Research Square

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

GamerLex: A sentiment lexicon of gaming slang and jargon derived from Steam reviews

Tibor Guzsvinecz

guzsvinecz.tibor@zek.uni-pannon.hu

University of Pannonia

Article

Keywords: gaming slang, sentiment analysis, sentiment lexicon, Steam, VADER, video game reviews

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11503520/v1

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

Additional Declarations: No competing interests reported.

Abstract

Sentiment lexicons such as VADER were built for general and social media English, and they misread the vocabulary of video game reviews. They miss slang and jargon such as cashgrab, unplayable or goated, and they score descriptive words such as kill, scary or brutal as strongly negative. This study presents GamerLex, a sentiment lexicon of gaming slang and game jargon for player reviews, built from 4.69 million Steam reviews. Candidate terms came from embedding-based expansion of slang seeds, two rounds of manual review and an audit against a glossary of game terminology. Each term receives a valence on the VADER scale from its association with players' recommendations. This association is estimated on a training split as a Mantel-Haenszel log odds ratio stratified by review length and calibrated on 1,449 VADER words used as anchors. The release contains 4,223 hand-audited entries in three tiers and an optional automatic tier of 3,823 frequent review words. On 99,389 held-out reviews, adding GamerLex to VADER raised the length-stratified AUC for predicting recommendations from 0.825 to 0.893. On 1,462 review sentences rated by three raters, the Spearman correlation with the median rating rose from 0.579 to 0.664, above two uncurated lexicons induced from the same data (0.634 and 0.620). The gain over VADER held when the review label was controlled for, in an approximately unselected subset of sentences and in sentences without terms added during the lexicon audit. The lexicon, its corpus statistics and the evaluation code are openly available.

1. Introduction

Player reviews are one of the largest sources of evaluative text about video games. Steam, the dominant digital distribution platform for PC games, lists more than 100,000 released titles (SteamDB, 2026), and every review on it carries the reviewer's binary recommendation. This makes the platform a natural test bed for sentiment analysis at scale. Studies have used it to describe review length and emotional content across top-level genres (TLGs) in more than 30 million reviews (Guzsvinecz & Szűcs, 2023), to diagnose why sentiment classifiers fail on 12 million reviews (Viggiato et al., 2022), to predict review helpfulness (Eberhard et al., 2018; Wang et al., 2021) and to rate game aspects (Yuan et al., 2025).

Most of this work relies on sentiment tools built for other kinds of text. VADER (Hutto & Gilbert, 2014) was constructed from general and social media vocabulary rated by crowd workers, with social media text as its primary target. The NRC Word-Emotion Association Lexicon (Mohammad & Turney, 2013), which has been applied to Steam reviews (Guzsvinecz, 2023; Guzsvinecz & Szűcs, 2023), covers general English. Neither was designed for the vocabulary of game reviews. The cost of this mismatch has been measured. Viggiato et al. (2022) applied Stanford CoreNLP, the machine-learning classifier of NLTK and SentiStrength to Steam reviews and found a median AUC of 0.70 for the best of the three tools. In a manual analysis of 382 reviews that all three tools misclassified, 23% of the errors were attributed to negative terminology: words such as kill that are negative in general English but not in games. The authors called for techniques that deal automatically with the game-related issues of reviews.

Game reviews mix two kinds of vocabulary that general lexicons handle poorly. The first is community slang and evaluative shorthand. *Goated*, *banger* and *mid* praise or dismiss a game in one word, *cashgrab* and *p2w* (pay to win) criticise its business model, and *gg*, *rekt* or *tilted* carry reactions over from multiplayer play. The second is game jargon, the descriptive vocabulary of mechanics and genres. Here the problem is often the reverse: VADER knows the word but gives it a sentiment that does not fit the domain. It rates kill at -3.7, brutal at -3.1, terrifying at -2.7 and scary at -2.2 on its − 4 to + 4 scale, yet in a review of a shooter or a horror game these words usually describe the content, and they often praise it.

Two sentences from our benchmark show both errors. *"Clunky controls, very sparse and vague tutorial, cringe worthy dialogue."* receives a positive VADER compound score of 0.30, because worthy carries a positive value while clunky and cringe are missing from VADER's vocabulary; the median of three human ratings is -2. *"Definitely a good way to kill time."* scores − 0.03, because kill cancels good, although the median human rating is + 2. With GamerLex added to VADER the two scores become − 0.61 and + 0.65. Such vocabulary is common: 62.8% of a random sample of 99,389 Steam reviews contain at least one term of the curated GamerLex vocabulary.

Sentiment resources for games exist, but they were built for other text types. Bergsma et al. (2020) adapted an affective lexicon to the dialogue of a single role-playing game. Lexicons of Twitch emotes and chat vocabulary have been built from viewer ratings (Kobs et al., 2020) and from word embeddings (Dolin et al., 2021; Ismail et al., 2025). Islam et al. (2020) induced domain lexicons from Amazon star ratings, one of them for video game products. To our knowledge, no openly released lexicon has been built for the slang and jargon of player reviews on a game platform, with valence on the VADER scale and an evaluation against human ratings.

This paper introduces GamerLex, a sentiment lexicon of gaming slang and game jargon built from 4.69 million Steam reviews. Candidate terms came from embedding-based expansion of slang seeds, two rounds of manual review and an audit against a glossary of game terminology. Ratings are not collected for each term. Valence comes from the company a term keeps: its association with players' recommendations, estimated on a training split, adjusted for review length and mapped to the VADER scale through anchor words, frequent words that VADER rates and GamerLex does not contain. The released lexicon has three hand-audited tiers and an optional automatic tier of frequent review words selected by a fixed rule. It is evaluated on held-out reviews and on 1,462 review sentences rated by three raters, against VADER and against two uncurated lexicons induced from the same data.

The paper makes three contributions:

- GamerLex v1.0, with 8,046 entries (4,223 hand-audited and 3,823 automatic), valence on the VADER scale, corpus statistics for every entry and a helper module for use with VADER, released under an open licence.
- A procedure that derives valence from review-level recommendations, corrects for review length and calibrates the result to an existing lexicon, so that domain terms and general words share one scale.

- An evaluation design that separates the contribution of the curated vocabulary from that of corpus-derived valence as such, by comparing GamerLex with uncurated lexicons built by the same procedure. The design includes checks for sentence selection and for the influence of the review label on human ratings.

Section 2 reviews related work. Section 3 describes the corpus, the construction of the candidate vocabulary, the valence procedure and the evaluation design. Section 4 reports the results, Section 5 discusses them together with the limitations, and Section 6 concludes.

2. Related work

The work presented in this paper sits at the intersection of four research strands: general-purpose sentiment lexicons, domain-specific lexicon induction, slang-aware sentiment analysis, and the computational study of gaming discourse. Each of these works is reviewed in turn, the limitations that motivate the present contribution are identified, and the work is situated in the unfilled space at their intersection.

2.1. General-purpose sentiment lexicons

Early sentiment lexicons such as the General Inquirer (Stone et al., 1966), LIWC (Pennebaker et al., 2001) and the opinion lexicon of Hu and Liu (2004) assign words to positive and negative categories without intensity. Later resources added graded scores: ANEW (Bradley & Lang, 1999) collected valence, arousal and dominance ratings, and SentiWordNet (Baccianella et al., 2010) propagated scores over WordNet synsets. VADER (Hutto & Gilbert, 2014) combined vocabulary from such sources with emoticons, acronyms and common slang, collected ten crowd ratings for each candidate feature and kept just over 7,500 features with validated valence on a -4 to + 4 scale. Its rules handle negation, degree modifiers, contrastive conjunctions, capitalisation and punctuation, and on tweets it outperformed individual human raters (*F1* = 0.96 against 0.84). The NRC Word-Emotion Association Lexicon (Mohammad & Turney, 2013) associates words with basic emotions and with positive and negative polarity, and it has been used on Steam reviews (Guzsvinecz, 2023; Guzsvinecz & Szűcs, 2023). All of these resources were built from general English or social media vocabulary. GamerLex is designed as an extension of VADER: it keeps VADER's rules and vocabulary and adds or replaces word values for the review domain.

2.2. Domain-specific lexicon induction

Because word polarity depends on the domain, many studies induce lexicons from in-domain text. Corpus-based methods propagate polarity from seed words through co-occurrence statistics (Turney & Littman, 2003), lexical graphs (Velikovich et al., 2010) or word embeddings. SentProp (Hamilton et al., 2016) combined domain-specific embeddings with label propagation and produced sentiment lexicons for 250 Reddit communities, showing that the same word can carry different polarity in different communities. A second line of work uses document labels as supervision. Islam et al. (2020) induced lexicons for 24 domains from the star ratings of Amazon reviews, one of them video games. GamerLex

follows this second line. Steam’s recommendation flag is a document label written by the reviewer and available for millions of reviews, so word valence can be estimated directly, without seed words or embedding neighbourhoods. Embeddings were used only to propose candidate terms (Section 3.2).

2.3. Slang-aware sentiment analysis

Slang presents a distinct challenge from domain specificity. Where domain-specific vocabulary tends to be relatively stable within a domain, slang evolves rapidly, varies across communities, frequently inverts the conventional polarity of its source words (sick and fire as positive evaluations; mid as negative), and propagates across community boundaries in ways that destabilize previously calibrated lexical resources (Eble, 1996; Eisenstein, 2013). Out-of-vocabulary slang terms are a documented source of error in sentiment classification (Maity et al., 2016; Babanejad et al., 2020).

The first dedicated effort to address slang in sentiment analysis was SlangSD (Wu et al., 2016), constructed at Arizona State University through a multi-stage process exploiting Urban Dictionary's entry structure, synonym relations, and existing sentiment lexicons. The first version of SlangSD contains 96,462 slang words and phrases with sentiment scores. The authors validate SlangSD on Twitter sentiment classification, demonstrating modest but consistent improvements over generic lexicons. However, SlangSD is general-purpose: it was constructed without domain filtering and was not validated on any specialized domain. Its coverage of gaming-specific terms is incidental rather than systematic.

Sun et al. (2021) developed a computational framework for slang generation using a publicly available subset of Urban Dictionary alongside the Online Slang Dictionary and Green's Dictionary of Slang. Their framework focuses on modeling the speaker's word choice in slang context through contrastive learning over conventional and slang senses. Their treatment focuses on slang generation rather than detection or sentiment classification.

More recent work has begun to explore large language models for slang processing. Marcondes et al. (2024) asked ChatGPT to annotate the VADER lexicon and reported fair performance, enough for a first annotation that domain specialists then refine. Their test did not involve the vocabulary of a particular community, such as gaming slang; the present work combines curated term lists with valences estimated from corpus evidence.

Slang resources built for gaming are rare. The CONDA dataset (Weld et al., 2021) tags game slang at the token level in Dota 2 chat but assigns no sentiment, and the gaming lexicons reviewed in Section 2.4 cover Twitch emotes and chat or the dialogue of a single game rather than player reviews.

2.4. Computational analysis of gaming discourse

The application of natural language processing methods to gaming text has grown largely over the past decade and now constitutes a recognizable subdomain of computational social science and games research. The existing literature concentrates in two areas that lie adjacent to (but do not overlap with) the present study.

In-game chat and toxicity detection is the more mature of these areas. Kwak and Blackburn (2015) conducted a pioneering linguistic analysis of over 590,000 League of Legends cases involving accused toxic players, identifying distinctive lexical and syntactic patterns of toxic behavior and documenting how players transition from typical to toxic interaction styles. Märtens et al. (2015) studied profanity and toxicity detection in the chat of multiplayer online battle arena games. The CONDA dataset (Weld et al., 2021) provides 44,869 utterances from 1,921 Dota 2 matches with token-level annotations across six slot labels: T (Toxicity), C (Character), D (Dota-specific), S (Game Slang), P (Pronoun), and O (Other). CONDA is the only published dataset to our knowledge that annotates game slang at the token level. However, the dataset is limited to a single game (Dota 2), uses short in-match utterances rather than long-form reviews, focuses on toxicity classification rather than sentiment, and does not assign sentiment polarity to slang terms.

Subsequent toxicity work has produced BERT-based classifiers such as ToxBuster (Yang et al., 2023), attention-based sequence taggers (Jia et al., 2023), and comparisons of embeddings, fine-tuned transformers and large language models for gaming chat (Tereshchenko & Hämäläinen, 2025). Across this literature, the focus remains on toxicity classification rather than slang lexicon construction or sentiment-aware analysis.

The second area of computational gaming work is Steam review analysis. Early studies in this strand focused on helpfulness prediction (Eberhard et al., 2018; Kang et al., 2017; Wang et al., 2021) and opinion extraction (Sirbu et al., 2016). Aspect-based sentiment analysis on Steam reviews has been examined by Urriza and Clariño (2021) and Yuan et al. (2025), the latter comparing VADER and TextBlob on Grand Theft Auto V and Cyberpunk 2077 reviews. At a larger scale, two lines of work merit detailed discussion. Viggiato et al. (2022) examined about 12 million Steam reviews, of which 6.85 million were in English, with three sentiment classifiers: Stanford CoreNLP, the machine-learning classifier of NLTK and SentiStrength. The median AUC of the best tool was 0.70. Training per genre helped, with median AUC values between 0.69 and 0.74 depending on the genre. A manual analysis of 382 reviews that all three tools misclassified identified six causes: contrast conjunctions, comparison with other games, negative terminology, unclear reviews, sarcasm and a recommendation that does not match the text. Negative terminology, words such as kill that are not negative in games, accounted for 23% of the cases. The authors called for techniques that deal automatically with the game-related issues of reviews.

Guzsvinecz and Szűcs (2023), working with more than 30 million Steam reviews across eleven TLGs, used the NRC Word-Emotion Association Lexicon to characterize emotional content variation across genres. They report that role-playing, simulation, and survival games generate longer reviews with deeper engagement, while racing games produce shorter and more functional reviews. The study established baseline emotional patterns across the Steam corpus with a general-purpose emotion lexicon. The present work builds on it by adding a sentiment resource tailored to the vocabulary of game reviews. The companion study by Guzsvinecz (2023) on Souls-like role-playing game reviews extends this approach to a single genre subcategory, applying similar lexical techniques to a more homogeneous corpus.

Beyond review analysis specifically, computational gaming work has examined the characteristics of Steam reviews and their reception (Lin et al., 2019), popularity prediction (De Luisa et al., 2021), and the relationship between emotional review content and platform-level recommendation outcomes (Andreev et al., 2021). Across all of this work, sentiment analysis is treated as a stable input to downstream questions rather than a methodology requiring domain-specific calibration.

A few sentiment resources have been built for games and gaming platforms. Bergsma et al. (2020) adapted an existing affective lexicon to the dialogue of The Elder Scrolls V: Skyrim by assigning game-specific words the valence of their nearest neighbours in an embedding space trained on in-game lore; they described this as the first adaptation of a sentiment lexicon to the video game domain. On Twitch, Kobs et al. (2020) obtained sentiment labels for the 100 most frequent emotes through a survey of users, Dolin et al. (2021) inferred scores for more than 22,000 emotes from their embedding neighbours in VADER and an emoji lexicon, and Ismail et al. (2025) extended VADER with emotes and embedding-based valences to label toxic chat. Islam et al. (2020) induced lexicons from Amazon star ratings, including a video game product domain. These resources differ from GamerLex in text type (in-game dialogue, live chat, product reviews), in vocabulary (emotes rather than slang and jargon) or in the source of valence (embedding neighbourhoods or product ratings rather than the recommendations of players on a game platform).

2.5. Positioning of the Present Work

General-purpose lexicons, VADER among them, lack the slang and jargon of game reviews and misjudge the descriptive vocabulary of games. The gaming lexicons that exist were built for live chat, emotes or the dialogue of one game, and none of them was evaluated on player reviews. Lexicon induction from document labels has been demonstrated on product reviews, but the labels used there were star ratings of products sold in a general store, and confounds such as review length were not addressed. GamerLex combines a curated vocabulary for player reviews with valence estimated from recommendations, adjusted for review length. Its evaluation compares the curated vocabulary with uncurated lexicons built from the same data, so that the contribution of curation can be separated from that of corpus-derived valence. The lexicon is meant as an extension of VADER for researchers who need an interpretable, rule-based score, not as a replacement for supervised classifiers.

3. Materials and methods

3.1. Corpus

The corpus derives from the collection of more than 30 million English-language Steam review records analysed by Guzsvinecz and Szűcs (2023), which was collected between mid-August and mid-September 2022. In this collection a review is listed once for each TLG of its game. After records with missing fields were removed, duplicates were identified on the review text together with the reviewer's playtime at the time of the review, and only the first occurrence was kept. This left 4,691,606 unique reviews; 85.11% of the records were duplicates. Each review carries the reviewer's binary

recommendation and one TLG label. Because the records are ordered alphabetically by TLG, the retained label is the alphabetically first TLG of the game: a review labelled Adventure, for example, comes from a game that is not tagged as Action. All counts and term matching use one reference tokenizer: lower-casing followed by the regular expression *[a-z0-9'*-]+*, which keeps apostrophes, hyphens and the asterisks of censored words inside tokens. Multi-word terms are matched as token sequences.

31,252 reviews without tokens and 2,148 reviews containing a sentence of the evaluation benchmark (Section 3.6) were discarded, so that no benchmark text could influence valence estimation. The remaining reviews were split at random (seed 42) into a training part (3,261,496 reviews, 70%), a validation part (465,514, 10%) and a test part (931,196, 20%). The share of recommending reviews was 0.814 in both the training and the test part. All statistics that enter the lexicon were computed on the training part only.

3.2. Candidate vocabulary

Seed term candidates came from four sources: a manual list of 101 common gaming slang terms; 608 headwords of the Wikipedia Glossary of video game terms; 4,335 entries of the Urban Dictionary Words And Definitions dataset on Kaggle (therohk, n.d.) with at least 50 up-votes, kept when the headword was a known gaming term or the definition contained an explicit gaming reference or at least two gaming keywords; and 1,885 of the 2,000 most frequent tokens in a random sample of 500,000 reviews after stop-word removal, once overlaps with the other sources were removed. Each of the 6,929 candidates was labelled as gaming slang or jargon or not. This gave 3,012 seeds: all 101 manual terms, 361 glossary headwords, 2,513 Urban Dictionary entries and 37 frequent corpus tokens.

A skip-gram FastText model (Bojanowski et al., 2017) was trained on the review corpus with 150 dimensions, a context window of 7, 10 epochs, a minimum count of 10, 5 negative samples, character n-grams of 2 to 6 characters and a sampling threshold of 10^{-4}; its vocabulary contains 123,961 words. In the first round, each seed was expanded with up to 30 nearest neighbours with a cosine similarity of at least 0.55, which gave 38,961 entries including the seeds. After punctuation variants were normalised and merged and URLs, file paths, numeric and very short tokens and common English words were removed, 26,544 unique candidates remained. Of these, 1,001 were accepted as gaming slang or jargon. In the second round the 1,001 accepted terms served as seeds with the same settings, which gave 13,149 entries and, after the same cleaning, 9,081 candidates; 3,584 were accepted. Together the two rounds gave 3,880 unique terms, each assigned a coarse category (mechanic, reaction, industry, skill, toxic or meta). In both expansion rounds, a rule-based script pre-filled the labels from the source and the similarity of each candidate, and the author reviewed them; the second round was reviewed in six batches.

After the scope of the lexicon was widened from slang to slang and game jargon, the 3,880 terms were audited together with 1,857 additional candidates. Most of the additions (1,406) were terms that the earlier rounds had rejected as not being slang: 1,123 first-round candidates, 234 seed candidates and 49 second-round candidates. They included basic game vocabulary, resource and economy terms of

strategy games, combat and horror vocabulary and slang with reversed polarity. The remaining 451 were added by hand: 366 headwords of the Wikipedia Glossary of video game terms (Wikipedia, 2026) that were still missing (for example boss, level, gameplay, patch), 52 words taken from the VADER lexicon for the override tier (for example terrifying), and 33 terms proposed by the author. Each of the 5,737 candidates was removed or assigned to one of three tiers:

- *core*: gaming slang and game jargon;
- *extended*: general evaluative vocabulary absent from VADER, including frequent misspellings;
- *vader_override*: words that VADER rates as negative but that describe content or express praise in game reviews, such as combat vocabulary (kill, battle), horror vocabulary (scary, terrifying) and slang with reversed polarity (sick, insane).

The audit removed 1,393 candidates (e.g. general English words without a gaming sense, proper names and fragments) and left 4,344 entries. Two kinds of information were visible during the audit that later also entered the evaluation: corpus statistics computed on all reviews, and the number of benchmark sentences that contain each candidate. They could have influenced which terms were kept, although not their valences, which were estimated on the training part only (Section 3.3); the automatic tier was selected by rule without either.

3.3. Valence estimation

Valence is estimated from the association between a term and the recommendation of the reviews that contain it. A raw comparison is biased by review length: short reviews are mostly positive and long reviews more often negative (in the training part, 90.2% of reviews with 1 to 5 tokens recommend the game against 69.7% of reviews with more than 320 tokens), and a frequent term is more likely to occur in a long review. Training reviews were therefore stratified into eight length bands (1–5, 6–10, 11–20, 21–40, 41–80, 81–160, 161–320 and more than 320 tokens). For each term *w*, the Mantel-Haenszel common odds ratio (Mantel & Haenszel, 1959) of recommendation was computed for reviews with and without *w*:

$$\lambda_w = log\left[\frac{\left(\frac{\Sigma_b A_b D_b}{T_b + 0.5}\right)}{\left(\frac{\Sigma_b B_b C_b}{T_b + 0.5}\right)}\right],$$

where, in band *b*, A_b and B_b are the recommending and non-recommending reviews that contain *w*, C_b and D_b those that do not, and T_b is the number of reviews in the band. The constant 0.5 keeps the ratio finite for rare terms. A standardized difference z_w is also reported, comparing the observed number of recommending reviews containing *w* with the expected number. The expectation uses the recommending share of each band.

To place λ_w on the VADER scale, shrinkage towards zero was applied for rare terms, and the result was multiplied by a slope *k*:

$$v_w = clip\left(k\,\lambda_w \frac{n_w}{(n_w + 25)},\ -4,\ 4\right),$$

where n_w is the number of training reviews that contain *w*. The slope was fitted by least squares through the origin on anchor words, that is, words that VADER contains, GamerLex does not, and that occur in at least 500 training reviews. Two mappings were fitted: a single slope, and separate slopes for positive and negative λ. The mapping was chosen on a validation sample of 49,475 reviews by the AUC of VADER extended with GamerLex in predicting the recommendation.

Three filters, all based on the training part, were applied before release. Entries that the reference tokenizer cannot match were removed. So were entries with fewer than five training reviews. A *vader_override* candidate whose corpus association was also clearly negative (observed minus length-expected recommending share of -0.10 or less, and $z \le -3$) was removed so that VADER's value applies.

To use GamerLex, its entries are added to VADER's lexicon, replacing VADER's value where both contain a word, and multi-word terms are joined with underscores before scoring. VADER's rules for negation, degree modifiers, contrast, capitalization and punctuation are unchanged; *vaderSentiment* 3.3.2 was used.

3.4. Automatic tier

An optional fourth tier adds frequent review words selected by a fixed rule. The candidates are the 30,000 most frequent tokens in a random sample of 300,000 reviews; this sample uses frequencies only, not labels. Valence is computed exactly as for the curated tiers. A token enters the tier if it meets all of the following conditions:

- it consists of letters, with hyphens allowed inside;
- it is not a function word, that is, not in the English stop list of *scikit-learn* (Pedregosa et al., 2011), not one of VADER's negations or degree modifiers, and not a contraction with n't;
- it is at least three characters long;
- it is not a probable proper name, that is, fewer than half of its mid-sentence occurrences are capitalised (checked when at least 20 such occurrences were observed);
- it is absent from VADER and from the curated tiers;
- it occurs in at least 100 training reviews;
- its $|z|$ is at least 5 and its $|valence|$ at least 0.5.

The thresholds are constants of the build script and were not tuned on the test data. The word-form condition was added after an exploratory analysis of the test sample had shown that the candidate pool contained list markers, quotation artefacts and censored spellings (for example *-fun*, *'game'*, *f*ck*); its effect on the evaluation was not examined before the final run.

3.5. Comparison lexicons

Nine scoring variants were compared:

- VADER alone;
- VADER extended cumulatively with the core tier, with core and extended, and with all three curated tiers (GamerLex curated);
- VADER with the automatic tier only;
- VADER with all four tiers (GamerLex full);
- GamerLex full without VADER's vocabulary;
- two uncurated baselines.

The uncurated baselines take the 30,000 most frequent tokens of the label-free sample, compute valence with the procedure of Section 3.3 and keep every token with $|z| \geq 3$ and at least five training reviews. The first baseline adds only the 12,414 tokens that VADER does not contain (induced, new words). The second adds all 14,429 and also replaces VADER's values for the 2,015 tokens that both contain (induced, all frequent words). Unlike the automatic tier, neither baseline excludes function words, digits or proper names. The baselines show what corpus-derived valence achieves without curation.

3.6. Evaluation

A random sample of 99,389 reviews from the test part was scored with each variant. The VADER compound score was used to rank reviews, with the recommendation as the target. The ROC AUC and a length-stratified AUC were reported, the latter being the AUC within each length band, averaged with weights equal to the number of recommending and non-recommending pairs in the band. The balanced accuracy was also reported, treating a score of zero or above as a recommendation, together with the coverage, defined as the share of reviews with a non-zero score. Differences between variants come with 95% percentile intervals from 200 paired bootstrap resamples of the test sample. For each genre with at least 200 test reviews (nine of the eleven TLGs; Experimental and Tabletop fell below this threshold), the stratified AUC was computed. Across genres, the Spearman correlation between the recommending share and the mean score was then computed.

The benchmark sentences were drawn from a random pool of 30,000 reviews of the deduplicated corpus. Reviews were split into sentences at sentence-final punctuation, and sentences of 6 to 60 whitespace-separated tokens that contained at least one term of the 3,880-term candidate list of Section 3.2 were kept. From these, 525 sentences of recommending reviews, 525 of non-recommending reviews and 450 further sentences regardless of the label were drawn at random (seed 42), 1,500 in total; 846 come from recommending reviews. The candidate list still contained the function word the, so 507 of the evaluated sentences entered the sample only through this token and were in effect selected without regard to gaming vocabulary. Two annotators with gaming experience rated the sentiment of each sentence on a five-point scale from − 2 to + 2 following a written guideline; they also marked slang tokens, which are not analyzed in this study. Their annotation sheet showed whether the review recommended the game. Nineteen sentences were unreadable after a spreadsheet conversion error, and

one annotator left nineteen further sentences unrated, which leaves 1,462 sentences with two ratings. The author rated all 1,481 readable sentences as a third rater, blind to the review label, to the annotators' ratings and to the lexicon scores, in a shuffled order.

Inter-rater agreement is reported as pairwise Spearman correlation, exact agreement, agreement within one point, and Krippendorff's alpha on interval and ordinal scales (Krippendorff, 2004). The primary gold standard is the median of the three ratings. The mean of the two annotators and the blind third rating alone are also reported. For illustration, six example sentences were drawn at random (seed 2026) from sentences of at most 20 words: four from those in which GamerLex corrected the sign of VADER's score, and two from those in which it reversed a correct VADER score. Each variant is evaluated by the Spearman correlation between its compound score and the gold standard, and by its sign accuracy, defined as the share of sentences with a non-zero gold rating that receive a score of the same sign. Differences between variants come with 95% percentile intervals from paired bootstrap resampling of sentences: 1,000 resamples in the main comparison against the median of three raters, and 2,000 in the comparisons against the blind rater and in the robustness checks.

GamerLex valences were learned from recommendations, and two raters saw the recommendation while rating. The agreement between a lexicon and these raters could therefore partly run through the label. Two checks were carried out to assess this. First, the Spearman correlation was computed within recommending and within non-recommending reviews, and the two values were averaged with weights equal to group size, using bootstrap resampling within each group. Second, the evaluation was repeated with the blind third rater as the only gold standard. To check for selection effects, the 507 sentences that entered the benchmark only through the token "the" were also evaluated. Finally, because the number of benchmark sentences containing each candidate was visible during the audit (Section 3.2), the evaluation was repeated on the 889 sentences that contain none of the 1,719 curated terms added in the audit. All checks use 2,000 bootstrap resamples.

3.7. Implementation and availability

The lexicon is built by one script that performs the split, the estimation, the filtering, the calibration and the evaluation, and that writes the release files together with a machine-readable record of every reported figure. The release contains the lexicon in CSV, JSON and VADER format, the corpus statistics of every entry, the anchor words, the entries removed by the filters with the reason, a helper module and the build report (see Data availability). Review texts are not redistributed.

4. Results

4.1. The lexicon

GamerLex v1.0 contains 8,046 entries: 4,223 in the three curated tiers and 3,823 in the automatic tier (Table 1). Of the 4,344 audited candidates, the build filters removed 121. Of these, 85 occurred in fewer than five training reviews, 19 could not be matched by the reference tokenizer, and 17 override

candidates, among them *killed*, *destroyed* and *ridiculous*, showed a clearly negative association in the corpus and keep their VADER value. The median entry occurs in 436.5 training reviews. Of the curated entries, 292 also occur in VADER, and for 178 of them the GamerLex valence has the opposite sign.

Table 1
Composition of GamerLex v1.0.

Tier	Entries	Multi-word	Valence ≤ -0.5	Valence ≥ + 0.5	Median training reviews	Examples (valence)
core	3,593	281	865	754	124	*abandonware* (-3.79), *unplayable* (-3.18), *banger* (+ 2.67), *goated* (+ 2.51)
extended	436	0	210	89	77.5	*unfinished* (-2.19), *mediocre* (-1.81), *catchy* (+ 1.68), *quirky* (+ 1.12)
vader_override	194	0	1	77	1,859	*kill* (-0.29), *brutal* (+ 0.99), *terrifying* (+ 1.92), *horror* (+ 0.86)
auto	3,823	0	1,695	2,128	826	*uninstalled* (-3.11), *money* (-1.34), *unique* (+ 1.45), *underrated* (+ 3.63)
Total	8,046	281	2,771	3,048	436.5	-

The calibration used 1,449 anchor words. On these words the corpus measure and VADER agreed in sign for 72.9% (Pearson 0.535, Spearman 0.542). The fitted slopes were 1.293 for a single mapping and 1.275 and 1.320 for positive and negative log odds; the separate slopes gave a slightly higher validation AUC (0.7805 against 0.7796) and were used.

In the override tier, the mean VADER valence is -2.00 and the mean GamerLex valence + 0.37, and the sign differs for 72.7% of the terms. *Kill* moves from − 3.7 in VADER to -0.29, *brutal* from − 3.1 to + 0.99 and *terrifying* from − 2.7 to + 1.92.

The automatic tier contains 1,695 negative and 2,128 positive words. Its most frequent entries are aspect and discourse words such as *just* (-0.76), *story* (+ 0.75), *buy* (-0.68), *music* (+ 0.93) and *money* (-1.34). Its strongest entries include genuinely evaluative words missing from VADER (*underrated*, *must-play*, *addictive*, *unforgettable*), but also words whose valence comes from fixed expressions rather than from their own meaning, such as *requested* (-3.10, as in a requested refund), *steaming* (-2.90) and *straw* (-2.69).

4.2. Review-level evaluation

Table 2 reports the review-level results; 62.8% of the test reviews contain at least one curated term and 81.2% a term of any tier. Each curated tier added to VADER improved the length-stratified AUC, from 0.8245 for VADER to 0.8674 with all three curated tiers. The override tier lowered the unstratified AUC (0.7961 to 0.7856) while raising the stratified AUC (0.8538 to 0.8674). This pattern is consistent with the length effect of Section 3.3: descriptive words are frequent, frequent words are more likely to occur in long reviews, and long reviews recommend less often. With all four tiers the stratified AUC reached 0.8930. GamerLex without VADER's vocabulary reached a stratified AUC of 0.7933, below VADER alone, so the lexicon works as an extension of VADER rather than on its own.

Table 2
Review-level results on 99,389 held-out reviews. *Δ* is the difference in length-stratified AUC from VADER with 95% bootstrap interval.

Variant	**Added entries**	**AUC**	**Stratified AUC**	***Δ* stratified AUC**	**Balanced accuracy**	**Coverage**
VADER	0	0.7495	0.8245	-	0.6974	0.8952
+ *core*	3,593	0.7912	0.8509	+ 0.0264 [+ 0.0249, + 0.0276]	0.7438	0.9217
+ *core* + *extended*	4,029	0.7961	0.8538	+ 0.0293 [+ 0.0278, + 0.0307]	0.7484	0.9228
+ GamerLex curated	4,223	0.7856	0.8674	+ 0.0430 [+ 0.0414, + 0.0446]	0.7385	0.9230
+ automatic tier only	3,823	0.8088	0.8656	+ 0.0412 [+ 0.0396, + 0.0427]	0.7520	0.9365
+ GamerLex full	8,046	0.8292	0.8930	+ 0.0685 [+ 0.0665, + 0.0706]	0.7789	0.9521
GamerLex full only	8,046	0.8003	0.7933	-0.0311 [-0.0361, -0.0257]	0.7468	0.8085
+ induced (new words)	12,414	0.8594	0.8915	+ 0.0671 [+ 0.0641, + 0.0697]	0.8069	0.9769
+ induced (all frequent words)	14,429	0.8985	0.9137	+ 0.0892 [+ 0.0857, + 0.0925]	0.8378	0.9767

The uncurated baselines were strong at this level. GamerLex full matched the baseline restricted to new words in stratified AUC (difference + 0.0015 [-0.0000, + 0.0032]) with fewer entries, and it remained below the baseline that also re-scores the words GamerLex shares with VADER (-0.0207 [-0.0227,

-0.0189]). Both differences exceed the corresponding unstratified differences (-0.0302 and − 0.0693), so part of the baselines' advantage comes from review length. Within GamerLex, the curated and the automatic tiers were complementary: the full lexicon exceeded both the curated tiers (+ 0.0256 [+ 0.0244, + 0.0269]) and the automatic tier alone (+ 0.0274 [+ 0.0260, + 0.0287]).

4.3. Genres

Adding GamerLex raised the stratified AUC in each of the nine genres with at least 200 test reviews (Table 3), from 0.8190–0.8852 with VADER to 0.8888–0.9432 with the full lexicon. Between genres the picture is different. The Spearman correlation between a genre's recommending share and its mean score was + 0.700 with VADER and + 0.583 with GamerLex. With nine genres this difference carries little information, but it gives no sign that GamerLex makes genre-level averages more faithful to recommendations.

Table 3
Length-stratified AUC by genre. Experimental (173 test reviews) and Tabletop (25 test reviews) are not shown because they fall below the threshold of 200 test reviews; the nine genres shown cover 99,191 of the 99,389 test reviews.

Genre	Test reviews	Recommending share	VADER	+ GamerLex curated	+ GamerLex full
Action	60,008	0.808	0.8207	0.8669	0.8908
Adventure	17,634	0.834	0.8275	0.8613	0.8928
Casual	7,237	0.849	0.8266	0.8615	0.8888
Simulation	6,935	0.790	0.8291	0.8762	0.8993
RPG	3,027	0.786	0.8255	0.8721	0.9013
Strategy	2,191	0.772	0.8190	0.8624	0.8956
Racing	1,049	0.802	0.8628	0.8833	0.9049
Puzzle	860	0.855	0.8461	0.8783	0.8991
Sports	250	0.828	0.8852	0.9229	0.9432

4.4. Sentence-level evaluation

Agreement between the two annotators was low: Spearman 0.543, exact agreement 28.5%, agreement within one point 73.4%, and Krippendorff's alpha 0.452 on the interval scale. The third rater agreed more closely with each of them (Spearman 0.727 with the first and 0.646 with the second annotator), and alpha for the three raters was 0.583. This is below the level of 0.667 that Krippendorff (2004) considers the minimum for tentative conclusions. Therefore, the sentence-level results are treated as indicative. The median of the three ratings serves as the primary gold standard, and the blind third rating is reported separately.

Table 4
Spearman correlation with human ratings on 1,462 benchmark sentences.

Variant	Median of three raters	Mean of two annotators	Third rater (blind)	Sign accuracy (median)	Coverage
VADER	0.5787	0.5706	0.5380	0.7256	0.8454
+ *core*	0.6176	0.6149	0.5720	0.7980	0.9371
+ *core* + *extended*	0.6241	0.6199	0.5764	0.8090	0.9405
+ GamerLex curated	0.6455	0.6392	0.6048	0.8035	0.9405
+ automatic tier only	0.6225	0.6152	0.5874	0.7947	0.9425
+ GamerLex full	0.6638	0.6604	0.6290	0.8364	0.9795
GamerLex full only	0.4073	0.4158	0.4045	0.6509	0.9008
+ induced (new words)	0.6338	0.6257	0.6075	0.8364	0.9993
+ induced (all frequent words)	0.6197	0.6249	0.6028	0.8430	0.9993

GamerLex full had the highest correlation with all three gold standards (Table 4). Against the median of three raters it exceeded VADER by 0.085 [0.058, 0.112], the baseline with new words by 0.030 [0.009, 0.051], the baseline with all frequent words by 0.044 [0.016, 0.074], and the curated tiers by 0.018 [0.003, 0.034]. The curated tiers alone did not differ clearly from the baselines (+ 0.012 [-0.015, + 0.038] and + 0.026 [-0.008, + 0.060]). With the blind third rater as the gold standard (2,000 resamples), the advantage of GamerLex full over the baselines was smaller and its interval reached zero (+ 0.022 [-0.000, + 0.044] and + 0.026 [-0.003, + 0.055]).

Sign accuracy tells a different story from correlation: GamerLex full and the baseline with new words reached the same sign accuracy (0.836) on the 911 sentences with a non-zero median rating, and the baseline with all frequent words was slightly higher (0.843). The advantage of GamerLex therefore lies in grading the strength of sentiment rather than in its direction. Of the 911 sentences, VADER had the opposite sign to the raters while GamerLex full agreed with them in 78 cases; the reverse occurred in 41.

Table 7 shows six of these sentences, drawn at random as described in Section 3.6. In the corrections, curated terms (*sub-par*, *crashes*) and words of the automatic tier (*story*, *minutes*, *flat*, *memories*) supply evaluative weight that VADER lacks. The two failures show typical mechanisms. In the first, the override tier removes VADER's negative value for combat, so VADER's positive value for joke decides the score of a sentence that all three raters rated − 2. In the second, VADER's negation rule turns words that GamerLex rates as negative (rhyme − 1.55, reason − 1.05) into positive evidence after no.

4.5. Robustness

The first annotator's ratings were strongly correlated with the review label that the annotation sheet displayed (Spearman 0.750), the second annotator's and the blind third rater's less so (0.347 and 0.477). Within label groups, GamerLex full still exceeded VADER, by 0.050 [0.020, 0.078] against the median of three raters and by 0.067 [0.034, 0.099] against the blind rater, and the baseline with all frequent words. Against the baseline with new words the within-label advantage was significant for the median of three raters (+ 0.031 [+ 0.006, + 0.056]) but not for the blind rater (+ 0.023 [-0.002, + 0.048]). Within label groups the automatic tier added nothing over the curated tiers (-0.008 [-0.026, + 0.010] and + 0.004 [-0.017, + 0.024]); its contribution to the overall correlation thus runs largely through the review label.

In the 507 sentences selected without regard to gaming vocabulary, GamerLex full reached Spearman correlations of 0.679 with the median of three raters and 0.656 with the blind rater, against 0.610 and 0.573 for VADER. Here it exceeded VADER, both baselines and the curated tiers alone, with all intervals above zero (Table 5).

Table 5
Robustness checks: difference in Spearman correlation between GamerLex full and the comparison variant, with 95% bootstrap interval (2,000 resamples).

Comparison	Within label, median of three	Within label, blind rater	Unselected subset, median of three	Unselected subset, blind rater
VADER	+ 0.050 [+ 0.020, + 0.078]	+ 0.067 [+ 0.034, + 0.099]	+ 0.069 [+ 0.032, + 0.106]	+ 0.083 [+ 0.046, + 0.123]
Induced (new words)	+ 0.031 [+ 0.006, + 0.056]	+ 0.023 [-0.002, + 0.048]	+ 0.044 [+ 0.010, + 0.079]	+ 0.043 [+ 0.007, + 0.079]
Induced (all frequent words)	+ 0.075 [+ 0.043, + 0.110]	+ 0.050 [+ 0.018, + 0.082]	+ 0.079 [+ 0.026, + 0.135]	+ 0.057 [+ 0.002, + 0.107]
GamerLex curated	-0.008 [-0.026, + 0.010]	+ 0.004 [-0.017, + 0.024]	+ 0.041 [+ 0.007, + 0.074]	+ 0.048 [+ 0.016, + 0.082]

The 889 sentences that contain no term added during the audit gave the same picture (Table 6). GamerLex full exceeded VADER and both uncurated baselines with both gold standards, including the blind rater (+ 0.036 [+ 0.010, + 0.062] against the baseline with new words), and the curated tiers alone still exceeded VADER (+ 0.057 with both gold standards). The audit, which had the benchmark counts in view, therefore does not explain the advantage of GamerLex.

Table 6
Benchmark sentences without terms added in the audit (n = 889): difference in Spearman correlation between the first and the second variant, with 95% bootstrap interval (2,000 resamples).

Comparison	Median of three raters	Third rater (blind)
GamerLex full minus VADER	+ 0.069 [+ 0.038, + 0.099]	+ 0.071 [+ 0.040, + 0.103]
GamerLex full minus induced (new words)	+ 0.043 [+ 0.018, + 0.070]	+ 0.036 [+ 0.010, + 0.062]
GamerLex full minus induced (all frequent words)	+ 0.060 [+ 0.024, + 0.094]	+ 0.047 [+ 0.011, + 0.085]
GamerLex full minus GamerLex curated	+ 0.012 [-0.012, + 0.035]	+ 0.014 [-0.010, + 0.036]
GamerLex curated minus VADER	+ 0.057 [+ 0.035, + 0.079]	+ 0.057 [+ 0.035, + 0.081]

Table 7
Example benchmark sentences with VADER and VADER + GamerLex full compound scores and the median of three human ratings (-2 to + 2). The first four are random draws from the 78 sentences where GamerLex corrected the sign of VADER's score, the last two from the 41 where it reversed a correct VADER score.

Sentence	Human (median)	VADER	+ GamerLex full	GamerLex terms (valence)
An alright idea for an adventure game, but a sub-par implementation of one.	-1	0.28	-0.60	*adventure game* (+ 1.10), *idea* (-0.99), *sub-par* (-1.56), *implementation* (-0.76)
the story seems interesting, but the game crashes every few minutes...	-1	0.21	-0.66	*story* (+ 0.75), *crashes* (-1.91), *minutes* (-1.18)
However, most of the jokes just fall flat.	-2	0.31	-0.18	*just* (-0.76), *flat* (-1.23)
Brings memories of games like Big Fight, Legionnaire, 64th Street, Knuckle Bash, Silent Dragon etc.	+ 1	-0.03	0.69	*brings* (+ 1.03), *memories* (+ 1.23), *fight* (-0.09)
this game's combat is kind of a joke.	-2	-0.05	0.32	*combat* (+ 0.11; VADER − 1.4)
There's no rhyme or reason behind controlling the trocar.	-1	-0.30	0.44	*rhyme* (-1.55), *reason* (-1.05)

5. Discussion

5.1. GamerLex as an extension of VADER

Adding GamerLex to VADER improved both examined tasks. The length-stratified AUC for ranking reviews by recommendation rose from 0.825 to 0.893, and the Spearman correlation with the median human rating of single sentences rose from 0.579 to 0.664. Used on its own, GamerLex was weaker than VADER on the length-stratified AUC (0.793 against 0.825) and on human ratings (Spearman 0.407 against 0.579). It supplements VADER's general vocabulary and rules; it does not replace them. The released helper therefore loads the entries into VADER, and scoring text with the GamerLex entries alone is discouraged.

5.2. What curation adds

The comparison with uncurated lexicons built by the same procedure qualifies these gains. At review level, a lexicon of the 30,000 most frequent tokens predicted recommendations better than GamerLex (stratified AUC 0.914 against 0.893). That baseline also re-scores 2,015 words that VADER already contains, and it may give valence to any frequent token that happens to co-occur with positive or negative reviews. Predicting the label from which the valences were estimated is exactly the task such a lexicon is built for.

Against human ratings of individual sentences the order reversed. GamerLex with all four tiers exceeded both baselines, by 0.030 and 0.044 against the median of three raters, while the curated tiers alone did not differ clearly from them. The gain thus comes from combining curated slang and jargon with a rule-selected layer of frequent content words. Sign accuracy adds a second qualification: GamerLex and the baseline with new words gave the correct direction for the same share of sentences (0.836). Where GamerLex does better is in how strongly it scores one sentence relative to another.

Part of the value of curation lies outside these metrics. Only 1,463 of the 4,223 curated entries (34.6%) appear in the larger uncurated lexicon; the others are multi-word terms, terms too rare for its frequency cut-off, or terms without a clear association with recommendations. The median core entry occurs in 124 training reviews, against 826 for the automatic tier. Curated entries are also interpretable. Each is a term a reader of game reviews would recognise, and its tier states why it is in the lexicon. A researcher who studies how players write about monetisation (*cashgrab*, *p2w*) or difficulty can inspect, filter and extend these entries. The automatic tier lacks this property, which is why it is released as a separate, optional layer.

5.3. Valence from recommendations

Deriving valence from recommendations replaces human ratings of each term with a measure of the company the term keeps. This scales to thousands of terms and yields values for slang that raters outside the community might not know. It has two consequences that users should keep in mind.

First, the valence reflects how a word is used in reviews rather than what it means. The automatic tier contains *requested* (-3.10), *steaming* (-2.90) and *straw* (-2.69) because these words occur in fixed

expressions of complaint, and frequent aspect words such as *story* (+ 0.75) or *money* (-1.34) carry the average verdict of the reviews that mention them.

Second, an evaluation against human ratings can partly reproduce the label. In our benchmark the two annotators saw the recommendation, and the ratings of one of them followed it closely (Spearman 0.750). Within label groups, GamerLex still exceeded VADER (by 0.050 against the median of three raters and by 0.067 against the blind rater), but the automatic tier no longer added anything to the curated tiers. Its contribution at sentence level thus runs largely through the review label, while the contribution of the curated tiers does not. For studies that measure the sentiment of text separately from the reviewer's verdict, the curated tiers are the safer choice; for studies that use sentiment as a proxy for the verdict, the full lexicon fits better.

Review length had to be taken into account. The share of recommending reviews fell from 90.2% among the shortest reviews to 69.7% among the longest, and a frequent word is more likely to occur in a long review. Without stratification the override tier appeared to harm prediction (AUC 0.796 to 0.786); within length bands it helped (0.854 to 0.867). An evaluation of review-level sentiment tools on Steam that ignores review length can therefore reach the wrong conclusion about which vocabulary helps.

5.4. Implications for research on player reviews

Within every genre with enough test reviews, GamerLex improved the ranking of reviews. Mean scores per genre, however, moved in different directions. With GamerLex the mean score rose in Action (0.388 to 0.422) and Adventure (0.498 to 0.532) and fell in Racing (0.439 to 0.401), Simulation (0.413 to 0.383) and Sports (0.470 to 0.435). Each genre has its own vocabulary of mechanics and content, and a general lexicon reads part of that vocabulary as emotion. These genre groups are not pure: each review carries the alphabetically first genre of its game, so the Action group contains every game also tagged as Action, and the other groups contain only games without an alphabetically earlier genre. Comparisons of mean sentiment between genres therefore depend on the lexicon. Neither lexicon tracked the recommending share of genres well (Spearman 0.700 with VADER and 0.583 with GamerLex over nine genres), so mean lexicon scores should not be used to rank genres by player satisfaction without a check against the recommendations themselves. Comparisons within genres, or comparisons that adjust for genre vocabulary, are safer.

The descriptive vocabulary of games raises the same issue at the level of single words. VADER rates *kill*, *brutal* and *terrifying* as strongly negative, and such words are common in reviews of shooters and horror games. When reviews of different kinds of games are compared with VADER, violence and fear in the game's content can be recorded as negativity in the player's evaluation. The override tier makes this distinction explicit, and because it can be switched on and off, researchers can test how much a conclusion depends on it.

5.5. Limitations

The lexicon was built from Steam reviews only, from the collection described by Guzsvinecz and Szűcs (2023), and gaming slang changes quickly; the valences describe use in this corpus and may drift. The corpus has no game identifier, so the random split placed reviews of the same games in the training and the test part. Associations between a term and a particular game can therefore carry over, and the review-level results may be somewhat optimistic. The genre analysis rests on one genre label per review, the alphabetically first genre of the game, so it does not capture games that belong to several genres and overrepresents Action (60,008 of the 99,389 test reviews). The same rule leaves Experimental and Tabletop with too few test reviews (173 and 25) for a genre-level estimate, so the lexicon has not been evaluated separately on these two genres.

The benchmark has three weaknesses. Agreement between raters was moderate (Krippendorff's alpha 0.583 with three raters), below the level usually required for firm conclusions. The sentences were sampled with an early candidate list, which favours sentences that contain gaming vocabulary; the 507 sentences that entered through the token the reduce but do not remove this concern. The third rater is the author, who rated blind to labels and lexicon scores but had worked with the lexicon before rating.

GamerLex assigns one valence per word form. Negation, sense ambiguity and irony are left to VADER's rules, so not scary and sick of can still be misread, and only 281 entries are multi-word terms. The automatic tier was not reviewed by hand, and its word-form rule was added after an exploratory look at the test sample. During the audit, corpus statistics computed on all reviews and the number of benchmark sentences containing each candidate were visible. Valences and the build filters use the training part only, but the choice of which curated terms to keep may have been influenced, which could favor the curated tiers on the benchmark; the automatic tier and the uncurated baselines are free of this influence. On the 889 benchmark sentences that contain no term added in the audit the results did not change (Table 6), but decisions about the original terms cannot be checked in the same way. Labelling and audit decisions were reviewed by the author.

Finally, GamerLex was compared with VADER and with lexicons induced from the same data, not with supervised classifiers. A fine-tuned transformer will likely predict recommendations better; GamerLex is meant for analyses that need a transparent, word-level score that can be inspected and adapted.

Future work can extend the procedure to other platforms and languages, add sense-specific entries for words such as sick, and track how the valence of slang changes over time by repeating the estimation on reviews from different periods.

6. Conclusions

GamerLex is a sentiment lexicon of gaming slang and game jargon for player reviews. It contains 4,223 hand-audited entries in three tiers and 3,823 rule-selected review words, all with valence on the VADER scale derived from a corpus of 4.69 million Steam reviews. Valence comes from the association between a term and players' recommendations, adjusted for review length and calibrated to VADER through anchor words. Added to VADER, the lexicon improved the ranking of held-out reviews by

recommendation and the agreement with human ratings of review sentences, and its advantage over VADER held when the review label was controlled for and in a subset of sentences selected without regard to gaming vocabulary.

Two findings matter for anyone who builds or uses such resources. An automatic lexicon induced from the same labels is a strong competitor for predicting those labels, so a domain lexicon should be evaluated against human judgements and against uncurated baselines, not only against a general-purpose lexicon. Also, review length confounds word-level associations with recommendations on Steam strongly enough to reverse the apparent effect of a vocabulary layer.

CRediT authorship contribution statement

Tibor Guzsvinecz: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration.

Declarations

Declaration of competing interest

The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This research has received no funding.

Ethical approval

Review and/or approval by an ethics committee was not required for this study because it relied exclusively on publicly available, anonymized user-generated content obtained via the official Steam Web API. The research did not involve any interaction with human subjects, collection of personal data, or interventions of any kind.

Informed consent statement

Not applicable as the research did not involve any interaction with human subjects.

Data availability

For review, GamerLex v1.0 and the build and evaluation scripts are available at https://github.com/zyzyx02/GamerLex. On acceptance, the release will be archived with a DOI.

Use of generative AI

During the preparation of this manuscript, the author used Claude Opus 5 and Claude Opus 5.5 (Anthropic) to assist with language and structural refinement, including improvements to readability, clarity, coherence, logical flow, and the organization of the manuscript. The AI tools were used solely in an editorial capacity and did not generate research data, scientific results, or substantive interpretations. All AI-assisted suggestions and revisions were critically reviewed and edited by the author, who retains full responsibility for the final content and conclusions of the manuscript.

Author Contribution

Tibor Guzsvinecz: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration.

Data Availability

For review, GamerLex v1.0 and the build and evaluation scripts are available at https://github.com/zyzyx02/GamerLex. On acceptance, the release will be archived with a DOI.

References

1. Andreev, G., Saxena, D. & Verma, J. K. Impact of review sentiment and magnitude on customers' recommendations for video games. 2021 International Conference on Computational Performance Evaluation (ComPE), 992–995. (2021). https://doi.org/10.1109/ComPE53109.2021.9752380
2. Babanejad, N., Agrawal, A., An, A. & Papagelis, M. A comprehensive analysis of preprocessing for word representation learning in affective tasks. Proceedings of ACL 2020, 5799–5810. (2020). https://doi.org/10.18653/v1/2020.acl-main.514
3. Baccianella, S., Esuli, A. & Sebastiani, F. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. Proceedings of LREC-10, 2200–2204. (2010).
4. Bergsma, T., van Stegeren, J. & Theune, M. Creating a sentiment lexicon with game-specific words for analyzing NPC dialogue in The Elder Scrolls V: Skyrim. In Proceedings of the Workshop on Games and Natural Language Processing (pp. 1–9). European Language Resources Association. https://aclanthology.org/2020.gamnlp-1.1/ (2020).
5. Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. Enriching word vectors with subword information. *Trans. Association Comput. Linguistics*. **5**, 135–146. https://doi.org/10.1162/tacl_a_00051 (2017).
6. Bradley, M. M. & Lang, P. J. Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical Report C-1, University of Florida. (1999).

7. De Luisa, A. et al. Predicting the popularity of games on Steam. arXiv preprint arXiv:2110.02896. (2021).
8. Dolin, P., d'Hauthuille, L. & Vattani, A. FeelsGoodMan: Inferring semantics of Twitch neologisms. arXiv:2108.08411. (2021). https://arxiv.org/abs/2108.08411
9. Eberhard, L., Kasper, P., Koncar, P. & Gütl, C. Investigating helpfulness of video game reviews on the Steam platform. 2018 5th International Conference on Social Networks Analysis, Management and Security (SNAMS), 43–50. (2018). https://doi.org/10.1109/SNAMS.2018.8554542
10. Eble, C. *Slang and sociability: In-group language among college students* (University of North Carolina Press, 1996).
11. Eisenstein, J. What to do about bad language on the internet. Proceedings of NAACL-HLT 2013, 359–369. (2013).
12. Guzsvinecz, T. Sentiment analysis of Souls-like role-playing video game reviews. *Acta Polytech. Hungarica*. **20** (5), 149–168. https://doi.org/10.12700/APH.20.5.2023.5.10 (2023).
13. Guzsvinecz, T. & Szűcs, J. Length and sentiment analysis of reviews about top-level video game genres on the Steam platform. *Comput. Hum. Behav.* **149**, 107955. https://doi.org/10.1016/j.chb.2023.107955 (2023).
14. Hamilton, W. L., Clark, K., Leskovec, J. & Jurafsky, D. Inducing domain-specific sentiment lexicons from unlabeled corpora. Proceedings of EMNLP 2016, 595–605. (2016). https://doi.org/10.18653/v1/D16-1057
15. Hu, M. & Liu, B. Mining and summarizing customer reviews. Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '04), 168–177. (2004). https://doi.org/10.1145/1014052.1014073
16. Hutto, C. J. & Gilbert, E. VADER: A parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the International AAAI Conference on Web and Social Media, 8(1), 216–225. (2014). https://doi.org/10.1609/icwsm.v8i1.14550
17. Islam, S. M. M., Dong, X. & de Melo, G. Domain-specific sentiment lexicons induced from labeled documents. In Proceedings of the 28th International Conference on Computational Linguistics (pp. 6576–6587). (2020). https://doi.org/10.18653/v1/2020.coling-main.578
18. Ismail, H., Khalil, A. & Jasmy, A. Enhancing online toxicity detection on gaming networks: A novel embeddings-based valence lexicon approach. *Int. J. Data Sci. Analytics*. **20** (5), 4489–4500. https://doi.org/10.1007/s41060-025-00730-1 (2025).
19. Jia, Y., Wu, W., Cao, F. & Han, S. C. In-game toxic language detection: Shared task and attention residuals (Student abstract). Proceedings of the AAAI Conference on Artificial Intelligence, 37(13), 16238–16239. (2023). https://doi.org/10.1609/aaai.v37i13.26979
20. 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. Manage. Technol.*, **8**(2), 90–94. https://doi.org/10.18178/ijimt.2017.8.2.709 (2017).

21. Kobs, K. et al. Emote-controlled: Obtaining implicit viewer feedback through emote-based sentiment analysis on comments of popular Twitch.tv channels. *ACM Trans. Social Comput.* **3** (2). https://doi.org/10.1145/3365523 (2020). Article 7.
22. Krippendorff, K. *Content Analysis: An Introduction to Its Methodology* 2nd edn (Sage, 2004).
23. Kwak, H. & Blackburn, J. Linguistic analysis of toxic behavior in an online video game. In Social Informatics (SocInfo 2014 International Workshops), Lecture Notes in Computer Science (pp. 209–217). Springer. (2015). https://doi.org/10.1007/978-3-319-15168-7_26
24. Lin, D., Bezemer, C. P., Zou, Y. & Hassan, A. E. An empirical study of game reviews on the Steam platform. *Empir. Softw. Eng.* **24** (1), 170–207. https://doi.org/10.1007/s10664-018-9627-4 (2019).
25. Maity, S. K., Ghuku, B., Upmanyu, A. & Mukherjee, A. Out of vocabulary words decrease, running texts prevail and hashtags coalesce: Twitter as an evolving sociolinguistic system. Proceedings of HICSS 2016, 1681–1690. (2016).
26. Mantel, N. & Haenszel, W. Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl Cancer Inst.* **22** (4), 719–748. https://doi.org/10.1093/jnci/22.4.719 (1959).
27. Marcondes, F. S., Gala, A. C., Rodrigues, M., Almeida, J. J. & Novais, P. Lexicon annotation with LLM: A proof of concept with ChatGPT. In Hybrid Artificial Intelligent Systems (HAIS 2024), Lecture Notes in Computer Science, Vol. 14858 (pp. 190–200). Springer. (2024). https://doi.org/10.1007/978-3-031-74186-9_16
28. Märtens, M., Shen, S., Iosup, A. & Kuipers, F. Toxicity detection in multiplayer online games. Proceedings of NetGames 2015, 1–6. (2015). https://doi.org/10.1109/NetGames.2015.7382991
29. Mohammad, S. M. & Turney, P. D. Crowdsourcing a word-emotion association lexicon. *Comput. Intell.* **29** (3), 436–465 (2013).
30. Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
31. Pennebaker, J. W., Francis, M. E. & Booth, R. J. *Linguistic Inquiry and Word Count: LIWC 2001* (Erlbaum, 2001).
32. Sirbu, D. et al. Extracting gamers' opinions from reviews. Proceedings of SYNASC 2016, 227–232. (2016). https://doi.org/10.1109/SYNASC.2016.044
33. SteamDB. Steam game releases by year. Retrieved October 9, 2026, from (2026). https://steamdb.info/stats/releases/
34. Stone, P. J., Dunphy, D. C., Smith, M. S. & Ogilvie, D. M. *The General Inquirer: A computer approach to content analysis* (MIT Press, 1966).
35. Sun, Z., Zemel, R. & Xu, Y. A computational framework for slang generation. *Trans. Association Comput. Linguistics*. **9**, 462–478. https://doi.org/10.1162/tacl_a_00378 (2021).
36. Tereshchenko, Y. & Hämäläinen, M. Efficient toxicity detection in gaming chats: A comparative study of embeddings, fine-tuned transformers and LLMs. arXiv:2510.17924. (2025). https://arxiv.org/abs/2510.17924

37. therohk. (n.d.). Urban Dictionary Words And Definitions [Data set]. Kaggle. Retrieved. from May (2026). https://www.kaggle.com/datasets/therohk/urban-dictionary-words-dataset
38. Turney, P. D. & Littman, M. L. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inform. Syst.* **21** (4), 315–346. https://doi.org/10.1145/944012.944013 (2003).
39. Urriza, I. M. & Clariño, M. A. A. Aspect-based sentiment analysis of user created game reviews. Proceedings of O-COCOSDA 2021, 76–81. (2021). https://doi.org/10.1109/O-COCOSDA202152914.2021.9660559
40. Velikovich, L., Blair-Goldensohn, S., Hannan, K. & McDonald, R. The viability of web-derived polarity lexicons. Proceedings of NAACL HLT 2010, 777–785. (2010).
41. Viggiato, M., Lin, D., Hindle, A. & Bezemer, C. P. What causes wrong sentiment classifications of game reviews? *IEEE Trans. Games*. **14** (3), 350–363. https://doi.org/10.1109/TG.2021.3072545 (2022).
42. Wang, Z., Chang, V. & Horvath, G. Explaining and predicting helpfulness and funniness of online reviews on the Steam platform. *J. Global Inform. Manage.* **29** (6), 1–22. https://doi.org/10.4018/JGIM.20211101.oa16 (2021).
43. Weld, H. et al. CONDA: A CONtextual Dual-Annotated dataset for in-game toxicity understanding and detection. Proceedings of ACL-IJCNLP 2021 (Findings), 2406–2416. (2021). https://doi.org/10.18653/v1/2021.findings-acl.213
44. Wikipedia. Glossary of video game terms. Retrieved October 5, 2026, from (2026). https://en.wikipedia.org/wiki/Glossary_of_video_game_terms
45. Wu, L., Morstatter, F. & Liu, H. SlangSD: Building and using a sentiment dictionary of slang words for short-text sentiment classification. *arXiv preprint arXiv* :160805129. (2016).
46. Yang, Z., Maricar, Y., Davari, M., Grenon-Godbout, N. & Rabbany, R. ToxBuster: In-game chat toxicity buster with BERT. arXiv preprint arXiv:2305.12542. (2023).
47. Yuan, L., Ding, M., Meng, F. & Tian, Y. Sentiment analysis and rating video game dimensions via NLP. *Appl. Comput. Eng.* **132** (1), 1–10. https://doi.org/10.54254/2755-2721/2024.20540 (2025).