. Good location, terrible food: Detecting feature sentiment in user-generated reviews. *Soc. Netw. Anal. Min.* **2013**, *3*, 1149–1163. [CrossRef]
49. Werder, K.; Brinkkemper, S. MEME: Toward a method for emotions extraction from GitHub. In *Proceedings of the 3rd International Workshop on Emotion Awareness in Software Engineering*; IEEE: New York, NY, USA; ACM: New York, NY, USA, 2018; pp. 20–24. [CrossRef]
50. Johnson, D.; Nacke, L.E.; Wyeth, P. All about that base: Differing player experiences in video game genres and the unique case of MOBA games. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2015; pp. 2265–2274. [CrossRef]
51. Jagoda, P. On difficulty in video games: Mechanics, interpretation, affect. *Crit. Inq.* **2018**, *45*, 199–233. [CrossRef]
52. Toh, W. The player experience and design implications of narrative games. *Int. J. Hum.-Comput. Interact.* **2023**, *39*, 2742–2769. [CrossRef]
53. Wright, S.; Denisova, A. "It's a terrible choice to make but also a necessary one": Exploring player experiences with moral decision making mechanics in video games. *Comput. Hum. Behav.* **2024**, *161*, 108424. [CrossRef]
54. Gowler, C.P.R.; Iacovides, I. "Horror, guilt and shame": Uncomfortable experiences in digital games. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*; ACM: New York, NY, USA, 2019; pp. 325–337. [CrossRef]
55. Dahl, G.; Kraus, M. Measuring how game feel is influenced by the player avatar's acceleration and deceleration. In *Proceedings of the 19th International Academic Mindtrek Conference*; ACM: New York, NY, USA, 2015; pp. 41–46. [CrossRef]
56. Wehbe, R.R.; Mekler, E.D.; Schaekermann, M.; Lank, E.; Nacke, L.E. Testing incremental difficulty design in platformer games. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2017; pp. 5109–5113. [CrossRef]
57. Summerville, A.; Mariño, J.R.H.; Snodgrass, S.; Ontañón, S.; Lelis, L.H.S. Understanding Mario: An evaluation of design metrics for platformers. In *Proceedings of the 12th International Conference on the Foundations of Digital Games*; ACM: New York, NY, USA, 2017; pp. 1–10. [CrossRef]
58. Liao, S.-C.; Wu, F.-G.; Feng, C.-H.; Shuai, C.-Y. A research on how to enhance user experience by improving arcade joystick in side-scrolling shooter games. In *Communications in Computer and Information Science*; Springer: Cham, Switzerland, 2020; pp. 431–437. [CrossRef]
59. Maubert Crotte, A.; Hepting, D.H.; Roshchina, A. Left-handed control configuration for side-scrolling games. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2019; pp. 1–6. [CrossRef]
60. Ho, J.C.F.; Zhang, X. Strategies for marketing really new products to the mass market: A text mining-based case study of virtual reality games. *J. Open Innov. Technol. Mark. Complex.* **2020**, *6*, 1. [CrossRef]
61. Zheng, L. A study on targeted advertising strategies for virtual goods on social media. *Commun. Humanit. Res.* **2024**, *63*, 48–55. [CrossRef]
62. Blancaflor, E.; San Miguel, J.M.G. Analyzing digital game distribution in gaming industry: A case study. In *Proceedings of the 2nd Indian International Conference on Industrial Engineering and Operations Management*; IEOM Society International: Warangal, India, 2022; pp. 674–681. [CrossRef]

63. Fox, J.; Gilbert, M.; Tang, W.Y. Player experiences in a massively multiplayer online game: A diary study of performance, motivation, and social interaction. *New Media Soc.* **2018**, *20*, 4056–4073. [CrossRef]
64. Mora-Cantallops, M.; Sicilia, M.-Á. Exploring player experience in ranked League of Legends. *Behav. Inf. Technol.* **2018**, *37*, 1224–1236. [CrossRef]
65. Abbasi, A.Z.; Rehman, U.; Fayyaz, M.S.; Ting, D.H.; Shah, M.U.; Fatima, R. Using the playful consumption experience model to uncover behavioral intention to play multiplayer online battle arena games. *Data Technol. Appl.* **2022**, *56*, 223–246. [CrossRef]
66. Zhu, J.; Ontañón, S. Experience management in multi-player games. In *Proceedings of the 2019 IEEE Conference on Games (CoG)*; IEEE: New York, NY, USA, 2019; pp. 1–6. [CrossRef]
67. Andreean, A.F.F.; Anugrah, M.F.; Tho, C. Exploring of emotional character in online games toward player. *Procedia Comput. Sci.* **2023**, *227*, 743–752. [CrossRef]
68. Gao, Y.; Chen, A.; Chi, S.; Zhang, G.; Hao, A. Analysis of emotional tendency and syntactic properties of VR game reviews. In *Proceedings of the 2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*; IEEE: New York, NY, USA, 2022; pp. 648–649. [CrossRef]
69. Guzsvinecz, T.; Németh, K.; Szűcs, J. Cybersickness in commercial first-person virtual reality video games: Analysis of player experience in Steam reviews. *Virtual Real.* **2026**, *30*, 83. [CrossRef]
70. Zsila, Á.; Shabahang, R.; Aruguete, M.S.; Orosz, G. Toxic behaviors in online multiplayer games: Prevalence, perception, risk factors of victimization, and psychological consequences. *Aggress. Behav.* **2022**, *48*, 356–364. [CrossRef] [PubMed]
71. Cook, C.L.; Tang, S.Y.C.; Lin, J.H.T. Comparing shades of darkness: Trolling victims' experiences on social media vs. online gaming. *Front. Psychol.* **2023**, *14*, 1163244. [CrossRef] [PubMed]
72. Frommel, J.; Johnson, D.; Mandryk, R.L. How perceived toxicity of gaming communities is associated with social capital, satisfaction of relatedness, and loneliness. *Comput. Hum. Behav. Rep.* **2023**, *10*, 100302. [CrossRef]
73. Beres, N.A.; Frommel, J.; Reid, E.; Mandryk, R.L.; Klarkowski, M. Don't you know that you're toxic: Normalization of toxicity in online gaming. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2021; pp. 1–15. [CrossRef]
74. Türkay, S.; Formosa, J.; Adinolf, S.; Cuthbert, R.; Altizer, R. See no evil, hear no evil, speak no evil: How collegiate players define, experience and cope with toxicity. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2020; pp. 1–13. [CrossRef]
75. Adinolf, S.; Turkay, S. Toxic behaviors in esports games: Player perceptions and coping strategies. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*; ACM: New York, NY, USA, 2018; pp. 365–372. [CrossRef]
76. Achterbosch, L.; Vamplew, P.; March, E. Assessing the impact of griefing in MMORPGs using self-determination theory. *Comput. Hum. Behav.* **2024**, *161*, 108388. [CrossRef]
77. Suler, J. The online disinhibition effect. *Cyberpsychol. Behav.* **2004**, *7*, 321–326. [CrossRef] [PubMed]
78. Wells, G.; Romhanyi, A.; Steinkuehler, C. Hate speech and hate-based harassment in online games. *Front. Psychol.* **2024**, *15*, 1422422. [CrossRef] [PubMed]
79. Kowert, R.; Kilmer, E.; Newhouse, A. Taking it to the extreme: Prevalence and nature of extremist sentiment in games. *Front. Psychol.* **2024**, *15*, 1410620. [CrossRef] [PubMed]
80. Cantone, G.G.; Tomaselli, V.; Mazzeo, V. Review bombing: Ideology-driven polarisation in online ratings: The case study of The Last of Us (part II). *Qual. Quant.* **2025**, *59*, 315–341. [CrossRef]
81. Cheuque, G.; Guzmán, J.; Parra, D. Recommender systems for online video game platforms: The case of STEAM. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion)*; ACM: New York, NY, USA, 2019; pp. 763–771. [CrossRef]
82. Hsu, H.; O'Hagan, M. International players' perceptions of localization in their gameplay experiences: An explorative study with Steam user game reviews. *Transl. Spaces* **2025**, *14*, 1–24. [CrossRef]
83. Petrovskaya, E.; Deterding, S.; Zendle, D.I. Prevalence and salience of problematic microtransactions in top-grossing mobile and PC games: A content analysis of user reviews. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2022; pp. 1–12. [CrossRef]
84. Genc-Nayebi, N.; Abran, A. A systematic literature review: Opinion mining studies from mobile app store user reviews. *J. Syst. Softw.* **2017**, *125*, 207–219. [CrossRef]
85. Jha, N.; Mahmoud, A. Mining non-functional requirements from app store reviews. *Empir. Softw. Eng.* **2019**, *24*, 3659–3695. [CrossRef]
86. Al-Subaihin, A.A.; Sarro, F.; Black, S.; Capra, L.; Harman, M. App store effects on software engineering practices. *IEEE Trans. Softw. Eng.* **2021**, *47*, 300–319. [CrossRef]
87. Maalej, W.; Kurtanović, Z.; Nabil, H.; Stanik, C. On the automatic classification of app reviews. *Requir. Eng.* **2016**, *21*, 311–331. [CrossRef]

88. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef] [PubMed]
89. Petersen, K.; Feldt, R.; Mujtaba, S.; Mattsson, M. Systematic mapping studies in software engineering. In *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering (EASE)*; BCS: Swindon, UK, 2008; pp. 68–77. [CrossRef]
90. Petersen, K.; Vakkalanka, S.; Kuzniarz, L. Guidelines for conducting systematic mapping studies in software engineering: An update. *Inf. Softw. Technol.* **2015**, *64*, 1–18. [CrossRef]
91. Piwowar, H.; Priem, J.; Larivière, V.; Alperin, J.P.; Matthias, L.; Norlander, B.; Farley, A.; West, J.; Haustein, S. The state of OA: A large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ* **2018**, *6*, e4375. [CrossRef] [PubMed]
92. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
93. Fleiss, J.L.; Cohen, J.; Everitt, B.S. Large sample standard errors of kappa and weighted kappa. *Psychol. Bull.* **1969**, *72*, 323–327. [CrossRef]
94. Wilson, E.B. Probable inference, the law of succession, and statistical inference. *J. Am. Stat. Assoc.* **1927**, *22*, 209–212. [CrossRef]
95. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159–174. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.