UNIVERSIDADE FEDERAL DE VIÇOSA

Defusing the Bomb: A Large-Scale Analysis and Machine Learning Approach to Review Bombing in Video Games on Steam

Marcus Vinicius Guerra Ribeiro
Magister Scientiae

VIÇOSA - MINAS GERAIS
2026

MARCUS VINICIUS GUERRA RIBEIRO

Defusing the Bomb: A Large-Scale Analysis and Machine Learning Approach to Review Bombing in Video Games on Steam

Dissertation submitted to the Computer Science Graduate Program of the Universidade Federal de Viçosa in partial fulfillment of the requirements for the degree of *Magister Scientiae*.

Adviser: Philipe de Freitas Melo

VIÇOSA - MINAS GERAIS
2026

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T

R484d
2026

Ribeiro, Marcus Vinicius Guerra, 2001-

Defusing the bomb: a large-scale analysis and machine learning approach to review bombing in video games on Steam / Marcus Vinicius Guerra Ribeiro. – Viçosa, MG, 2026.

1 dissertação eletrônica (89 f.): il. (algumas color.).

Texto em inglês.

Orientador: Philipe de Freitas Melo.

Dissertação (mestrado) - Universidade Federal de Viçosa, Departamento de Informática, 2026.

Referências bibliográficas: f. 83-89.

DOI: https://doi.org/10.47328/ufvbbt.2026.466

Modo de acesso: World Wide Web.

1. Interfaces de usuário (Sistemas de computação) - Aspectos psicológicos. 2. Jogadores de videogames - Aspectos sociológicos. 3. Discurso de ódio na Internet. 4. Processamento de linguagem natural (Computação). I. Melo, Philipe de Freitas, 1990-. II. Universidade Federal de Viçosa. Departamento de Informática. Programa de Pós-Graduação em Ciência da Computação. III. Título.

CDD 22. ed. 005.437019

Bibliotecário(a) responsável: Euzébio Luiz Pinto CRB-6/3317

MARCUS VINICIUS GUERRA RIBEIRO

Defusing the Bomb: A Large-Scale Analysis and Machine Learning Approach to Review Bombing in Video Games on Steam

Dissertation submitted to the Computer Science Graduate Program of the Universidade Federal de Viçosa in partial fulfillment of the requirements for the degree of *Magister Scientiae*.

APPROVED: July 9, 2026.

Assent:

_____

Marcus Vinicius Guerra Ribeiro
Author

_____

Philipe de Freitas Melo
Adviser

Gostaria de expressar minha sincera gratidão à minha família pelo apoio incondicional ao longo de todo o meu mestrado. Também sou profundamente grato ao meu orientador por acreditar nesta ideia de pesquisa e por oferecer o suporte acadêmico necessário para desenvolvê-la até se tornar esta dissertação.

ACKNOWLEDGMENTS

This work has been sponsored by the following Brazilian research agencies: Coordination for the Improvement of Higher Education Personnel (CAPES; Financing code 001), Minas Gerais State Foundation for Research Aid (FAPEMIG) and National Council of Scientific and Technological Development (CNPq).

ABSTRACT

RIBEIRO, Marcus Vinicius Guerra, M.Sc., Universidade Federal de Viçosa, July, 2026. **Defusing the Bomb: A Large-Scale Analysis and Machine Learning Approach to Review Bombing in Video Games on Steam**. Adviser: Philipe de Freitas Melo.

Steam is currently the largest digital distribution platform for PC games and hosts one of the most extensive publicly accessible ecosystems of user reviews. In this context, Review Bombing (RB) has emerged as an orchestrated form of collective action in which large volumes of reviews are posted within short periods of time with the intent of influencing a game's public evaluation. Despite its relevance, large scale empirical analyses of RB remain limited. In this work, we present a comprehensive longitudinal study of RB on Steam based on a dataset comprising 157,825 games and more than 103 million user reviews, validated against 89 RB incidents and 11,800,720 RB reviews identified by Steam's official moderation pipeline and subsequently confirmed through manual review. We analyze temporal patterns, user participation, and textual characteristics of RB incidents, as well as the thematic content of reviews through topic modeling. Our findings demonstrate that RB constitutes a distinct and empirically observable phenomenon, characterized by recurring temporal, behavioral, and linguistic patterns that significantly differentiate it from the organic flow of user reviews. We further show that these incidents are predominantly driven by established and highly engaged users and are frequently used as vehicles for expressing geopolitical, cultural, and ideological grievances, often framed through consumer oriented language such as refund requests and allegations of fraudulent practices. Moreover, we show that RB can be instrumentalized both to promote and to challenge hate speech, highlighting important tensions between consumer expression and platform governance. Finally, we develop and evaluate a machine learning approach for detecting RB incidents based on temporal, semantic, and behavioral features, achieving an F1 score of 0.67 using XGBoost. Our results indicate that semantic shifts in review content, temporal variations in reviewing activity, and the concentration of negative reviews are the most predictive signals for identifying RB incidents.

Keywords: review bombing; hate speech; Steam; natural language processing; topic modeling; gaming culture

RESUMO

RIBEIRO, Marcus Vinicius Guerra, M.Sc., Universidade Federal de Viçosa, julho de 2026. **Desarmando a Bomba: Uma Análise em Larga Escala e Abordagem de Aprendizado de Máquina para o Review Bombing em Jogos Digitais na Steam**. Orientador: Philipe de Freitas Melo.

A Steam é atualmente a maior plataforma de distribuição digital de jogos para PC e abriga um dos mais extensos ecossistemas de avaliações de usuários publicamente acessíveis. Nesse contexto, o Review Bombing (RB) emergiu como uma forma orquestrada de ação coletiva, na qual grandes volumes de avaliações são publicados em curtos intervalos de tempo com o objetivo de influenciar a avaliação pública de um jogo. Apesar de sua relevância, análises empíricas em larga escala sobre RB ainda são limitadas. Neste trabalho, apresentamos um estudo longitudinal abrangente de RB na Steam com base em um conjunto de dados de 157.825 jogos e mais de 103 milhões de avaliações, validados contra 89 episódios e 11.800.720 reviews confirmadas de RB e identificadas pelo pipeline oficial de moderação da Steam e verificadas por meio de inspeção manual. Analisamos padrões temporais, participação de usuários e características textuais dos episódios de RB, bem como o conteúdo temático das avaliações por meio de modelagem de tópicos. Nossos resultados demonstram que o RB constitui um fenômeno distinto e empiricamente observável, caracterizado por padrões temporais, comportamentais e linguísticos recorrentes que o diferenciam significativamente do fluxo orgânico de avaliações. Observamos ainda que esses episódios são predominantemente conduzidos por usuários estabelecidos e altamente engajados na plataforma, sendo frequentemente utilizados como veículos para a expressão de insatisfações geopolíticas, culturais e ideológicas, muitas vezes formuladas por meio de linguagem associada ao consumo, como pedidos de reembolso e alegações de práticas fraudulentas. Além disso, mostramos que o RB pode ser instrumentalizado tanto para promover quanto para contestar discursos de ódio, evidenciando desafios relevantes para o equilíbrio entre expressão dos consumidores e mecanismos de governança de plataformas digitais. Por fim, desenvolvemos e avaliamos uma abordagem de aprendizado de máquina para detecção de episódios de RB baseada em atributos temporais, semânticos e comportamentais, alcançando F1-score de 0,67 com XGBoost. Os resultados indicam que mudanças semânticas nas avaliações, variações temporais nos padrões de publicação e a concentração de avaliações negativas constituem os sinais preditivos mais relevantes para a identificação do fenômeno.

Palavras-chave: review bombing; discurso de ódio; Steam; processamento de linguagem natural; modelagem de tópicos; cultura gamer

List of Figures

List of Tables

List of abbreviations and acronyms

API	Application Programming Interface
BERTopic	Bidirectional Encoder Representations Topic Modeling
CDF	Cumulative Distribution Function
FN	False Negative
FP	False Positive
TN	True Negative
TP	True Positive
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
HTTP	Hypertext Transfer Protocol
JSON	JavaScript Object Notation
KDE	Kernel Density Estimation
LGBTQIA+	Lesbian, Gay, Bisexual, Transgender, Queer/Questioning, Intersex, Asexual and others
NLP	Natural Language Processing
Optuna	Hyperparameter Optimization Framework
PC	Personal Computer
RB	Review Bombing
RQ	Research Question
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Over-sampling Technique
t-SNE	t-distributed Stochastic Neighbor Embedding
TF-IDF	Term Frequency–Inverse Document Frequency
UMAP	Uniform Manifold Approximation and Projection
XGBoost	eXtreme Gradient Boosting

Contents

1 Introduction

Since its launch in 2003 by Valve Corporation, Steam[1] has evolved from a digital storefront into a vast socio-technical ecosystem and the largest hub for PC gaming, with millions of daily active users (PLUNKETT, 2013; Steam, 2025). Beyond its commercial role, the platform functions as a large social network in which community interaction—through forums, groups, and user profiles—is as central as the games themselves. In this environment, user reviews form the core of the platform's reputation system, acting as a trust mechanism that shapes purchasing decisions and, consequently, the long-term financial viability of games (MUNNA; RIFAT; BADRUDDUZA, 2020).

This democratic feedback loop is increasingly subverted by *Review Bombing* (RB), characterized by coordinated waves of negative reviews intended to damage a product's reputation by signaling grievances often external to the game's intrinsic quality (CANTONE; TOMASELLI; MAZZEO, 2025; PONISZEWSKA-MARAŃDA; HYNASIŃSKI; BOROWSKA, 2026). While sometimes framed as legitimate consumer protest (WIRZBURGER, 2025), RB can also function as a vehicle for coordinated hostility and harassment (CANTONE; TOMASELLI; MAZZEO, 2025). These campaigns frequently weaponize the review interface to propagate discourse rooted in sexism, racism, homophobia, and xenophobia (CANTONE; TOMASELLI; MAZZEO, 2025). A well-known ideology-driven harassment campaign in gaming was the *Gamergate* controversy, a coordinated online movement widely associated with misogynistic harassment against women developers, journalists, and critics in gaming communities (MORTENSEN, 2018). Beyond social media attacks, *Gamergate* also resulted in coordinated RB incidents, in which users mobilized to post large volumes of negative reviews motivated not only by opposition to perceived ideological positions within gaming culture, but also by misogynistic hostility toward women and feminist representation in games and gaming communities (CHESS; SHAW, 2015; FERGUSON; GLASGOW, 2021).

Despite RB also affecting domains such as movies, books, and e-commerce platforms (SCHUFF; MUDAMBI; WANG, 2024; PAYNE, 2024), digital games constitute a relevant setting due to the highly participatory and identity-driven nature of gaming communities (SALDANHA; SILVA; FERREIRA, 2023). Unlike traditional review systems, platforms such as Steam, where social networking features are deeply intertwined with reputation systems, RB emerges not merely as a form of consumer dissatisfaction, but as a socially coordinated practice shaped by community dynamics, ideological conflict, and platform affordances (WIRZBURGER, 2025). For these reasons, this study focuses exclusively on RB incidents in digital games, using Steam as a large-scale environment.

[1] <https://store.steampowered.com/>

1.1 Motivation

The integrity of user review systems is critical for the functioning of digital markets (BAOWALY; TU; CHEN, 2019). On platforms like Steam, aggregate review scores directly influence purchasing decisions, developer revenue, and long-term game visibility (KRAEMER; WEIGER; HEIDENREICH, 2025). When these systems are subverted by coordinated campaigns, the consequences extend beyond statistical noise: they affect the livelihoods of developers, distort public perception of products, and erode trust in the platform as a whole (FERGUSON; GLASGOW, 2021).

Beyond their economic impact, RB campaigns frequently operate as vehicles for hate speech and coordinated harassment, rooted in broader social conflicts — including misogyny, racism, and homophobia — that spill over from external communities into review platforms (CANTONE; TOMASELLI; MAZZEO, 2025). Games featuring LGBTQIA+ characters, female protagonists, or racial diversity have been disproportionately targeted, with negative reviews functioning less as consumer feedback and more as ideological sanctions (RIBEIRO; PIMENTEL; MELO, 2024; MELO; PIMENTEL, 2022), transforming a commercial infrastructure into a site of digitally mediated social conflict.

In response to rising RB incidents, Valve introduced automated filters in 2017 to identify off-topic RB and exclude their reviews (STEAM, 2019). However, these interventions largely treat RB as a merely statistical anomaly rather than as a socially and politically embedded phenomenon. By framing coordinated review attacks primarily as disruptions to platform functionality and marketplace reliability, such approaches risk overlooking the broader ideological conflicts, toxic speech practices, and collective forms of harassment that frequently motivate these campaigns (CANTONE; TOMASELLI; MAZZEO, 2025). Moreover, reviews on platforms such as Steam are not merely indicators of product quality, but also mechanisms of visibility, participation, and political expression within digital communities. Understanding RB therefore requires examining not only review volume, but also the temporal, behavioral, semantic, and sociopolitical dynamics that underpin these episodes.

Motivated by these limitations, this study investigates RB on Steam through a large-scale empirical analysis of user reviews, combining quantitative pattern analysis, topic modeling, and supervised machine learning to characterize and automatically detect coordinated review activity.

1.2 Research Questions and Hypotheses

To guide our investigation, we formulate the following research questions:

- **RQ1:** How does Review Bombing manifest at scale across the Steam ecosystem,

and what common structural patterns unify these distinct coordinated episodes?

- **RQ2:** To what extent do the temporal, behavioral, and game-level properties of Review Bombing episodes exhibit distinct signatures compared to organic feedback?
- **RQ3:** What topics and sociopolitical motivations drive users to exploit review systems during RB episodes?
- **RQ4:** How effectively can machine learning models distinguish RB from organic negative feedback given shifting contexts and heterogeneous user expression?
- **RQ5:** Which semantic, temporal, and behavioral features contribute most to the automated detection of RB episodes?

We hypothesize that RB episodes share identifiable structural, temporal, and semantic patterns that both unify them under a common pattern and set them apart from organic negative feedback, making automated detection through temporal and semantically machine learning models feasible.

1.3 Proposed Approach

To answer these questions, we analyze a large dataset comprising over 103 million user reviews from 157,825 Steam games. We rely on Steam's official "off-topic" (STEAM, 2019) moderation labels as ground truth to identify and validate 89 RB episodes. Our methodology combines longitudinal quantitative analysis of temporal and behavioral patterns with transformer-based topic modeling (BERTopic (GROOTENDORST, 2022)) to examine review content.

We further enrich the supervised learning setting by engineering ten additional behavioral and temporal features, alongside semantic embedding representations. To better understand feature separability and representation quality, we conduct exploratory feature analyses using Kernel Density Estimation (KDE), a non-parametric density estimation method (PARZEN, 1962), t-distributed Stochastic Neighbor Embedding (t-SNE) (MAATEN; HINTON, 2008), and SHapley Additive exPlanations (SHAP) (LUNDBERG et al., 2020). Finally, we develop supervised classifiers—including XGBoost (CHEN; GUESTRIN, 2016), Random Forests (BREIMAN, 2001), and neural networks (FINE, 1999)—to evaluate the feasibility of automated RB detection.

1.4 Contributions

Our findings portray RB as a complex social and geopolitical phenomenon that extends beyond simple negativity or noise. RB is largely driven by established, engaged

users who use the review infrastructure for collective positioning and, at times, hostile expression. These episodes employ a recognizable protest lexicon in which commercial complaints, such as refund demands, often mask deeper cultural, ideological, or geopolitical conflicts. Our machine learning experiments further introduce a framework that models RB as coordinated collective action unfolding over time. The proposed temporal and feature-enhanced representation enabled reliable RB detection, with XGBoost achieving the best performance and reaching an F1-score of 0.677. Feature importance analyses reveal that semantic deviation from a game's historical review discourse, abrupt changes relative to previous review periods, and concentrations of negative reviews are the strongest indicators of RB activity. Although reliably distinguishing RB from organic feedback remains challenging due to evolving discourse.

Understanding these dynamics is essential for addressing current design gaps in platform governance. When review systems are weaponized to propagate coordinated hate speech or enforce exclusionary norms, they can silence diverse voices and discourage inclusive development. By identifying the motivations and tactical adaptations behind RB—including strategies such as positive review surges to bypass moderation—this work provides an empirical foundation for more nuanced moderation approaches that move beyond reactive filtering.

1.5 Organization

The remainder of this work is organized as follows. Chapter 2 discusses the theorical foundation of this work; Chapter 3 discusses related work; Chapter 4 introduces the RB phenomenon and Steam's moderation framework; Chapter 5 presents the dataset and methodology; Chapter 6 analyzes RB dynamics; Chapter 7 examines review content through topic modeling; Chapter 8 presents the detection models; and Chapter 9 concludes with implications and future directions.

2 Theoretical Foundation

This chapter presents the theoretical background underlying the techniques used throughout this dissertation. It is organized around three broad families of methods: (i) unsupervised techniques for discovering latent structure in text and numerical data, namely topic modeling, clustering, and dimensionality reduction; (ii) supervised learning algorithms for classification, together with the strategies used to handle imbalanced data, tune hyperparameters, and validate results; and (iii) the statistical and interpretability tools that allow a trained model's behavior to be examined and its predictions to be explained. Rather than describing how each method was applied, the goal here is to explain, from first principles, why each method works, what assumptions it makes, and what its strengths and limitations are.

2.1 Topic Modeling

Topic modeling refers to a family of unsupervised techniques that aim to discover the latent thematic structure of a collection of documents. Classical approaches, such as Latent Dirichlet Allocation (LDA), model each document as a probabilistic mixture of topics and each topic as a probability distribution over words, estimated purely from word co-occurrence statistics. While effective on long, well-curated corpora, these bag-of-words models struggle with short, noisy, or highly contextual texts, since they ignore word order and semantic relationships between terms that never literally co-occur.

BERTopic (GROOTENDORST, 2022) addresses this limitation by decoupling the problem into a modular pipeline that first represents documents in a semantically rich embedding space, then discovers clusters of similar documents in that space, and finally derives an interpretable, word-based description for each cluster. Because it relies on contextual embeddings rather than raw word frequencies, it can group together documents that express the same idea using entirely different vocabulary. The pipeline is composed of four conceptually independent stages, described below.

2.1.1 Sentence Embeddings

The first stage converts each document into a fixed-length dense vector, or *embedding*, that captures its meaning. This is achieved with a pre-trained sentence-transformer model, a neural network trained (typically via a contrastive or Siamese objective) so that sentences with similar meaning are mapped to nearby points in the embedding space, while unrelated sentences are mapped far apart. Unlike traditional representations such as bag-of-words or TF-IDF vectors, which are sparse and depend directly on vocabulary

overlap, sentence embeddings are dense, low-dimensional, and encode semantic similarity independently of the exact words used.

The model `paraphrase-multilingual-MiniLM-L12-v2` (REIMERS; GUREVYCH, 2019) is a distilled, multilingual variant of a transformer encoder: it retains most of the semantic capacity of larger models such as BERT while requiring substantially less computation, and it maps any input sentence, regardless of language, into a shared 384-dimensional vector space. This multilingual alignment is particularly relevant when a corpus mixes several languages, since semantically equivalent sentences in different languages are still placed close together.

2.1.2 Dimensionality Reduction with UMAP

Embeddings produced by sentence-transformer models typically live in spaces of a few hundred dimensions. Although this dimensionality is necessary to capture fine-grained semantic nuances, most clustering algorithms degrade in high-dimensional spaces due to the so-called *curse of dimensionality*: as the number of dimensions grows, distances between points become increasingly uniform, making it hard to distinguish genuine clusters from noise.

UMAP (Uniform Manifold Approximation and Projection) (MCINNES, 2018) mitigates this problem by projecting high-dimensional data onto a much lower-dimensional space (commonly two to ten dimensions) while preserving as much of the original topological structure as possible. Conceptually, UMAP first builds a weighted graph in which each point is connected to its nearest neighbors, with edge weights reflecting the probability that two points belong to the same local neighborhood. It then searches for a low-dimensional layout whose corresponding neighborhood graph is as similar as possible to the original one, formalized through a cross-entropy loss between the two graphs' edge-weight distributions. In contrast to linear techniques such as Principal Component Analysis (PCA), UMAP is non-linear and is explicitly designed to preserve both local neighborhoods (which points are close to which) and, to a reasonable extent, global structure (how clusters of points relate to one another), which makes it particularly well suited as a pre-processing step before density-based clustering.

2.1.3 Density-Based Clustering with HDBSCAN

Once documents are represented in a low-dimensional space, the next step is to group them into clusters that hopefully correspond to coherent topics. Centroid-based algorithms such as K-means require the number of clusters to be specified in advance and assume clusters are roughly spherical and of similar size and density — assumptions that rarely hold for real text corpora, where some topics are discussed far more often, and with far more variety, than others.

HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) (MCINNES; HEALY; ASTELS, 2016) instead defines clusters as regions of high point density separated by regions of low density, without requiring the number of clusters as input. It works by estimating, for every point, a notion of local density based on the distance to its nearest neighbors, and then building a hierarchy of nested clusters as the density threshold is gradually relaxed, in a manner analogous to single-linkage hierarchical clustering but weighted by density. From this hierarchy, HDBSCAN extracts the set of clusters that are the most *stable* — that is, the clusters that persist across the widest range of density thresholds — rather than simply cutting the hierarchy at a fixed height. A crucial consequence of this density-based formulation is that HDBSCAN can identify clusters of very different sizes and shapes simultaneously, and, importantly, it can explicitly label points that do not belong to any dense region as *noise*, instead of forcing every point into some cluster. This is especially valuable for topic modeling, since not every document is expected to belong to a well-defined topic.

2.1.4 Topic Representation via Class-Based TF-IDF

Once documents have been partitioned into clusters, each cluster must still be translated into a human-interpretable description, typically a small set of representative keywords. BERTopic accomplishes this with a class-based adaptation of **TF-IDF** (Term Frequency–Inverse Document Frequency) (JONES, 1972).

In its classical form, TF-IDF measures how important a term is to an individual document within a collection by multiplying how frequently the term occurs in that document (term frequency) by a factor that penalizes terms that occur in many documents across the collection (inverse document frequency), on the intuition that a term which appears everywhere carries little discriminative information. The class-based variant applies this same logic, but treats every cluster — rather than every individual document — as a single unit, concatenating all documents that were assigned to that cluster into one large "document." The term-frequency component then reflects how often a word appears within that cluster, and the inverse-document-frequency component penalizes words that are common across many clusters. The words with the highest resulting scores in a cluster form its topic representation: they are the words that occur unusually often in that cluster relative to how often they occur elsewhere, and are therefore the words that best distinguish that topic from the others.

2.2 Data Balancing Techniques

Many real-world classification problems are affected by *class imbalance*, a situation in which one class (the majority class) is represented by far more examples than the other (the minority class). This is problematic because most learning algorithms are implicitly

optimized to minimize overall error, which they can often achieve simply by favoring the majority class and largely ignoring the minority one — yielding a model with deceptively high accuracy but very poor ability to detect the class of actual interest. Resampling techniques address this issue by altering the class distribution of the training data itself, either by removing majority-class examples, adding synthetic minority-class examples, or both.

2.2.1 Tomek Links

A **Tomek link** (Imbalanced Learn Developers, 2025) is defined as a pair of examples belonging to different classes that are each other's nearest neighbor: no other example in the dataset lies closer to either of the two than they lie to each other. Such pairs typically sit right at the boundary between classes, and often correspond to noisy examples, mislabeled instances, or genuinely ambiguous cases that blur the decision boundary between classes.

The technique built around this concept works by identifying all such pairs in the training set and removing the majority-class example from each one, while keeping the minority-class example untouched. The resulting effect is a form of *data cleaning* rather than balancing in the strict numerical sense: it does not necessarily equalize the size of the two classes, but it removes majority-class points that lie dangerously close to the minority class, producing a cleaner, more clearly separated decision boundary for a subsequent classifier to learn.

2.2.2 SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) (Imbalanced Learn Developers, 2026) takes the complementary approach of increasing the representation of the minority class by generating new, synthetic examples, rather than simply duplicating existing ones (which would add no new information and increase the risk of overfitting). For each minority-class instance selected for oversampling, SMOTE identifies one of its nearest neighbors within the same class and creates a new synthetic point by interpolating linearly between the two: a random point is chosen along the line segment that connects the original instance to its neighbor in feature space. By construction, the synthetic examples lie within the region already spanned by real minority-class instances, which allows the minority class to be enlarged while preserving the overall shape of its distribution in feature space, rather than simply repeating existing points.

In practice, Tomek Links and SMOTE are frequently combined: SMOTE first enlarges the minority class with synthetic examples, and Tomek Links are then used to remove any majority-class points left uncomfortably close to the newly expanded minority region, yielding a training set that is both better balanced and better separated.

2.3 Clustering and Similarity Measures for Feature Engineering

Beyond their use in topic discovery, clustering and similarity computations over text embeddings can themselves be turned into predictive features, for instance to quantify how semantically concentrated a set of texts is, or how much redundancy exists among them. This section presents two techniques that support such feature construction.

2.3.1 Mini-Batch K-Means

Standard K-means partitions a dataset into K clusters by iteratively assigning each point to its nearest centroid and then recomputing each centroid as the mean of the points assigned to it, repeating until convergence. Its main computational drawback is that, at every iteration, it must examine the entire dataset in order to update the centroids, which becomes expensive for very large datasets.

Mini-Batch K-Means (SCULLEY, 2010) addresses this cost by updating the centroids using only a small, randomly sampled subset (a mini-batch) of the data at each iteration, rather than the full dataset. Each centroid is updated with a learning rate that decreases as more mini-batches are processed, analogous to stochastic gradient descent, so that later updates have progressively less influence and the centroids gradually stabilize. This substantially reduces both computation time and memory requirements, at the cost of a small amount of noise in the centroid estimates, which in practice tends to have negligible impact on the final clustering quality, making the algorithm well suited to large-scale text corpora.

2.3.2 Cosine Similarity

Cosine similarity quantifies how similar the *direction* of two vectors is, irrespective of their magnitude. For two non-zero vectors A and B, it is defined as the cosine of the angle θ between them (BAEZA-YATES; RIBEIRO-NETO et al., 1999):

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}.$$

The measure ranges from -1, indicating vectors pointing in exactly opposite directions, to $+1$, indicating vectors pointing in the same direction, with 0 indicating orthogonality (no linear directional relationship). Because it normalizes away the magnitude of the vectors, cosine similarity is particularly well suited for comparing text representations, where the "length" of a vector (for instance, one influenced by document length or overall word frequency) is often irrelevant to the semantic content that is actually of interest, and only the direction — which encodes meaning — should matter.

2.4 Feature Analysis and Interpretability

Once features have been engineered, it is important to understand their distribution, their relationships with one another and with the target variable, and, once a model has been trained, which features it actually relies on. This section introduces the statistical, visualization, and interpretability tools used for this purpose.

2.4.1 Kernel Density Estimation (KDE)

Kernel Density Estimation (PARZEN, 1962) is a non-parametric method for estimating the probability density function underlying a set of observed data points, without assuming the data follows any particular parametric distribution (such as a Gaussian). The estimate is built by placing a smooth kernel function — most commonly a Gaussian — centered at each observed data point, and summing these kernels together. Where many observations are concentrated, the overlapping kernels reinforce one another and produce a high estimated density; where observations are sparse, the estimated density is correspondingly low. The width of the kernel (the bandwidth) controls the trade-off between a smooth, potentially over-generalized estimate and a jagged, potentially overfitted one. KDE is particularly useful for visualizing how a variable, or a low-dimensional projection of a set of variables, is distributed, without imposing artificial assumptions about the shape of that distribution.

2.4.2 t-SNE

t-distributed Stochastic Neighbor Embedding (t-SNE) (MAATEN; HINTON, 2008) is a non-linear technique for visualizing high-dimensional data in two or three dimensions. It proceeds by first converting the pairwise distances between points in the original high-dimensional space into a probability distribution that reflects how likely each point is to pick another as its neighbor, modeled with a Gaussian kernel. It then searches for a corresponding low-dimensional configuration of points whose pairwise similarities, this time modeled with a heavier-tailed Student's t-distribution, match the original distribution as closely as possible, by minimizing the Kullback–Leibler divergence between the two distributions. The use of a heavy-tailed distribution in the low-dimensional space is what allows t-SNE to push dissimilar points comfortably far apart while still keeping similar points close together, which tends to produce visualizations in which clusters are visually well separated. It is important to note that t-SNE is primarily a visualization tool: distances between well-separated clusters in a t-SNE plot are not, in general, quantitatively meaningful, and the technique is not intended to preserve global structure in the way UMAP attempts to.

2.4.3 Pearson Correlation Coefficient

The **Pearson correlation coefficient** (PEARSON, 1895) quantifies the strength and direction of the *linear* relationship between two continuous variables X and Y. Given n paired observations, it is computed as

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

where $\bar{x}$ and $\bar{y}$ are the sample means of X and Y. The numerator is (proportional to) the covariance between the two variables, while the denominator normalizes this covariance by the variables' individual standard deviations, so that r_{XY} always lies between -1 and $+1$. A value close to $+1$ indicates a strong positive linear relationship, a value close to -1 indicates a strong negative linear relationship, and a value close to 0 indicates the absence of a linear relationship, though it does not rule out a strong non-linear one.

2.4.4 Correlation and Causation

A recurring pitfall in the interpretation of statistical associations is conflating correlation with causation. A high Pearson coefficient only indicates that two variables tend to vary together in a linear fashion; it says nothing about whether one causes the other. Three common alternative explanations for an observed correlation are worth keeping in mind: a genuine causal link may run in the opposite direction from the one assumed; a third, unmeasured variable (a confounder) may independently influence both variables, creating an association between them without either causing the other; or the correlation may simply be a product of chance, particularly when many variable pairs are tested. For this reason, correlations reported in an empirical analysis should be interpreted strictly as statistical associations, and any causal claim requires additional evidence beyond correlation alone, typically from controlled experiments or causal-inference methods designed specifically for observational data.

2.4.5 Model Interpretation with SHAP

Complex machine learning models, particularly ensembles of trees or neural networks, achieve strong predictive performance partly because they can capture intricate, non-linear interactions between features, but this same complexity makes it difficult to understand *why* a model produced a particular prediction. **SHAP** (SHapley Additive exPlanations) (LUNDBERG et al., 2020) addresses this by adapting a concept from cooperative game theory, the Shapley value, to the problem of feature attribution.

In game theory, the Shapley value provides a principled way of fairly distributing the total "payoff" of a cooperative game among its players, based on each player's

average marginal contribution across all possible orders in which players could join the game. SHAP reframes a model's prediction for a given instance as the payoff of a game in which the "players" are the input features, and computes each feature's Shapley value as its average marginal contribution to the prediction, considering all possible subsets of the other features. The result is a decomposition of the difference between the model's prediction for that instance and its average prediction over the whole dataset into additive contributions from each individual feature. This construction guarantees three properties considered desirable for an explanation method: *local accuracy* (the feature contributions sum exactly to the difference between the individual prediction and the average prediction), *consistency* (a feature's attributed importance cannot decrease if the model comes to depend on it more strongly), and correct handling of *missingness* (features that are absent or unused receive no attributed importance).

It is important to emphasize what SHAP values do and do not indicate. They quantify how much a model relies on a given feature to produce its predictions, given the patterns the model happened to learn from the training data — they do not necessarily reveal the true, real-world importance of that feature, nor any causal role it may play. A model may, for instance, assign substantial importance to a feature that is merely a proxy correlated with the true underlying cause of the outcome, without that feature having any causal bearing on it itself.

2.5 Classification Models

This section describes three supervised learning algorithms that are widely used for classification tasks and that represent distinct modeling paradigms: gradient-boosted trees, bagged trees, and artificial neural networks.

2.5.1 XGBoost

XGBoost (eXtreme Gradient Boosting) (CHEN; GUESTRIN, 2016) belongs to the family of *gradient boosting* algorithms, which build a strong predictive model as an ensemble of many weak learners — typically shallow regression trees — added sequentially, rather than trained independently. At each boosting round, a new tree is trained specifically to predict the residual errors (more precisely, the negative gradient of the loss function) left by the ensemble built so far, so that each successive tree incrementally corrects the mistakes of its predecessors. XGBoost's specific contribution is an efficient and regularized implementation of this idea: it augments the standard gradient-boosting objective with an explicit penalty on tree complexity (number of leaves and magnitude of leaf weights), which helps prevent overfitting, and it introduces algorithmic optimizations — such as approximate split-finding and native handling of sparse or missing data — that allow it to scale efficiently to large datasets and to support parallel computation

during tree construction. These properties have made it a consistently strong baseline for structured, tabular data.

2.5.2 Random Forest

Random Forest (BREIMAN, 2001) follows a different ensembling strategy, known as *bagging* (bootstrap aggregating), in which many decision trees are trained independently and in parallel, rather than sequentially. Each tree is trained on a different *bootstrap sample* of the training data — a sample of the same size as the original dataset, drawn with replacement — which ensures that no two trees see exactly the same data. In addition, at each split within each tree, only a random subset of the available features is considered as candidates, rather than all of them. This additional layer of randomization decorrelates the individual trees, so that their errors are less likely to be systematically similar. The forest's final prediction is obtained by aggregating the predictions of all its trees, through majority voting for classification tasks or averaging for regression tasks. Because each individual tree tends to overfit its own bootstrap sample in a slightly different way, averaging across many such trees substantially reduces the overall variance of the ensemble compared to a single decision tree, typically without a large increase in bias.

2.5.3 Feedforward Neural Networks

A **feedforward neural network** (FINE, 1999) is composed of layers of interconnected computational units (neurons), organized so that information flows strictly from an input layer, through one or more hidden layers, to an output layer, with no cycles or feedback connections. Each neuron computes a linear combination of its inputs, followed by a non-linear *activation function* (such as ReLU or sigmoid); it is precisely this non-linearity, applied repeatedly across layers, that allows a neural network to approximate complex, non-linear relationships between inputs and outputs, in a manner that purely linear models cannot. The network's parameters (the weights and biases of every connection) are learned through *back-propagation*, an algorithm that computes the gradient of a chosen loss function with respect to every parameter by applying the chain rule of calculus backward through the network, and then updates the parameters, typically via a variant of gradient descent, so as to progressively reduce the loss on the training data. For binary classification tasks specifically, the output layer commonly consists of a single neuron with a sigmoid activation function, which squashes its output into the $(0, 1)$ interval so that it can be interpreted as the estimated probability of the positive class, trained by minimizing the binary cross-entropy loss between this predicted probability and the true class label.

2.6 Hyperparameter Optimization and Validation

Beyond the parameters a model learns directly from data, most learning algorithms also expose *hyperparameters* — settings such as the number of trees, tree depth, or learning rate, which are not learned from data but must be chosen beforehand and strongly affect the model's final performance. This section presents the techniques used to search for good hyperparameter configurations and to estimate a model's performance reliably.

2.6.1 Optuna

Optuna (AKIBA et al., 2019) is a framework for automating the search over hyperparameter configurations. Rather than requiring the entire search space to be declared upfront, as in a fixed grid search, Optuna uses a *define-by-run* interface, in which the search space is defined dynamically, as the code that trains and evaluates the model is executed, allowing conditional and more flexible search spaces. It samples candidate configurations using algorithms such as the Tree-structured Parzen Estimator, a Bayesian-optimization method that models which regions of the hyperparameter space are more likely to yield good performance based on the results of previous trials, and preferentially samples new configurations from those promising regions rather than searching uniformly at random. Optuna also incorporates *pruning* strategies, which monitor a trial's performance during training and terminate it early if intermediate results indicate it is unlikely to outperform previously completed trials, avoiding wasted computation on clearly unpromising configurations. Together, these mechanisms make it possible to explore large hyperparameter spaces considerably more efficiently than an exhaustive search would allow.

2.6.2 Cross-Validation

Estimating how well a model will generalize to unseen data requires evaluating it on data it was not trained on. A single train/test split, however, produces an estimate that depends heavily on which particular examples happened to fall into the test set, especially when the available dataset is small. **K-fold cross-validation** (KOHAVI, 1995) addresses this by partitioning the dataset into K equally sized, non-overlapping folds; the model is then trained K times, each time using $K - 1$ of the folds for training and the remaining fold, held out, for validation, so that every observation is used for validation exactly once across the K repetitions. Averaging the validation metric across all K repetitions yields a more stable and reliable estimate of generalization performance than a single split would, while still making use of the entire available dataset for both training and validation purposes.

2.7 Evaluation Metrics for Classification

Evaluating a classifier's performance, especially under class imbalance, requires more than a single aggregate metric. This section presents the metrics used to assess classification performance throughout this work, all of which are derived from the fundamental counts of correct and incorrect predictions for each class: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

- **Accuracy** is the overall fraction of predictions the model gets right, across both classes:
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}.$$
Although intuitive, accuracy can be a misleading metric under class imbalance: a model that always predicts the majority class can achieve high accuracy while never correctly identifying a single minority-class instance.

- **Precision** measures, among all instances the model predicted as positive, what fraction actually were positive:
$$\text{Precision} = \frac{TP}{TP + FP}.$$
High precision indicates that positive predictions can be trusted, i.e., that false alarms are rare.

- **Recall** (also called sensitivity) measures, among all instances that were actually positive, what fraction the model correctly identified:
$$\text{Recall} = \frac{TP}{TP + FN}.$$
High recall indicates that the model misses few actual positive cases.

- **F1-score** combines precision and recall into a single metric by taking their harmonic mean:
$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$
Because it is a harmonic (rather than arithmetic) mean, the F1-score penalizes models that achieve a high value in one of the two component metrics only at the expense of a very low value in the other, making it a useful single-number summary when both precision and recall matter and neither should be sacrificed disproportionately for the other.

Precision and recall are generally in tension with one another: a classifier can typically be tuned to increase one at the cost of the other, for instance by adjusting

the decision threshold above which a prediction is considered positive. In imbalanced classification problems specifically, where failing to detect a positive instance (a false negative) is often considerably more costly than raising a false alarm (a false positive), recall is frequently prioritized as the metric of primary interest, with precision and F1-score monitored to ensure this does not come at an unacceptable cost in the rate of false positives.

3 Related Work

3.1 Review Bombing Analyses

Research on the RB phenomenon initially emerged through investigations of high-profile controversies within gaming communities. Early studies focused on individual incidents to understand how online communities mobilize around cultural, political, and ideological disputes. Examples include the #GamerGate harassment campaigns (CHATZAKOU et al., 2017b; CHATZAKOU et al., 2017a), the backlash against *Battlefield V* (#NotMyBattlefield) (HOWE; LIVINGSTON; LEE, 2019), the exclusionary reactions to LGBTQ+ character representation in *Overwatch* (VÄLISALO; RUOTSALAINEN, 2022), and the coordinated attacks targeting *The Last of Us Part II* (MELO; PIMENTEL, 2022; CANTONE; TOMASELLI; MAZZEO, 2025; LETIZI; NORMAN, 2024). These studies demonstrate that RB often extends beyond dissatisfaction with product quality and instead reflects broader social conflicts, ideological polarization, and identity-based disputes. Similar dynamics have also been observed in *Borderlands 2* (WIRZBURGER, 2025), *No Man's Sky* (LU et al., 2020), and *Cyberpunk 2077* (SIUDA et al., 2024), where community backlash was triggered by developer actions, controversial decisions, or unmet expectations.

While these case studies established the relevance of RB as a form of coordinated online action, recent research has shifted toward understanding the phenomenon at a broader scale. Rather than treating RB as a collection of isolated incidents, newer studies examine recurring patterns and common motivations across multiple events. For example, (RIBEIRO; PIMENTEL; MELO, 2024) analyzed approximately one million Metacritic[1] reviews and identified three recurring drivers of RB in games: dissatisfaction with corporate decisions, ideological or political disagreement, and frustration with franchise design choices. Similar findings reported by (CANTONE; TOMASELLI; MAZZEO, 2025) and (JUNIOR et al., 2025) suggest that ideological grievances, particularly those involving gender, diversity, and LGBTQIA+ representation, are frequently expressed through seemingly technical critiques of gameplay or narrative quality.

Beyond digital games, RB has also been studied in other review ecosystems. (SCHUFF; MUDAMBI; WANG, 2024) investigated RB in movies and television using Metacritic data and proposed the Bombed Index (BI), a metric based on divergences between critic and user evaluations. Their results indicate that bombed reviews tend to be shorter, more negative, and more frequently unrelated to the evaluated product. Similarly, (PAYNE, 2024) examined politically motivated review attacks against businesses on Yelp

[1] <https://www.metacritic.com/>

and Google Maps, showing how review systems can become instruments for broader social and political disputes. Together, these studies demonstrate that RB is not exclusive to gaming platforms but represents a broader challenge for online reputation systems.

Despite the growing body of literature, large-scale studies focusing on gaming platforms remain relatively scarce. Games constitute a particularly important setting for RB research due to the strong relationships between players, developers, and online communities (SALDANHA; SILVA; FERREIRA, 2023). Among gaming platforms, Steam occupies a unique position because of its scale, central role in the PC gaming market (THORHAUGE, 2024), and explicit recognition of the RB phenomenon through its off-topic review activity system (STEAM, 2019). Nevertheless, studies specifically examining RB on Steam remain limited. One notable exception is the work of (WIRZBURGER, 2025), who adopts a governmentality perspective to argue that Steam manages politically motivated RB episodes by labeling them as "off-topic" and reducing their visibility through interface and algorithmic mechanisms. This limited number of large-scale Steam studies highlights the need for broader empirical investigations capable of characterizing RB across multiple games, communities, and historical contexts.

3.2 Review Bombing Detection

Research on RB detection has largely emerged from related areas such as fake review detection, spam identification, and sentiment anomaly analysis. A recent literature review by (PONISZEWSKA-MARAŃDA; HYNASIŃSKI; BOROWSKA, 2026) categorizes existing approaches into statistical methods, traditional machine learning, deep learning, and graph-based models. For example, (CORONADO-BLÁZQUEZ, 2024) analyzed more than 513 thousand Metacritic reviews using TF-IDF features and traditional machine learning classifiers, achieving approximately 88% accuracy. Graph-based approaches have also shown promising results, such as the heterogeneous graph transformer model proposed by (CAI et al., 2024), which achieved 84.8% accuracy, and GAIM (ZHANG et al., 2022), which combines Graph Attention Networks with textual embeddings and obtained over 91% accuracy on movie review datasets. Similarly, deep learning approaches explored by (SHEN; LIU; SHEN, 2023) and (WANG; WU, 2020) leveraged neural architectures to model relationships between review content, ratings, and user behavior. Overall, the literature indicates that integrating semantic, temporal, and behavioral information improves detection performance, while also highlighting the scarcity of large-scale RB-specific datasets as one of the main open challenges in the field (PONISZEWSKA-MARAŃDA; HYNASIŃSKI; BOROWSKA, 2026).

3.3 Research Gap

Although research on the RB phenomenon has expanded in recent years, important gaps remain. Most existing studies focus on isolated RB episodes or small collections of highly publicized cases (CANTONE; TOMASELLI; MAZZEO, 2025; SCHUFF; MUDAMBI; WANG, 2024). While these investigations provide valuable insights, they offer limited evidence regarding whether their findings generalize across different games, communities, historical periods, and motivations. Consequently, there is still little understanding of the common structural patterns that characterize the RB phenomenon at the platform level.

The same challenge extends to RB detection. Existing approaches frequently rely on manually selected incidents or methods adapted from adjacent domains such as spam and fake review detection (PONISZEWSKA-MARAŃDA; HYNASIŃSKI; BOROWSKA, 2026). Although many studies report strong predictive performance (CORONADO-BLÁZQUEZ, 2024; SHEN; LIU; SHEN, 2023), their models are often evaluated on narrowly defined datasets and specific case studies, making it unclear whether the learned patterns truly capture the broader RB phenomenon or merely reflect characteristics of particular events. Furthermore, most approaches (CAI et al., 2024; ZHANG et al., 2022) frame detection as a game-level classification task, identifying whether a title experienced an RB episode while providing limited insight into when the episode occurred. Consequently, the scarcity of large-scale labeled datasets remains one of the main obstacles to developing robust and broadly applicable RB detection models (PONISZEWSKA-MARAŃDA; HYNASIŃSKI; BOROWSKA, 2026).

This dissertation addresses these gaps through a large-scale empirical investigation of RB on Steam, combining platform-wide characterization, topic modeling, and machine learning-based detection using more than 103 million reviews, 37,636,245 unique users and 157,825 games collected.

4 The Review Bombing Phenomenon and Steam

4.1 The Steam Platform

Steam is one of the largest digital distribution platforms for PC gaming, developed and maintained by Valve Corporation (PLUNKETT, 2013; Steam, 2025). Launched in 2003, it has evolved from a simple storefront and update client into a comprehensive ecosystem that encompasses game purchasing, automatic patching, community forums, user-generated content (such as mods via the Steam Workshop), social networking, and detailed player statistics. As of 2024, Steam hosts tens of thousands of titles and serves as the primary marketplace for PC games, making its integrated user review system one of the most influential sources of consumer feedback in the industry (Steam, 2025).

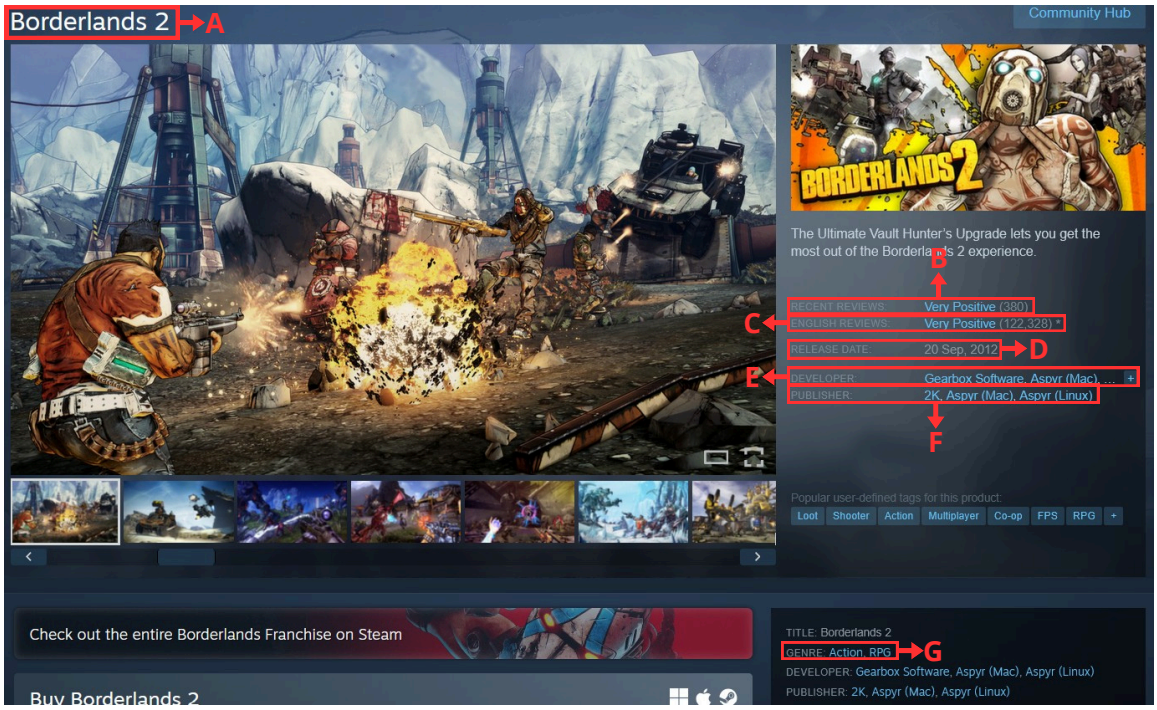


Figure 1 – Example of Steam's homepage for the game *Borderlands 2*.

Figure 1 displays the Steam store page for *Borderlands 2*, illustrating the standard layout of a game's storefront. The page is anchored by the game title (**A**). Item **B** shows the *recent review score*, indicating how the game is recently being evaluated by users, while item **C** presents the *all-time review score*, summarizing the cumulative user reviews across the game's lifespan. Both summaries are computed from binary user recommendations (discussed below) and help prospective buyers decide whether the game is worth purchasing. The remaining metadata includes: **D** – the original release date; **E**

– the developer; **F** – the publisher. It is worth noting the distinction between developer and publisher: the developer is the studio that designs and creates the game, whereas the publisher handles the manufacturing, marketing, distribution, and often the financial investment required to bring the title to market. Finally, item **G** lists the genre tags, which are assigned by the developer when the game is first registered on the platform.

The computation underlying elements **B** and **C** is straightforward. Steam's review summaries are derived from the binary recommendations submitted by users (discussed below). The platform categorizes overall sentiment using a combination of public thresholds and proprietary, non-disclosed factors (KREMERS, 2026); however, the publicly documented component is based on the proportion of positive and negative reviews. This results in the following classification scheme about the porcentage of positive reviews: Overwhelmingly Positive (95%), Very Positive (80%), Positive (80% but with a lower total number of reviews), Mostly Positive (70%), Mixed (40%), Mostly Negative (20%), Negative (20%), and Very Negative (<20%). These discrete thresholds imply that even relatively small changes in the proportion of positive reviews can cause a title to transition between categories, a characteristic that becomes particularly significant in the context of RB episodes.

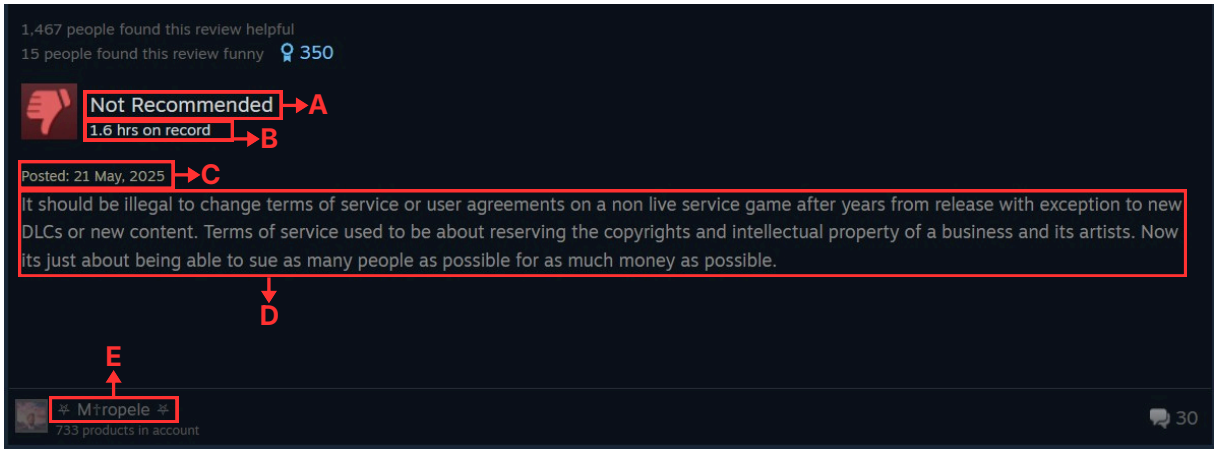


Figure 2 – Example of a Steam user review for *Borderlands 2*.

Figure 2 shows an individual user review. Only users who have acquired the game through Steam (by purchase, gift, or promotional license) are permitted to write a review (STEAM, 2026). The annotated components are: **A** – the binary verdict "Recommended" (Positive) or "Not Recommended" (Negative); **B** – the number of hours the user had played at the time the review was posted; **C** – the date of publication; **D** – the free-text review body; and **E** – the user's public username. Throughout this study, reviews marked as "Recommended" are considered positive reviews, whereas reviews marked as "Not Recommended" are considered negative reviews.

Figure 3 illustrates the visual marker Steam applies when it detects an RB episode. The histogram represents the volume of reviews over time, with each bar corresponding

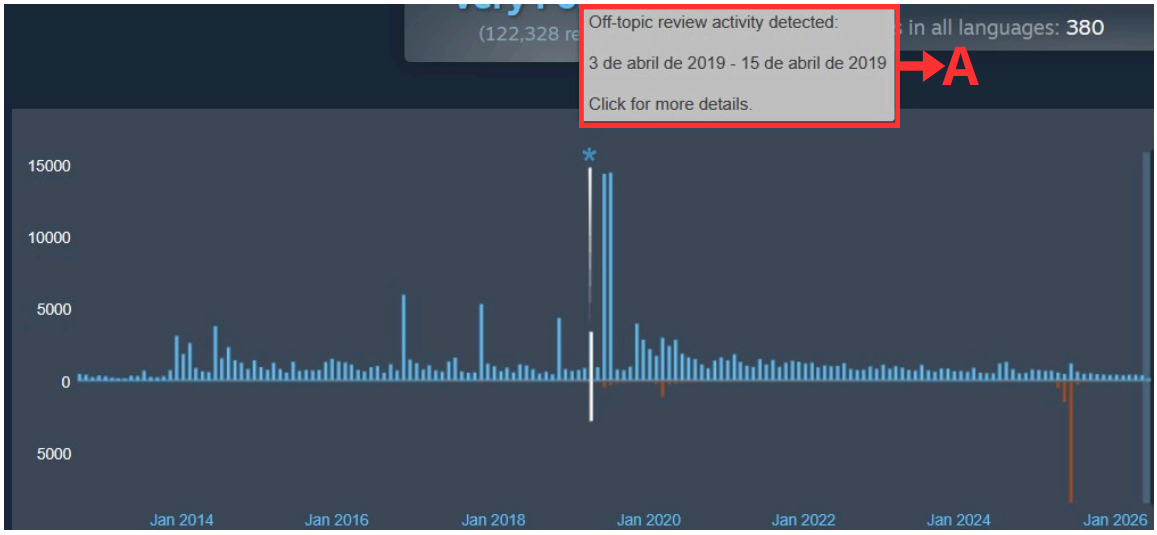


Figure 3 – Example of the visual marker Steam applies during an RB episode, shown on the game page of *Borderlands 2*.

to a defined period (typically one month). Positive reviews are displayed as blue bars extending above the horizontal axis, while negative reviews appear as red bars below it. The area labelled **A** highlights the timeframe of a detected RB episode for *Borderlands 2*, spanning from 3 April 2019 to 15 April 2019. During this interval, the histogram bars turn white, signalling that Steam's automated system has flagged the reviews from this period as off-topic.

4.2 Review Bombing History

Although studies on RB have only recently begun to appear, RB itself is not a new phenomenon. One of the earliest documented uses of the term dates back to 2008, in an article published by the website *Ars Technica* discussing the game *Spore* (KUCHERA, 2008). The article reported that *Spore* received a wave of negative user reviews after the developers introduced an online activation key requirement; in response, users took to Amazon and subjected the game to a coordinated campaign of negative reviews.

In the following years, the term gradually gained wider acceptance within gaming communities. In 2011, the titles *Toy Soldiers: Cold War* and *Bastion* (CARTER, 2011) experienced a flood of zero-score reviews on Metacritic, apparently instigated by trolls whose sole objective was to lower the games' aggregate scores. Over time, however, the notion of RB evolved and became more closely associated with reviews that are disconnected from a game's intrinsic quality and instead reflect conflicts between communities and developers (Wikipedia Contributors, 2026). The year 2015 witnessed three prominent cases that illustrate this shift. *Titan Souls* was targeted by RB on Steam after a dispute between the developer and the influential content creator *TotalBiscuit* (HERNANDEZ, 2015); his community subsequently orchestrated a wave of negative reviews. In a similar

vein, *The Elder Scrolls V: Skyrim* and *Fallout 4* (MCKEAND, 2017) were both subjected to RB on Steam following the announcement that *mods*—user-generated additions to the game—would become paid products, a decision that sparked widespread community outrage and a deluge of negative reviews on the platform.

This pattern continued in 2017, when a significant wave of RB was carried out by Chinese players. These players felt aggrieved after the developers of *Crusader Kings II* raised the game's price in certain regions, including China, and by the absence of a Chinese localization for *Nier: Automata* (FENLON, 2017). In both instances, the players used Steam as a forum for protest, exploiting the review system as a platform to voice their demands. *PlayerUnknown's Battlegrounds* was also affected by RB from Chinese players after the developer introduced in-game advertisements for a VPN service, which many players viewed as an attempt to monetize ongoing connectivity problems rather than solve them (HALL, 2017). An important element in those episodes is the recognition by the gaming media of the phenomenon's importance and impact: PC Gamer wrote: "Flooding Steam reviews has become a popular form of protest, and now Chinese players are using reviews to demand localization" (FENLON, 2017). This observation underscores how the Steam review system had been transformed into a recognized instrument of collective bargaining, with players strategically wielding review scores to pressure developers for changes and improvements.

Due to the size and centrality of Steam, more and more episodes of RB were occurring on the platform (Wikipedia Contributors, 2026), indicating that the phenomenon indeed generated visibility and undesirable effects for the platform. In 2017, Steam began developing a system designed to identify and mitigate the impact of such episodes (STEAM, 2019), a topic that will be explored in greater detail in the following section.

However, other platforms were also being RB. In December 2018, the launch of the *Epic Games Store*—a competing digital distribution platform—introduced a new dynamic. Major established titles such as *Metro Exodus* and *Borderlands 3* became exclusive to that store and were removed from Steam, thereby shifting the target of RB to the storefront itself (CHALK, 2019). In 2019, RB also affected games on Metacritic. *Call of Duty: Modern Warfare* received a large number of negative user reviews following controversy over its portrayal of Russian military actions in the campaign (HALL, 2019), while *Pokémon Sword and Shield* was targeted after the announcement that several Pokémon species would be excluded from the game (IGGY, 2019).

The 2020s registered some of the largest RB episodes in the history of the gaming industry. In 2020, *Warcraft III: Reforged* became one of the most emblematic cases, receiving overwhelmingly negative evaluations due to the poor quality of the remaster and the removal of features present in the original game (TASSI, 2020). That same year, *The Last of Us Part II* was targeted by thousands of negative reviews driven by controversies related

to gender and sexuality issues within the game's narrative and characters (LOUREIRO, 2020). In 2021, *Grand Theft Auto: The Trilogy – The Definitive Edition* suffered RB because of its technical shortcomings and graphical issues (JAMES, 2021). In 2022, *Gran Turismo 7* received massive criticism directed at aggressive monetization and changes to game progression (WILLIAMS, 2022). In 2023, *Overwatch 2* became one of the worst-rated games on Steam after controversies related to the cancellation of promised content, the introduction of microtransactions, and conflicts involving Chinese players (CARTER, 2023). Most recently, in 2024, *Helldivers 2* was at the center of one of the largest RB campaigns on the platform when Sony announced the mandatory linking of a PlayStation Network account (LOWRY, 2026). This decision generated over 200,000 negative reviews in just a few days and ultimately led the company to reverse its policy. These cases also illustrate the growing scale of RB campaigns, which have evolved from thousands of negative reviews in the early 2010s to hundreds of thousands of reviews capable of influencing public perception, media coverage, and even corporate decisions (Wikipedia Contributors, 2026).

Beyond video games, RB also manifested in films, series, and other audiovisual media across platforms such as Yelp[1], Metacritic, Rotten Tomatoes[2], and IMDb[3], often intertwined with controversies over social representation, diversity, and ideological disputes. Representative examples include *Ghostbusters* (2016), which was met with negative reviews following the decision to feature an all-female lead cast (LUNA, 2021); *Captain Marvel* (2019), targeted by campaigns linked to public statements made by actress Brie Larson (SIMS, 2019); and *Star Wars: The Last Jedi* (2017), whose reception was marked by intense polarization among fans (MCMAHON, 2020). In the television domain, productions such as *Batwoman* (2019) (GRAMUGLIA, 2019), *The Lord of the Rings: The Rings of Power* (2022) (TASSI, 2022), *The Last of Us* (2023) (PINHEIRO, 2025), and *The Acolyte* (2024) (WATERCUTTER, 2024) were also hit by coordinated negative review campaigns, frequently driven by issues of diversity, representation, or ideological disagreement. These cases demonstrate that even though RB started in video games, it has transcended the video game sector and now broadly affects other forms of mainstream digital entertainment, especially when they become associated with wider cultural and social debates.

A constant force runs through all these episodes: online communities. Rather than consisting of isolated acts of dissatisfaction, RB campaigns have consistently relied on the capacity of groups of users to coordinate collective action through digital platforms (WIKIPEDIA, 2026). Across the cases discussed, communities have used review systems as instruments to express grievances, organize protests, and exert pressure

[1] <https://www.yelp.com/>
[2] <https://www.rottentomatoes.com/>
[3] <https://www.imdb.com/>

on developers, publishers, and platform operators. Over time, these mobilization dynamics have become increasingly rapid and large in scale, culminating in campaigns capable of generating hundreds of thousands of reviews and, in some cases, influencing corporate decisions (LOWRY, 2026). Moreover, the phenomenon is not restricted to a single type of community: gaming communities, fan groups, national player bases, and ideologically motivated audiences have all employed RB to pursue distinct objectives. Thus, while the motivations behind RB vary considerably, its effectiveness fundamentally depends on the existence of organized online communities capable of transforming individual dissatisfaction into coordinated collective action.

In summary, the historical trajectory of RB reveals a phenomenon in which coordinated user evaluations, driven by passionate online communities, are employed as a protest tool, often detached from the intrinsic quality of the product. Throughout this history, Steam has occupied a central and defining role: its review system repeatedly served as both gathering place and amplifier for these communities, fueled by its massive user base and the perceived influence of review scores on sales and reputation (Wikipedia Contributors, 2026).

4.3 Review Bombing Structure

Structurally, RB is not a spontaneous expression of individual dissatisfaction; rather, it usually represents a structured collective action that often precedes the act of rating or even playing the target game (CANTONE; TOMASELLI; MAZZEO, 2025). The phenomenon rarely originates on the target platform. Instead, it typically follows some stages: **Coordination:** The process begins in external "staging areas" such as Reddit, 4chan forums, Discord servers, or social media threads, where a specific narrative is identified and amplified by the community (KUCHERA, 2017; CHESS; SHAW, 2015). **Mobilization:** Next, a "call to action" is issued, effectively redirecting the community collective energy toward specific store pages. This mobilization shifts the activity from external complaints to on-platform intervention (FERGUSON; GLASGOW, 2021; WIKIPEDIA, 2026). **Execution:** The primary mechanism of execution involves mobilized users submitting negative ratings massively to deliberately depress the game aggregate score (KUCHERA, 2017; CANTONE; TOMASELLI; MAZZEO, 2025; WIRZBURGER, 2025).

4.4 Literature and Media Definitions of Review Bombing

As discussed in the previous section, one of the earliest references to RB appears in 2008 (KUCHERA, 2008), which reports that *Spore* experienced an episode of RB but does not provide a clear definition of the phenomenon. As RB became more frequent

and visible across digital platforms, gaming media began to characterize the phenomenon more explicitly.

Media outlets generally describe RB as the coordinated use of user review systems to express dissatisfaction with a product, company, or decision. Wikipedia (WIKIPEDIA, 2026), for example, defines RB as a situation in which a large number of users submit reviews in response to actions taken by creators, publishers, or distributors rather than to the intrinsic quality of the work itself. Similarly, Polygon (KUCHERA, 2017) portrays RB as an organized effort to flood a game with negative reviews in order to protest decisions made by developers, emphasizing the recurring nature of the phenomenon on platforms such as Steam. Medium (SANDERS, 2022) also characterizes RB as a tactic employed by dissatisfied groups seeking to reduce a product's rating while drawing public attention to a specific grievance.

Although the academic literature on RB remains relatively recent and limited, its definitions largely converge with those proposed by the media. (SCHUFF; MUDAMBI; WANG, 2024; WIRZBURGER, 2025) similarly describes RB as the submission of large numbers of negative reviews intended to lower a product's rating, often employing language focused on social, cultural, or political controversies rather than on the product's quality. Likewise, (PONISZEWSKA-MARAŃDA; HYNASIŃSKI; BOROWSKA, 2026) characterizes the phenomenon as a coordinated action in which users rapidly flood a platform with negative reviews motivated by ideological or cultural disputes, aiming to damage the reputation of a product or service. Similarly, (RIBEIRO; PIMENTEL; MELO, 2024; JUNIOR et al., 2025) defines RB as a coordinated attack in which users post large volumes of negative reviews in response to controversies involving a game, its developers, or associated companies, with the objective of influencing public perception and altering the work's reputation. Finally, (CANTONE; TOMASELLI; MAZZEO, 2025) also emphasizes the collective and orchestrated nature of the phenomenon, highlighting the intention to manipulate review metrics and public opinion through concentrated reviewing activity.

Taken together, these definitions reveal a broad consensus: RB is a coordinated effort by a group of users to submit a large volume of reviews within a short period in order to influence a product's reputation, public perception, or evaluation metrics, often in response to controversies that extend beyond the product's intrinsic quality.

4.5 Steam Definition of Review Bombing

Since this work focuses on Steam's RB labels, it is important to consider the platform's own definition of the phenomenon. Steam defines RB as a situation in which players post a large number of reviews within a short period of time with the intention of lowering a game's Review Score (STEAM, 2019). The platform further identifies a subset of these

episodes as *off-topic review activity*, in which the ***"content of the reviews centers on issues considered unrelated to whether future purchasers are likely to enjoy the game, and therefore should not meaningfully influence the aggregate score"***. Within Steam's moderation framework, periods classified as off-topic review activity are treated as instances of RB and may be excluded from the calculation of the game's overall rating.

Among the platforms discussed in this study that have experienced RB episodes, such as Metacritic, Amazon, and the Epic Games Store, Steam is the only one that provides an explicit formal definition of RB and an associated moderation taxonomy. The other platforms acknowledge or are reported to experience RB phenomena, but do not publicly specify a definition or formal criteria equivalent to Steam's off-topic review classification.

To identify such events, Steam uses a hybrid automated–human moderation pipeline. Valve maintains an internal monitoring system that continuously tracks review activity across the store and automatically flags titles that show abnormal surges in review volume over short time windows (STEAM, 2019). According to Steam, this automated layer does not attempt to determine the underlying cause of the spike; instead, it serves only as a trigger for manual inspection. Once flagged, a specialized moderation team evaluates the activity, deciding whether the surge reflects coordinated off-topic behavior or not. If confirmed, the corresponding interval is marked as a period of off-topic RB activity, and the game is publicly labeled as having undergone an off-topic RB, as illustrated in Figure 3, which presents the distribution of reviews during the years for the game *The Witcher 3*, where we can see the positive reviews in blue and negative in red.

4.6 Operational Definition of Review Bombing

In this study, the operational definition of RB is not based exclusively on Steam's internal characterization of the phenomenon, allowing it to be applied to other gaming review platforms as well. This choice is motivated by the fact that platform specific definitions may reflect the political and economic positions of the industry, particularly in how certain forms of user behavior are framed, classified, or excluded from aggregate scoring mechanisms. Relying solely on such a definition could therefore introduce a normative bias into the analysis.

Therefore, we adopt a broader operational definition grounded in academic, media, and industry perspectives. Under this definition, RB refers to a coordinated surge of negative user reviews posted within a relatively short time window with the intention of influencing a product's perceived quality, reputation, or aggregate rating, rather than strictly reflecting the intrinsic experience of the product. This behavior is typically

associated with collective mobilization processes and may arise from disputes involving developers, as well as from broader cultural, ideological, or community driven conflicts.

5 Data and Methodology

This chapter outlines the datasets and methodology used in the study, including the development of a large-scale collection of Steam user reviews and the identification of RB episodes. It describes the scope and temporal coverage of the review dataset, explains how Steam-labeled RB episodes were collected, and details the filtering and validation steps applied to align them with our definition of RB. In addition, the chapter presents the methodological framework adopted for the analysis of reviews, users, and games, as well as the machine learning methodology developed for automated RB detection. Figure 4 summarizes the complete data collection and processing pipeline, while Figure 5 summarizes the RB classifier pipeline.

Figure 4 – Overview of data collection methodology.

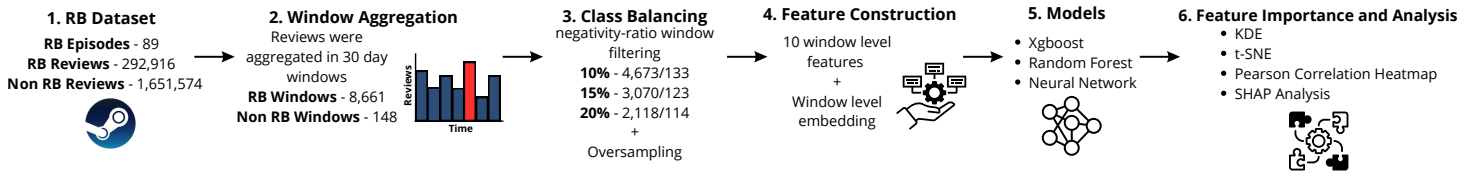


Figure 5 – Overview of automated RB detection via machine learning methodology.

5.1 Steam User Review Dataset

To construct our dataset, we adopted two complementary data-access strategies. The first leverages the official Steam Web API,[1] which returns structured JSON responses and is therefore easy to process. The second relies on direct HTTP requests paired with HTML parsing, using the Python libraries `requests`[2] and `BeautifulSoup`[3]. Whenever possible, we favored the API for its practicality. However, when the API imposed rate limits that were too restrictive for large-scale crawling (e.g., fetching millions of reviews) or when the required information was not exposed by any API endpoint (e.g., the off-topic RB tag), we resorted to `requests` and `BeautifulSoup`.

[1] <https://developer.valvesoftware.com/wiki/Steam_Web_API>
[2] <https://pypi.org/project/requests/>
[3] <https://beautiful-soup-4.readthedocs.io/en/latest/>

Due to the scale of the collection process, data acquisition was divided into stages. The first stage consisted of collecting all games available on Steam. For this purpose, we used the API endpoint <https://api.steampowered.com/IStoreService/GetAppList/v1/&max_results=500&last_appid=0>, which returns the application identifier (`appid`) and name of each game. The endpoint supports pagination through the `last_appid` parameter, allowing iterative retrieval of batches of games until the entire game catalog has been collected. This process resulted in a list of 157,825 games.

Because this endpoint provides only the application identifier (`appid`) and title, additional metadata had to be collected separately. Since no endpoint in the official Steam API provides information about whether a game experienced an RB episode, this stage relied on the `requests` and `BeautifulSoup`. Therefore, for each collected `appid`, we generated the corresponding store page URL (e.g., <https://store.steampowered.com/app/(appid)>) and retrieved its HTML content. We then parsed the page to extract the information illustrated in Figure 1, including: **A** the game title, **B** the total number of recent reviews, **C** the total number of overall reviews, **D** the release date, **E** the developer, **F** the publisher, and **G** the game genres. Additionally, from the information shown in Figure 3, we extracted Steam's RB activity warning, including its start date, end date, and presence status.

Among the 157,825 games collected during the first stage, we removed all entries with zero overall reviews, resulting in a final set of 83,420 games. Games without reviews cannot have experienced an RB episode, making them irrelevant to the objectives of this study. Although additional metadata were collected from the store pages, they were not directly used in the analyses presented in this work.

The next stage consisted of collecting all user reviews associated with the filtered set of games. Given the volume of requests required, we also did not use the Steam API for this task. Instead, for each of the 83,420 games, we queried the endpoint <https://steamcommunity.com/app/(appid)/homecontent/?userreviewscursor=>, which returns the first ten reviews associated with a game. Steam implements cursor-based pagination for review retrieval. Each response contains both the current cursor and the cursor required to retrieve the next batch of reviews. By recursively following these cursors, it is possible to iterate through the complete review history of a game until no additional reviews remain.

For each review, we extracted the information illustrated in Figure 2: **A** the binary recommendation label (recommended or not recommended), **B** the total playtime prior to submitting the review, **C** the review publication date, **D** the review text, and **E** the username. Through this process, we collected a total of **103,663,785 reviews** authored by **37,636,245 unique users**.

Data collection was conducted between July 2024 and February 2025. To the best of

our knowledge, this dataset represents the largest publicly documented collection of Steam user reviews currently available in the academic literature. The collection procedure did not violate Steam's Terms of Service, as only publicly accessible information was retrieved. In accordance with Steam's Terms of Service, user identifiers were anonymized prior to all analyses. Table 2 summarizes the collected metadata.

Metadata	Data Type	Description
`appid`	Integer	Unique identifier assigned by Steam to each game.
`game_title`	String	Official title of the game as displayed on the Steam store page.
`recent_reviews`	Integer	Number of reviews considered by Steam in the recent review summary.
`overall_reviews`	Integer	Total number of reviews received by the game.
`release_date`	Date	Official release date of the game on Steam.
`developer`	String	Name of the game developer.
`publisher`	String	Name of the game publisher.
`genres`	List[String]	List of genres associated with the game.
`rb_flag`	Boolean	Indicates whether Steam marked the game with an off-topic RB warning.
`rb_start_date`	Date	Starting date of the RB episode identified by Steam.
`rb_end_date`	Date	Ending date of the RB episode identified by Steam.
`recommended`	Boolean	Binary recommendation label of a review (recommended or not recommended).
`review_playtime`	Float	Total playtime accumulated by the user before posting the review, measured in hours.
`review_date`	Date	Publication date of the review.
`review_text`	Text	Written content of the review.
`username`	String	Steam username of the review author.

Table 1 – Summary of the collected Steam dataset metadata.

5.1.1 Supplementary Metadata Collection

Beyond the game and review data used in the analyses presented in this work, we collected additional user-level, social, and gameplay metadata used in other works.

For the **37,636,245 unique users** identified in the review dataset, we queried the Steam Web API endpoint <ISteamUser/GetPlayerSummaries>. This process retrieved each user's Steam identifier, username, profile URL, profile description, country, state, and city codes, current status, user banned status. User level information was collected through <IPlayerService/GetSteamLevel>, while ban information was obtained through <ISteamUser/GetPlayerBans>. Among all identified users, only **21,513,415** had public profiles, allowing these metadata to be retrieved.

For users with public profiles, we collected profile comments through the Steam Community endpoint <https://steamcommunity.com/comment/Profile/render/(steamid)>, which returns paginated comments posted on a user's profile page. Steam allows users to post public messages directly on another user's profile page. For each accessible profile,

we retrieved the comment text, author identifier, author username, publication date, and comment identifier. This process resulted in **138,037,843 profile comments** collected from **3,314,742** users whose profiles and comment sections were publicly accessible.

We also collected friendship information through the Steam Web API endpoint <ISteamUser/GetFriendList>. For each accessible account, we retrieved the identifiers of all friends. Because friend lists may be private, this information was available for only **14,950,537** users.

Finally, we collected user game libraries through the endpoint <IPlayerService/GetOwnedGames>. For each owned game, the retrieved information included the application identifier, total playtime, playtime during the previous two weeks, and ownership metadata provided by the API. Due to the large number of requests required, this collection was restricted to Brazilian users, resulting in **387,709 user game libraries** being collected.

Although these supplementary metadata were gathered during the data collection process, they were not used in the analyses presented in this work and are therefore not discussed further.

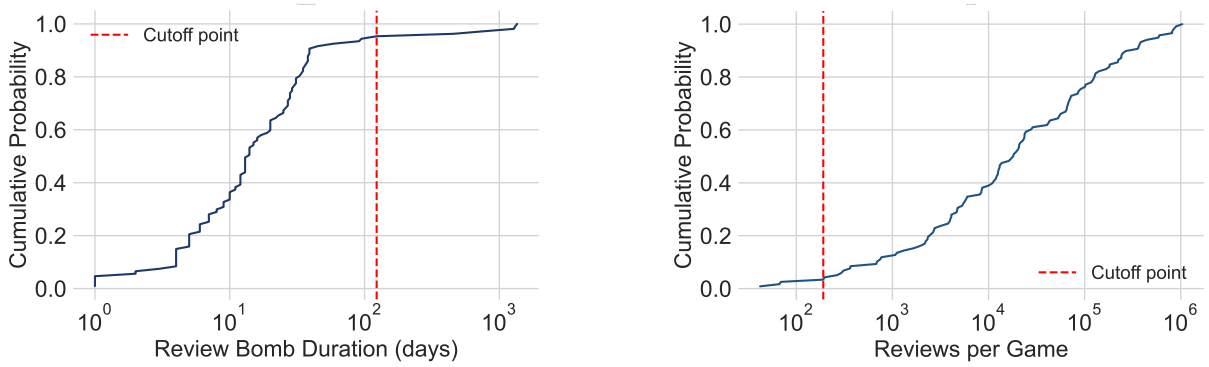
5.2 Review Bombing Labeling

As previously discussed, Steam employs a hybrid automated–human moderation pipeline to identify RB and publicly flags affected titles by highlighting periods of "off-topic" review activity. We utilize these platform-provided labels as the primary ground truth to build our initial dataset, resulting in 118 flagged games (Figure 3). However, relying solely on Steam's "off-topic" flag presents a potential methodological limitation. The platform's moderation algorithms and internal criteria for defining what exactly consists a off-topic activity remain largely opaque. Consequently, these labels inherently reflect Steam's proprietary and institutional conceptualization of RB. This black-box approach may overlook coordinated manipulation that evades platform detection, while simultaneously capturing instances that diverge from the operational definition of RB adopted in this study. Despite these constraints, Steam remains the most comprehensive, publicly accessible repository of labeled RB incidents in digital games, offering a valuable environment for large-scale empirical analysis.

To mitigate these platform specific biases, align the dataset with the operational definition of RB adopted in this study, and ensure data integrity, we implemented an additional step of manual validation following the initial automated collection. We examined all reviews posted between the RB start and end dates provided by Steam for each flagged game to confirm the presence of genuine RB behavior. Specifically, we aimed to identify concentrated off-topic grievances and complaints, such as targeted personal

attacks against developers, explicit political statements, or the presence of hate speech against vulnerable groups. Through this qualitative filtering, we identified 7 games where no consistent off-topic narrative could be established, making it impossible to confirm a clear RB motivation. For example, during the RB flagged period for *Tools Up!*, there was, in fact, a wave of negative reviews, but they mostly remained focused on core gameplay aspects, such as the lack of multiplayer functionality, rather than exhibiting genuine off-topic coordination of an RB.

Furthermore, we also identified 16 games, such as *Bless Global* and *Nova Empire*, in which the flagged review activity was mostly composed of positive reviews intended to artificially boost the game's visibility or reputation, often associated with platform incentives or commissioned reviewing practices. Since the core definition of RB adopted in this study involves coordinated activity aimed at **negatively** impacting a game's review score or public perception, these positive cases were also excluded from our dataset.



(a) CDF of total duration of each RB

(b) CDF of total reviews per game

Figure 6 – CDF analyses summarizing the total duration of RB and total reviews of RB games

Next, to refine our RB dataset, we evaluated whether each RB event instance met a minimum review volume to further analyses. Figure 6b shows the distribution of review volume per game. Approximately 5% of the games had fewer than 300 reviews in total. Games such as *The Adventures of Looppy*, with only 67 lifetime reviews, and *Chicken Bomb* with only 6 reviews, cannot meaningfully exhibit a coordinated spike in activity, falling outside any reasonable definition of RB. Therefore, all games with fewer than 300 total reviews were removed, corresponding to only 2 games removed. Additionally, the fact that *Chicken Bomb* was flagged as having experienced an RB episode despite having only six reviews highlights a limitation of Steam's detection system. Even with human review, a game with such a small review volume cannot reasonably exhibit the large-scale activity associated with the RB phenomenon, suggesting the presence of false positives.

Complementarily, we also examined the temporal extent of RB periods. As shown in Figure 6a, most RB time intervals flagged by Steam lasted between 10 and 100 days, consistent with prior descriptions of RB as a short-lived phenomenon (CANTONE;

TOMASELLI; MAZZEO, 2025). However, 4 games were flagged as being engaged in ongoing RB for over 3 years (e.g., *The Witcher: Enhanced Edition Director's Cut*, *The Witcher Adventure Game*, *Book of Demons* and *Shinobi Warfare*). Likely reflecting labeling errors rather than genuine sustained RB. We excluded all episodes with RB windows longer than 123 days, corresponding to the top 5% of the distribution (4 games).

After applying these validation and filtering steps, our final cleaned RB dataset contained **89 RB episodes**, consisted of **89 games**, with a total of **11,800,720 reviews** and **7,486,387 unique users**. From these, all the reviews posted during a RB episode are considered RB reviews, resulting in **292,916 RB reviews**. Due to privacy constraints, the dataset is not publicly released, and only aggregated and anonymized statistics are reported in this study.

5.3 Methodology for Understanding Review Bombing Dynamics

After developing the dataset, we conducted a multi-layered analysis focusing on reviews, users, and games. The objective of this stage was to characterize the RB phenomenon and identify behavioral differences between RB and non-RB activity on Steam.

First, we investigated the temporal dynamics of RB episodes. This analysis included the yearly distribution of RB episodes, the volume of reviews generated during each episode, the temporal relationship between game release and RB occurrence, and the evolution of negative review rates over time. Time-series visualizations were employed to identify recurring patterns and anomalous peaks associated with RB activity.

Second, we analyzed review-level characteristics. We examined the proportion of reviews generated during RB episodes, the distribution of recommendation rates, and textual properties such as review length. Cumulative Distribution Functions (CDFs) were used to compare RB and non-RB review distributions.

Third, we investigated user engagement and participation patterns. We compared users involved in RB episodes with non-RB users using metrics such as playtime at review submission, number of reviewed games, participation across multiple RB episodes, and historical recommendation behavior. Distributional differences were evaluated using descriptive statistics and the Kolmogorov–Smirnov test.

Finally, we analyzed the characteristics of games affected by the RB phenomenon. This analysis included genre distributions, developer distributions, review volumes, and the recurrence of RB episodes across games from the same developer.

5.3.1 Topic Modeling via BERTopic

To better understand the complaints and narratives expressed during RB episodes, we separated reviews into positive and negative subsets and trained an independent topic model for each group. This separation allows us to analyze whether positive and negative reviews exhibit distinct semantic patterns, motivations, and forms of discourse during coordinated review activity.

To perform this analysis, we applied a transformer-based topic modeling pipeline using **BERTopic** (GROOTENDORST, 2022). The pipeline consisted of the following stages:

(i) Embeddings: Given the global nature of Steam, we utilized the `paraphrase-multilingual-MiniLM-L12-v2` model to generate dense vector representations of reviews across multiple languages (REIMERS; GUREVYCH, 2019). These embeddings encode semantic information by placing reviews with similar meanings closer in vector space.

(ii) Dimensionality Reduction: We applied **UMAP** (MCINNES, 2018) to project embeddings into a lower-dimensional space. This step reduces computational complexity while maintaining meaningful semantic neighborhood information and improving cluster separability.

(iii) Clustering: Then, we applied **HDBSCAN** (MCINNES; HEALY; ASTELS, 2016) to identify dense semantic clusters. HDBSCAN automatically discovers groups with varying densities while identifying outlier instances as noise, a property particularly advantageous for user-generated review data.

After clustering, BERTopic derives representative terms from each cluster to construct interpretable topic representations. Because topics included multiple languages, we used `Gemini 1.5 Pro` to translate representative keywords and assist in generating descriptive topic labels from the semantic clusters.

5.4 Automated Review Bombing Detection Via Machine Learning

The final stage of this work consists of developing a machine learning framework capable of automatically identifying RB episodes. Beyond determining whether a game has experienced RB, our objective is also to identify when such episodes occur and to understand which semantic and temporal characteristics contribute most to the classification process. Figure 5 summarizes the complete machine learning pipeline adopted in this study.

5.4.1 Temporal Formulation and Window Construction

To transform the RB labels provided by Steam into a supervised learning problem, we represent each game as a sequence of temporal windows. Rather than treating RB as a game-level phenomenon, we model it at a time-window level. A game-level formulation simplifies the problem by assuming that all reviews associated with a RB game belong to the RB episode, which may incorrectly label legitimate reviews posted outside the affected period as RB-related. Therefore, reviews are grouped according to their publication dates and aggregated into fixed-size time intervals.

This periodic formulation allows the classifier to operate at a time-window level, enabling the identification of specific periods associated with RB activity. Moreover, it introduces a temporal structure in which windows are analyzed as consecutive segments of a game's review history. Since RB typically emerges as a sudden disruption in reviewing behavior, this representation allows the model to capture temporal variations in both review content and activity patterns. An additional advantage is that the classifier can identify not only whether a game experienced RB, but also when the event occurred. Consequently, the temporal window representation provides a more informative and practically useful formulation than a single game-level classification.

Steam already provides, for each flagged game, the exact start and end dates of RB episodes, as illustrated in Figure 3. However, this precise temporal information would not be available when applying the classifier to unseen games in a real-world scenario. Therefore, instead of directly modeling the exact RB intervals, we construct generic fixed-size temporal windows over the entire review timeline of each game. A window is labeled as RB (class 1) if it contains at least one day overlapping a Steam-identified RB interval; otherwise, it is labeled as non-RB (class 0). This process transforms the continuous stream of reviews into a sequence of fixed-size temporal windows, as illustrated in Figure 7.

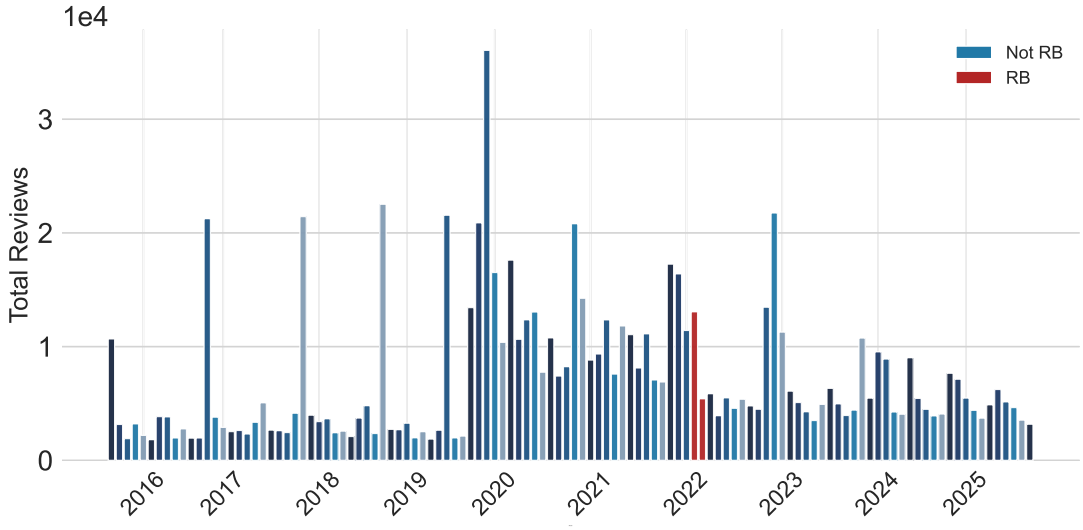


Figure 7 – Example of 30 days window segmentation on the game *The Witcher 3*.

5.4.2 Window Size Selection

The next step is to determine an appropriate window size. Each window should capture as many RB comments as possible while including the fewest non-RB reviews to minimize contamination. We initially explored fixed temporal segmentation strategies by partitioning each game's review timeline into windows of 10, 20, and 30 days, which, according to Figure 6a, should cover most RB episodes. While straightforward, this approach brings a tradeoff between the integrity of RB episodes and the contamination added to RB windows.

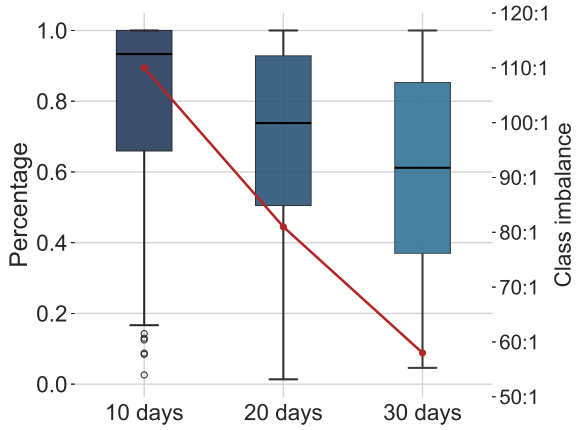


Figure 8 – Box plot of the proportion of RB reviews over total reviews for each day within each window.

Figure 8 illustrates the relationship between window size and label contamination: as the temporal window increases, a larger number of non-RB reviews are inevitably aggregated into windows labeled as RB. This effect is reflected in the differing distributions across window sizes. In particular, the 10-day window exhibits a high median purity of RB content (0.933), whereas the 30-day window shows a substantially lower median value (0.612), indicating higher contamination. Despite this lower purity, the 30-day window was adopted in the experiments because it consistently achieved a better classification performance in our tests. We also evaluated larger window sizes (40, 50, and 60 days), but they consistently produced inferior results and were therefore omitted from the subsequent analyses.

5.4.3 Class Imbalance and Data Processing

A major challenge associated with this formulation is the severe class imbalance, driven both by the relatively small number of RB periods compared to the overall lifespan of games and by the imbalance in the number of affected titles: among more than 150,000 games, only 89 experienced RB episodes. Since titles such as *The Witcher 3* remain active for more than 10 years while RB episodes are typically short-lived (e.g., less than a month), the resulting datasets contain substantially more non-RB windows than RB

windows. Under these conditions, classifiers tend to become biased toward the majority class (class 0), increasing the risk of overfitting and reducing their ability to generalize to RB instances.

Although shorter windows preserve higher purity of RB reviews, they also produce more imbalanced datasets and greater overfitting toward the majority class, while longer windows reduce this skew at the cost of incorporating non-RB reviews into RB windows. The selected 30-day window resulted in a dataset containing 8,809 temporal windows, of which 8,661 were labeled as non-RB and 148 as RB, illustrating the class imbalance inherent to the RB detection task. To mitigate this imbalance, we adopted two complementary strategies, rather than relying solely on resampling to correct it.

First, we applied a negativity-ratio filtering procedure (Figure 9). Since RB episodes are commonly characterized by unusually high concentrations of negative reviews, windows with less than 10%, 15%, and 20% negative reviews were progressively removed, thereby focusing on periods more likely to contain coordinated attacks. This filtering yielded class ratios of (4,673! :!133), (3,070! :!123), and (2,118! :!114), respectively.

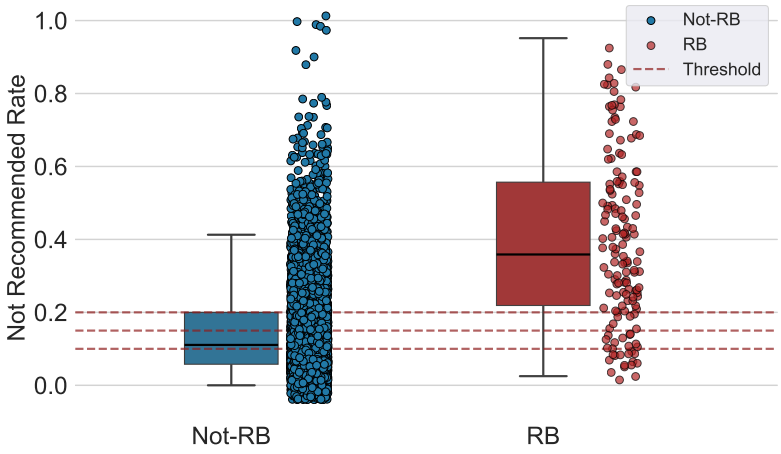


Figure 9 – Lower bound threshold filters of ratio of negative reviews for each time window.

Second, we evaluated different data resampling strategies: undersampling using Tomek Links, oversampling using SMOTE, and hybrid combinations of both methods. To avoid information leakage, all resampling procedures were applied exclusively to the training data, while the test folds remained unchanged.

5.4.4 Feature Representation

For semantic representation, every review was encoded using the multilingual sentence-transformer model `paraphrase-multilingual-MiniLM-L12-v2`[4], generating a 384-dimensional dense vector. The embeddings of all reviews belonging to the same tem-

[4] <https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2>

poral window were averaged to obtain a single semantic representation of the corresponding period.

In addition to the embedding representation, we extracted ten handcrafted features designed to characterize semantic concentration, temporal evolution, review polarity, and review volume. As no established feature set exists in the RB literature for temporal episode-level detection, these features were designed based on patterns observed throughout our exploratory analyses of RB episodes.

First, we included the proportion of negative reviews (`not_recommended_ratio`) within each temporal window. Since RB events are characterized by sudden increases in negative feedback, this feature provides a direct signal of abnormal review polarity concentration. Higher values indicate a stronger concentration of negative reviews, whereas lower values reflect predominantly positive feedback.

To measure semantic dispersion within a window, we computed the average variance across embedding dimensions (`embedding_variance`). This feature measures how much the embeddings of individual reviews deviate from one another within the same window. Lower values indicate that reviews tend to be semantically similar and concentrated around a common topic, whereas higher values suggest greater diversity in the discussed themes. Since RB campaigns are often driven by a specific grievance repeatedly expressed by many users, RB windows are expected to exhibit lower semantic variance than periods of regular review activity.

We also modeled textual concentration using a clustering-based feature (`cluster_density`). For each temporal window, review embeddings were grouped using MiniBatch K-Means clustering (SCULLEY, 2010), a scalable variant of the traditional K-Means algorithm that updates cluster centroids using small randomly sampled subsets (mini-batches) of the data rather than the full dataset at each iteration. This approach substantially reduces computational cost while preserving clustering quality comparable to standard K-Means, making it suitable for large collections of review embeddings. The number of clusters was dynamically defined as a function of the number of reviews in the window. The feature corresponds to the proportion of reviews assigned to the largest cluster. Higher values indicate that a substantial proportion of reviews concentrates around a unique dominant semantic theme, whereas lower values suggest discussions distributed across multiple topics.

To further capture redundancy and repetitive discourse, we computed the proportion of highly similar review pairs (`pct_highly_similar_texts`). This metric was estimated through pairwise cosine similarity comparisons between review embeddings inside the same temporal window. Previous work by (MARTES et al., 2025) evaluated multiple cosine similarity thresholds for transformer-based semantic similarity models, reporting optimal values ranging from approximately 0.61 to 0.87 depending on the underlying ar-

chitecture. Based on these findings, we adopted a more conservative threshold of 0.90 to capture only highly redundant or near-duplicate semantic content. The resulting feature represents the proportion of highly similar review pairs relative to the total number of evaluated pairs. Elevated values indicate repetitive semantic patterns, whereas values near zero indicate more diverse textual content. At a higher level, this feature captures the extent to which users converge on a common discourse, which is expected to increase during RB episodes.

In addition to intra-window characteristics, we introduced temporal semantic continuity features. Specifically, each window embedding was compared with the embeddings of its adjacent temporal windows using cosine similarity. The resulting features, `similarity_with_previous` and `similarity_with_next`, aim to quantify how semantically consistent a given period is relative to its immediate temporal context. Higher values indicate strong semantic continuity across consecutive periods, whereas lower values suggest abrupt changes in discussion topics that may indicate the emergence or conclusion of RB events.

We also incorporated two global semantic deviation features. The first, `cosine_dist_mean_id`, measures the average cosine distance between the current window and all other temporal windows of the same game, while `cosine_dist_centroid_id` computes the cosine distance between the current window and the semantic centroid of the game's entire review history. Both features quantify how much a given period deviates from the game's typical review discourse. Higher values indicate greater semantic divergence from the historical discussion, whereas lower values suggest that the reviewed topics remain consistent with the game's overall review history.

Finally, we incorporated two temporal and volume-based features. The first, `days_since_launch`, measures the number of days elapsed between the first temporal window of the game and the current window. Higher values correspond to periods further from the game's release date, whereas lower values represent early stages of the review lifecycle. The second, `period_entries_percentage`, represents the proportion of reviews contained in the current temporal window relative to the total number of reviews associated with the game. Since RB episodes frequently manifest as short-lived spikes of unusually concentrated activity, this feature captures abrupt increases in review volume localized within specific periods. Higher values indicate periods responsible for a large share of the game's reviews, whereas lower values represent comparatively inactive intervals. Table 2 summarizes these features.

5.4.5 Baseline and Alternative Configurations

As a reference scenario, we also evaluated a baseline representation based exclusively on individual review embeddings. Unlike the proposed temporal approach, the

Feature	**Description**
Embedding (384 dims.)	Mean sentence embedding of all reviews in the temporal window generated by the `paraphrase-multilingual-MiniLM-L12-v2` model.
not_recommended_ratio	Proportion of negative reviews within the temporal window.
embedding_variance	Average variance across embedding dimensions, measuring semantic dispersion among reviews.
cluster_density	Proportion of reviews assigned to the largest MiniBatch K-Means cluster within the window.
pct_highly_similar_texts	Proportion of review pairs with cosine similarity greater than 0.90.
similarity_with_previous	Cosine similarity between the current window embedding and the previous temporal window.
similarity_with_next	Cosine similarity between the current window embedding and the next temporal window.
cosine_dist_mean_id	Average cosine distance between the current window and all other windows of the same game.
cosine_dist_centroid_id	Cosine distance between the current window and the semantic centroid of the game's review history.
days_since_launch	Number of days elapsed since the first temporal window of the game.
period_entries_percentage	Proportion of reviews contained in the current window relative to the total reviews of the game.

Table 2 – Summary of the features used to represent each temporal window.

baseline does not aggregate reviews into windows and does not incorporate any temporal, structural, or handcrafted features. Reviews posted during Steam-labeled RB periods were assigned to class 1, while all remaining reviews were assigned to class 0. Each review was independently encoded using `paraphrase-multilingual-MiniLM-L12-v2`. This baseline allows us to assess whether semantic embeddings alone are sufficient for RB detection and to quantify the contribution of the proposed temporal representation and engineered features. Since most prior studies on machine learning for RB detection either examine single RB episodes or only tangentially engage with the problem, there is a lack of widely adopted baselines for large-scale RB detection; throughout this work, we refer to this approach as the *Baseline*.

According to the operational definition of RB adopted in this study, RB is characterized by coordinated negative review campaigns. Therefore, we also evaluated an alternative experimental setting in which all feature extraction procedures were performed exclusively on negative reviews. Under this configuration, positive reviews were discarded before feature construction, and all the features described in Section 5.4.4 were computed over the remaining subset. This experiment was designed to investigate whether removing potentially unrelated positive reviews improves the ability of the classifiers to identify RB patterns. Note that the negativity-ratio filtering strategy described in Section 5.4.2 is a distinct procedure, applied at the window level to remove entire windows with low negativity; it depends on aggregated window-level statistics and is therefore not applicable to the individual review-level baseline representation.

5.4.6 Models and Evaluation

We evaluated three supervised learning models: XGBoost[5], Random Forest [6], and a Feedforward Neural Network[7]. Hyperparameter optimization was performed using Optuna. All experiments were conducted on a workstation equipped with an 8-core AMD Ryzen 7 5800H processor with Radeon Graphics, 16 GB of RAM, and an NVIDIA GeForce GTX 1650 GPU with 4 GB of dedicated VRAM.

Classifier performance was evaluated through 5-fold cross-validation. Because overall accuracy can be misleading under highly imbalanced conditions, we report precision, recall, and F1-score separately for both classes, with particular emphasis on the RB class.

5.4.7 Feature and Model Analysis

Finally, to better understand the characteristics of the engineered feature space and the behavior of the proposed classifiers, we employed several complementary analysis techniques. Kernel density estimation (KDE) was used to examine the distribution of individual features and identify differences between RB and non-RB windows. t-SNE projections were generated to provide a visual assessment of class separability in the high-dimensional feature space. Pearson correlation analysis was performed to identify linear relationships and potential redundancies among engineered features. Finally, SHAP was applied to investigate which features most strongly influenced model predictions, providing insights into the decision-making process learned by the classifiers.

[5] Implemented using the Python `xgboost` package: <https://xgboost.readthedocs.io/en/stable/python/python_intro.html>.

[6] Implemented using `sklearn.ensemble.RandomForestClassifier`: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>.

[7] <https://pytorch.org/>.

6 Results of Dynamics of Review Bombing

In this chapter, we address RQ1 and RQ2 by examining how RB unfolds on Steam from three perspectives: the reviews produced, the users involved, and the targeted games.

To directly address RQ1, our panoramic and empirical analysis reveals a remarkably consistent structural signature across the 89 validated episodes. At scale, RB does not manifest merely as random platform noise or sustained, long-term harassment; rather, it emerges as highly concentrated, temporally bounded bursts of negative feedback that generated nearly 300,000 reviews. These coordinated waves share unifying structural traits: they are overwhelmingly post-release phenomena (with 87.6% occurring more than a year after launch) and are directed primarily at "Action" and "Indie" titles. Furthermore, rather than being driven by newly created bot accounts or a small core of persistent attackers, these campaigns are executed by established platform users who mobilize situationally.

After outlining these unifying patterns, the following subsections address **RQ2** by demonstrating how these RB periods distinctly contrast with organic non-RB feedback within the platform. Specifically, we detail the temporal dynamics to investigate when episodes occur and their concentration (Section 6.1), compare RB participants against standard reviewers in terms of activity and engagement (Section 6.2), and explore the characteristics of affected games, including genres, developers, and review volume (Section 6.3).

6.1 Temporal Dynamics and Patterns of Review Bombing

Figure 10 illustrates the temporal distribution of RB episodes and the review volume of affected games. Normal reviews are shown in blue, and reviews during RB episodes in red. The first RB episodes appear in 2018, reflecting Steam's RB detection system introduced in 2017.

After this, an irregular increase in RB episodes can be observed, with 2 occurrences in 2018, 9 in 2019, and 4 in 2020. This period is followed by a pronounced spike in 2022, when the number of episodes rose from 11, in 2021, to 38 in 2022. A plausible explanation is the geopolitical context of the Russia-Ukraine conflict. During this period, several developers publicly took positions on the conflict, which triggered backlash from portions of the player base, as documented in episodes such as *Book of Demons* (ADAMS, 2022a) and *Cyberpunk 2077* (WALKER, 2022). However, a higher number of RB episodes does not necessarily correspond to a larger total volume of reviews. For example, despite having only 9 RB episodes in 2023, the total number of RB reviews reached 130,743,

exceeding the 77,847 reviews from 2022. This increase was primarily driven by the RB incident involving War Thunder (PLUNKETT, 2023), which generated more than 100,000 reviews.

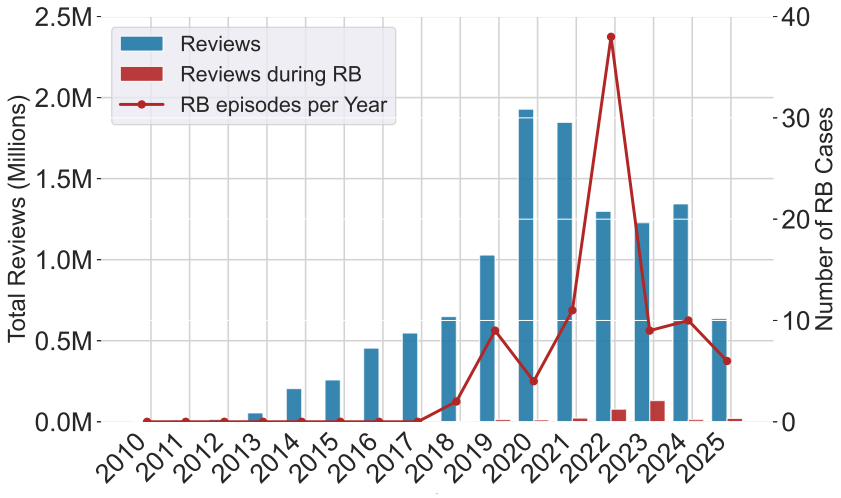


Figure 10 – Total of reviews of RB games per year and total of RB episodes per year.

Figure 11 provides a comprehensive panoramic view of the relative temporal dynamics of RB across all 89 affected titles. By ordering the games according to their total lifespan on the platform and plotting the proportion of negative reviews (y-axis) over time in months (x-axis), a distinct visual pattern of RB emerges. The baselines demonstrate that, for the vast majority of their lifespans, these games maintain relatively stable levels of negative feedback. However, the flagged month with a RB episode (marked in red) align perfectly with acute spikes in negativity. This visual evidence reinforces that RB does not represent a gradual shift in community opinion or a steady decline in game quality, but rather an anomalous and highly concentrated burst of coordinated action.

Furthermore, the distribution of the RB markers reveals a striking temporal disconnect between a game's release and its targeted attack. Visually, we noted very few instances of RB around zero-month mark (on the release of the game). That illustrates that, on Steam, RB is predominantly a post-release phenomenon. As the data shows, 87.6% of these episodes occur more than a year after launch. This long delay indicates that RB campaigns rarely occur in the initial state of the product, such as launch bugs or poor optimization. Instead, they work as reactive movements triggered by late-stage developments, such as controversial updates, DLCs, sudden monetization changes, or external sociopolitical and geopolitical events involving the developers.

Finally, the longitudinal ordering in Figure 11 highlights that a game's longevity and established community offer no immunity against coordinated backlash. The attacks are distributed across the entire spectrum of game lifespans, from more recently released titles at the top of the distribution, such as *KartRider: Drift* with only 11 months of lifespan, to legacy games at the bottom. Extreme cases, such as *The Witcher 3*, demonstrate

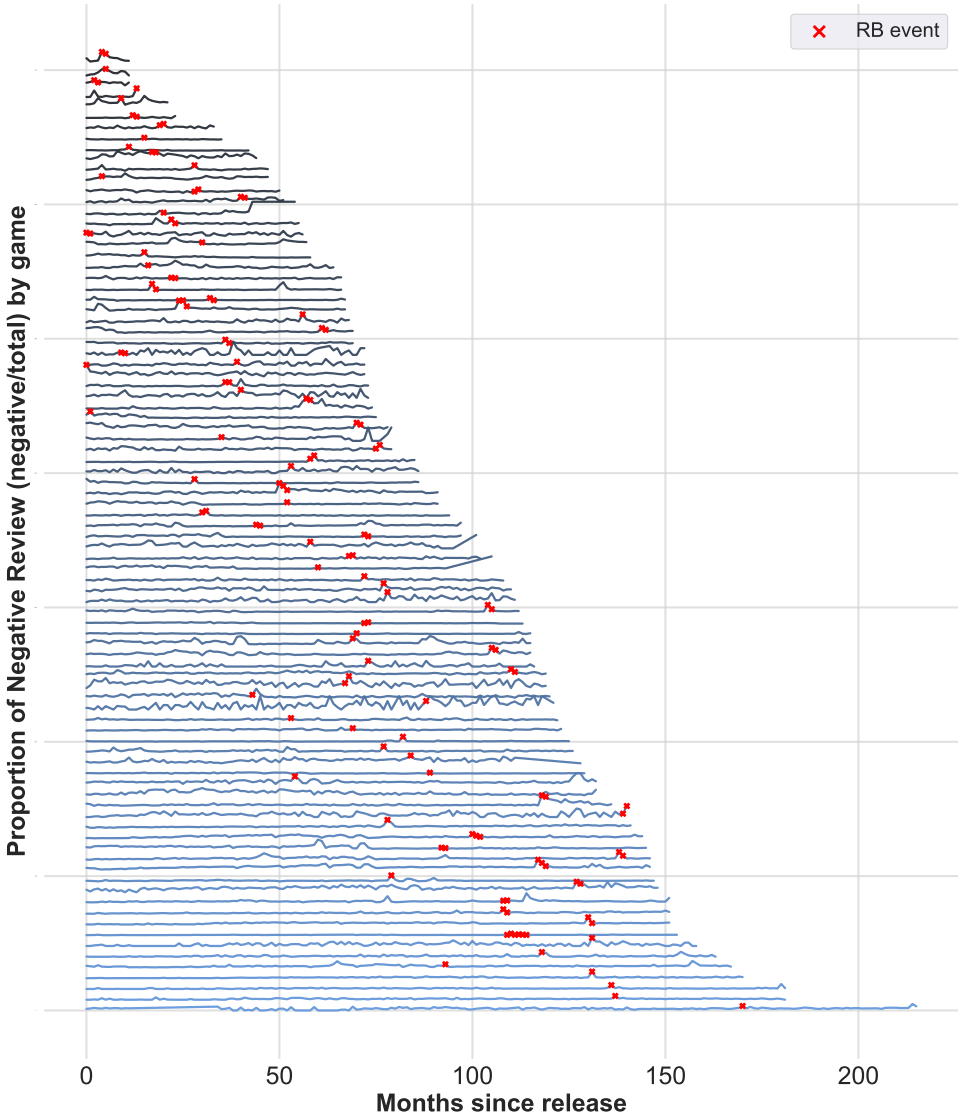
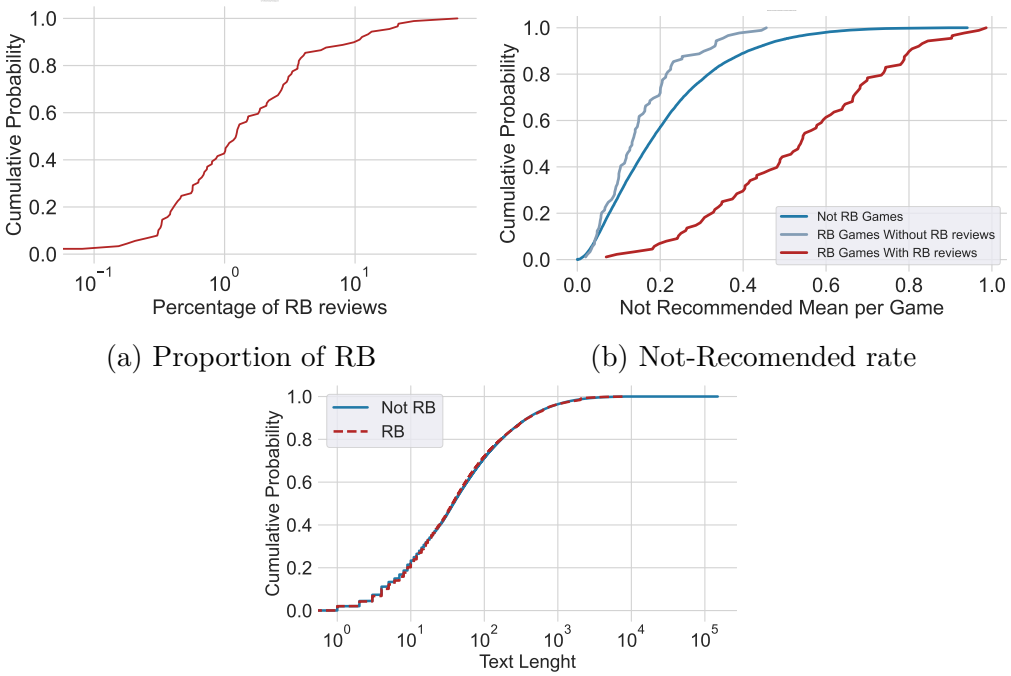


Figure 11 – Distribution of negative review peaks in RB games

that a game can exist peacefully on the platform for nearly a decade before a sudden external incident provokes an unprecedented wave of negative reviews. This underscores the volatility of review infrastructures and the unpredictable nature of digitally mediated protests.

6.2 Review Bombing Participant Profiles and Engagement

To better understand RB, it is essential to examine who participates in these events and whether their behavior differs from that of typical Steam users. Figure 12a shows the proportion of a game's total reviews posted during RB episodes. For 40% of games, RB reviews account for less than 1% of all reviews. While small, this is expected given many games' long lifespans. Even so, RB episodes can involve substantial activity: in *Halo Infinite*, the RB period represents only 0.25% of reviews but corresponds to 898



(a) Proportion of RB

(b) Not-Recomended rate

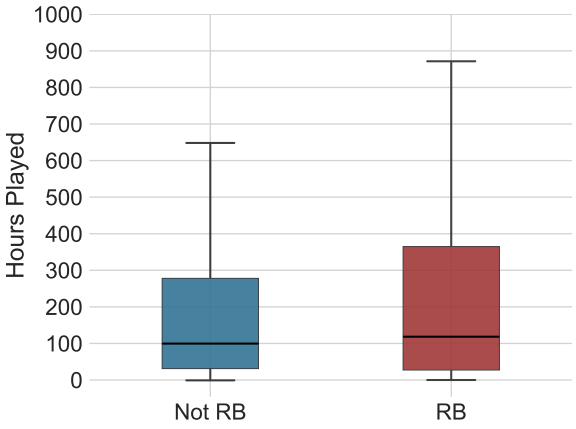
(c) String length

Figure 12 – CDFs of RB intensity, games recommendation rates and text length.

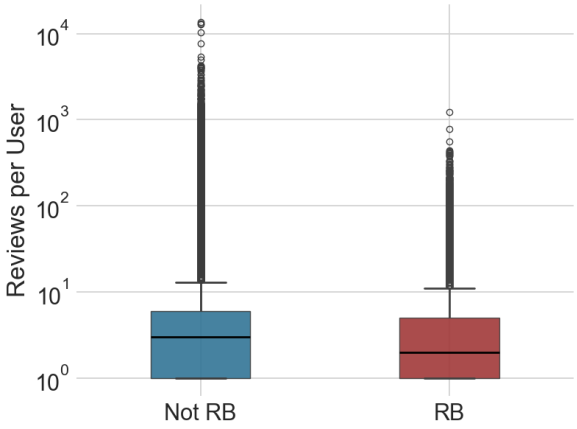
reviews. At the other extreme, some cases show highly concentrated bursts. In *Feed the Cups*, the RB episode lasted about one month and generated 7,752 reviews, nearly 60% of the game's all reviews, making it the largest in relative terms.

To further illustrate the severity of these coordinated events, Figure 12b presents the cumulative distribution function (CDF) of the proportion of negative reviews ("Not Recommended") per game. To contextualize the impact of RB, we evaluate three distinct scenarios: games never affected by RB (*Not RB games*), affected games during their organic periods (*RB games Without RB reviews*), and affected games strictly during the attack window (*RB games With RB reviews*). The baseline distributions reveal that Steam titles are very frequently well-received, with more than 60% of unflagged games having a negative review rate below 20%. Similarly, this positive consensus is largely shared by the targeted games during their normal (organic) periods, with roughly 80% of them also maintaining a negativity rate below the 20% threshold. This establishes that games targeted by RB are not inherently poorly rated or low-quality products. In contrast, the distribution of reviews posted strictly during RB episodes shifts drastically to the right. During these coordinated attacks, only about 8% of the episodes maintain a negative review rate below 20%, while the median proportion of negative reviews surges to over 50%. This noticeable divergence clearly isolates RB episodes as severe, coordinated anomalies of negativity rather than standard fluctuations.

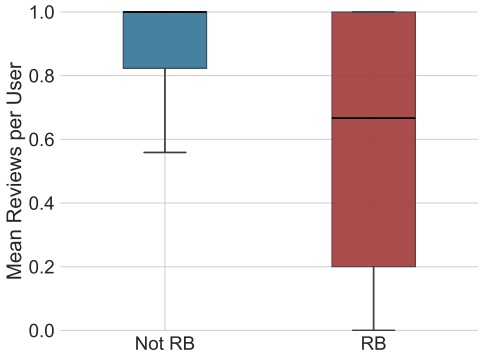
Beyond behavioral metrics, we also examined the textual characteristics of the reviews. Figure 12c presents the cumulative distribution of review lengths for games affected



(a) Hours played per reviewer user



(b) Number of reviewed games per user



(c) Average user recommendation rate score.

Figure 13 – Boxplot comparison of RB and non-RB users based on playtime, number of reviews, and recommendation patterns.

by RB, contrasting RB episodes with organic, non-RB periods. The two distributions are highly similar; however, the Kolmogorov test indicates a statistically significant difference between them. In both scenarios, approximately 70% of reviews contain fewer than 100 characters. This demonstrates that participation in a coordinated RB campaign does not alter user verbosity or the length of reviews.

Next, we analyzed overall gaming engagement of users posting RB reviews. Figure 13a compares hours played for RB and non-RB user reviewers. Reviews during RB episodes are similar to non-RB reviews but show slightly higher playtime. Non-RB reviews have a median of about 100 hours, whereas RB reviews show a median near 125 hours. Thus, RB participants tend to be relatively more engaged players than casual or new ones. We then examine the number of games reviewed per each user. Users who reviewed at least one RB-affected game are divided into two groups: those who posted at least one RB-period review and those who never did. This yields approximately 260,439 RB users and 7,142,825 non-RB users. Figure 13b shows the distribution of total games reviewed per user. Non-RB users are slightly more active overall, with a median of 7 games reviewed, compared to 5 for RB users.

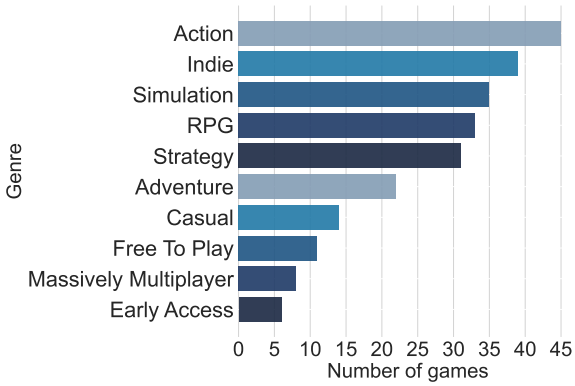
Further analysis of participation across multiple RB episodes reveals that RB is not driven by a small core of repeated attackers. Among all the 260,439 users involved in RB posts, only about 10% participated in more than one RB episode, and this number drops to just 1.38% for users involved in more than two events. This suggests that RB episodes are typically carried out by very distinct groups of users reacting to specific situations, rather than by a persistent set of actors repeatedly engaging in coordinated attacks. The clearest distinction between the two groups emerges when considering overall review recommendation. Figure 13c presents the average recommendation rate for each user across all of their Steam reviews. Non–RB users tend to be consistently positive, recommending the majority of the games they review. RB users, by contrast, display a more polarized pattern. Although they still recommend more games positively than they do negatively, their overall recommendation rate is lower, indicating a greater willingness to express dissatisfaction within the platform.

Overall, these findings provide a perspective on RB users that differs from much of the existing literature. Prior studies, such as (MELO; PIMENTEL, 2022; CANTONE; TOMASELLI; MAZZEO, 2025), characterize RB participants as newly created or low-engagement accounts primarily used to spread hostility in their case study. On Steam, however, users must own a game in order to post a review, which leads to a distinct dynamic. Although RB users post a higher proportion of negative reviews compared to non–RB users, they are otherwise remarkably similar in terms of total number of reviews, accumulated playtime, and review length. In some cases, RB users even exhibit slightly higher engagement, as reflected by their playtime prior to posting reviews. These results evidence that RB is a multilayered and distinctive phenomenon and often requires more complex strategies to be identified.

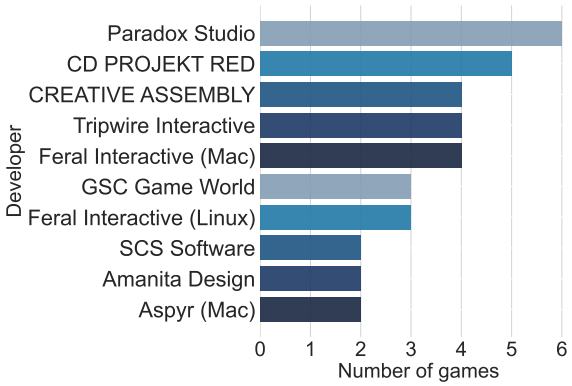
6.3 Targeted Games of Review Bombing

Complementing the patterns discussed above, we examine which games are most frequently affected by RB by analyzing genre, developer, and main games.

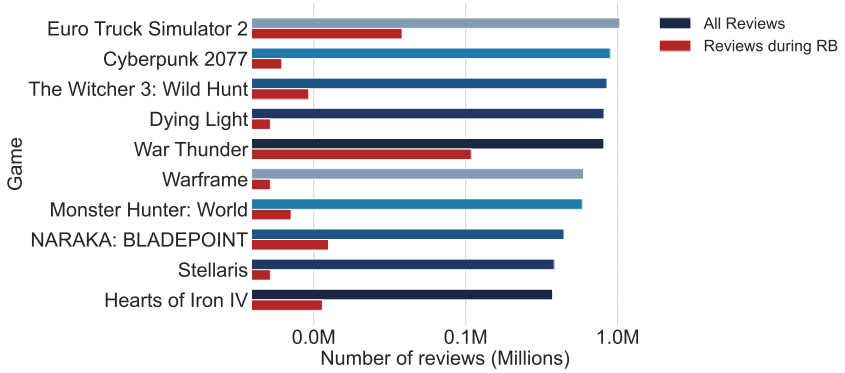
First, we analysed which genres and types of game are more frequently targeted by RB. Because Steam allows games to have multiple genre tags, each tag assigned to an RB game was considered independently. Figure 14a shows that “Action” and “Indie” games are the most frequently targeted. Among the 89 RB episodes, 45 ($\approx 50\%$) involve “Action” and 38 involve “Indie” titles. An interesting result is the prominence of Indie titles among affected games. Indie projects often maintain closer and more direct relationships between developers and their communities when compared to large studios. While this proximity can foster strong engagement and communication, it may also increase vulnerability to backlash. Decisions, public statements, or perceived missteps by developers can quickly circulate within tightly connected communities, drawing attention that ex-



(a) Top 10 RB episodes by genre



(b) Top 10 RB episodes by developer



(c) Top 10 most reviewed games that experienced RB

Figure 14 – Histograms describing the top 10 most affected games by RB, including genre, developer, and review volume distributions.

tends beyond the game itself and sometimes resulting in coordinated negative reviews. A well-known example is *Depression Quest* (MORTENSEN, 2018), an indie game associated with the *Gamergate* controversy whose developer became the target of widespread online harassment following personal accusations circulated in gaming communities. The case illustrates how the high visibility and personal exposure of indie developers can intensify community backlash and make both developers and their games particularly vulnerable to RB campaigns.

We next analyze developers associated with multiple RB episodes. Figure 14b shows that developers with more than one affected game often experience RB across all their titles for the same underlying reason. For instance, *Paradox Interactive*, the developer with the highest number of RB games with a total of six games, produces strategy and simulation games such as *Hearts of Iron IV* and *Crusader Kings III*. In both cases, RB episodes were triggered by the same grievance: Chinese players argued that Chinese territories were inaccurately represented across these titles (RACINOWSKA, 2025). Another example is *CD Projekt Red*, where backlash against the company's public stance during the Russia–Ukraine conflict led to RB episodes across 5 of its 41 titles (ADAMS, 2022b). This pattern reinforces that RB is not about the games themselves, but about the developers and their actions, decisions, and public statements. Combined

with earlier findings that most users participate in only one RB episode, this also indicates that Steam's requirement that users own a game before posting a review may partially limit the broader spread of RB across the platform. While coordinated backlash can still occur, this constraint appears to reduce its scale compared to environments where participation does not require prior ownership.

Finally, we examine the most reviewed RB games. Looking at absolute review volume highlights the presence of titles whose communities generate huge amounts of activity on Steam. For instance, *Euro Truck Simulator 2*, the biggest game in terms of reviews that got RB, stands out with more than one million reviews alone, illustrating the scale that long-running and widely played games can reach. Among these titles, one episode stands out in particular. When considering the absolute number of RB, *War Thunder* presents the biggest RB episode of steam in absolute review number. Despite being a massive title with more than 800k total reviews, it also exhibits an unusually large number of RB reviews, reaching close to 100k during the episode and persisting for 44 days. The scale and duration of this episode point to a highly engaged but also highly dissatisfied community. The backlash was largely directed at changes to the in-game economy, particularly after updates significantly increased the difficulty of acquiring certain items without monetary purchases (PLUNKETT, 2023).

7 Unveiling the Sociopolitical Motivations behind Review Bombing

Previous chapters examined who participates in RB and how these events manifest through behavioral and temporal patterns. However, understanding RB also requires analyzing what users are actually saying during these episodes and which motivations drive their participation. This chapter addresses RQ3 by analyzing the textual content of reviews posted during validated RB events using Bertopic modeling. Our goal is to identify recurring themes across different games and controversies, revealing the political, cultural, and economic grievances that motivate coordinated reviewing behavior. We begin by presenting the topic modeling procedure and then discuss the main topics identified within RB reviews. Finally, we examine how these findings contribute to understanding RB as a form of digitally mediated collective action.

7.1 Topic Modeling of Review Bombing Reviews

While temporal and behavioral patterns provide structural insights into RB, understanding it as a sociotechnical act requires decoding player discourse. We therefore move from metadata to semantic analysis, investigating the narratives, grievances, and collective positioning mobilizing users during RB episodes. Our analysis focuses on topic modeling of approximately 292,916 reviews posted within validated RB windows, comprising 168,213 negative and 124,703 positive entries. For this analysis, we split our RB dataset into two parts: recommend (positive) reviews and not-recommend (negative) reviews.

The resulting topics are presented in Table 3. All stages of topic extraction, clustering, and topic assignment were performed by our BERTopic pipeline, while ChatGPT (GPT-5.5, OpenAI) was used exclusively for auxiliary interpretation tasks: translating topic keywords and representative quotes, and generating human-readable topic labels. During translation, we explicitly instructed the model to preserve the original meaning as literally as possible in order to minimize interpretative bias. For topic naming, we provided the model with the topic keywords and also some representative review examples to improve contextual accuracy. Additionally, all generated topic labels were manually reviewed by the authors to ensure that they appropriately represented the underlying discussion themes. In some topics (e.g., *Game economy*), repeated words are expected because they correspond to the same term across different languages.

We first focus on the negative topics. By analyzing topic titles, their reviews ex-

amples, and their associated games, we contextualize both the semantic meaning of each topic and the events that triggered the corresponding RB episodes. The two most dominant topics, *Commercial Fraud* and *Refund Requests*, together account for approximately 75% of all negative reviews mapped. These topics appear in nearly all RB games, being present in 87 out of the 89 affected games. Importantly, reviews within these topics are not tied to any specific RB episodes; instead, they reflect a recurring behavioral pattern in which players, after expressing their criticisms or opinions, label the game as fraudulent and explicitly request refunds; for example, this quote from *Euro Truck Simulator 2*, where the user complains about the developer political position and asks for a refund.

> ***Euro Truck Simulator 2 (Russian):****"Don't get involved in politics in the gaming industry! Especially to write lies. If there was an opportunity, I would make a refund"*

More specific agendas emerge in topics such as *Game Economy*, *Store Exclusivity*, and *Developer Greed*, which together represent about 9.3% of mapped reviews. These correspond to documented RB events, including backlash against *Borderlands 2* and *Metro* (CHALK, 2019) following Epic Games Store exclusivity announcements, and protests against *War Thunder* (PLUNKETT, 2023) (Chapter 6).

> ***Borderlands 2 (English):****"Epic Store is trash. Exclusivity deals on PC gaming is trash. Don't support Epic or the companies that sign exclusivity deals. Steam ain't perfect, but exclusivity isn't competition. If Epic wanted to compete, they'd build a better platform. They're not trying to win people over, they're trying to strongarm people into using their platform. They don't care about the consumer. Epic is not good for gaming, it's divisive and anti-consumer."*

> ***War Thunder (English):*** *" "also, review bombing will not cause modiying or nullifying in-game prices - if the game is shut down, no one wins" well gaijin, i don't need to win. i just need YOU to lose."*

In the *War Thunder* example, RB is framed as deliberate protest rather than spontaneous reaction. The reviewer acknowledges RB and its potential influence on in-game decisions but pursues it to exert pressure through reputational harm. This suggests RB often functions as organized collective action aimed at damaging developers rather than improving the game.

Turning to geopolitical and cultural topics, we see another side of RB. *Chinese Discrimination*, *Censorship*, and *Chinese Identity* topics together account for roughly 10% of the mapped reviews and are largely associated with coordinated actions by Chinese player communities. Games such as *Tekken 8* (DAMMANN, 2024), and *Hearts of Iron IV* (RACINOWSKA, 2025) were RB following episodes including alleged unjustified bans of Chinese players (including professional competitors), perceived misrepresentation of Chinese territories, or the absence of Chinese localization. Reviews frequently frame

these issues as systemic disrespect toward Chinese players, as illustrated by the following example:

> ***Hearts of Iron IV (Chinese):*** *" Chinese players have been demanding Tibet cores for how many years now? Now even India has Tibet cores, yet China's two parties have none. You have never respected Chinese customers or taken our demands seriously, and this time you even locked Chinese IPs. Is this what you call freedom of speech? From now on, you can release whatever DLC you want—if I buy even one, I lose! Nobody likes spending money to buy shit and eat it. The South America DLC last time and this India DLC this time are both piles of shit. The Old Buddha already paid the money."*

Additional geopolitical topics, such as *Racism and War Propaganda* (approximately 1%), are tied to the Russia–Ukraine conflict. Although this proportion appears small, it is important to note that most of these reviews are in the first two topics, as illustrated by the *Euro Truck Simulator 2* quote. In these cases, developers' public stances prompted RB by players who perceived such statements as inappropriate political interference. For instance, *The Witcher 3* (ADAMS, 2022b) received hostile reviews following public positions taken by its developers:

> ***The Witcher 3 (Russian):*** *"Games, like sports, should be outside of politics, especially when you don't understand how our people live and how it is simply impossible to change anything without destroying EVERYTHING that we have. I hope that someday it will get through to you, Polish hypocrites, which you have been all along—though in principle that's unlikely."*

We also observe explicit hate-speech-related topics, such as *LGBTQIA+ Discourse and Political Speech*, together accounting for about 1.8% of negative reviews. These capture backlash against developers' public statements or perceived ideological positions. One example involves *Feed the Cups* (U/RANDOMDRAGONSLAYER, 2024), whose developer made sexist remarks in an interview, prompting backlash:

> ***Feed the Cups (Chinese):*** *"Next time remember to say in advance that the game has built-in "dick detection": if you don't have a dick, on opening day it automatically deducts 1,000 yuan in startup funds. Female players have to be stepped on; female players' money can't be left unearned. You got popular by relying on female players, then once you put the bowl down you turn around and curse female players. I don't even have two ounces of meat between my legs, I can't press the controller anymore. Refund! Refund! Refund! Refund! Refund!"*

Titles published by *Tripwire* were targeted following the CEO's public support for Texas' anti-abortion legislation (CHALK, 2022). In *Killing Floor*, political disagreement is framed through overt insults and ideological labeling:

> ***Killing Floor (Russian):*** *"The game itself is ok, but Tripwire interactive is a cuck company that complies to cancel culture. The CEO had a perfectly normal Christian stand and they sacked him. Expect liberal gaffotry in the future. TWI, may you go broke, and with all due respect (aka none), go fk yourselves"*

Importantly, these cases illustrate that RB operates as a community-driven response mechanism, rather than one inherently tied to a fixed ideological position. In the case of *Feed the Cups*, users mobilize collective outrage to sanction the developer for sexist remarks made in a public interview, framing their negative reviews as a form of moral accountability. By contrast, in *Killing Floor*, the community reaction unfolds in the opposite direction: players deploy RB to defend the CEO following political backlash, portraying his criticism and removal as an instance of so-called "cancel culture."

Turning to positive reviews, most topics reflect positive engagement. *Fun and General Gameplay* and *Truck Simulator* dominate, accounting for about 91% of mapped reviews. However, we also observe RB-related topics mirroring those in the negative model, including *Chinese Protest*, *Commercial Fraud*, *Ukraine Conflict*, and *Updated Economy*. These topics suggest a strategic use of positive recommendations to avoid triggering moderation or automatic RB detection, while still expressing dissatisfaction.

Notably, we also identify the *Military Misogyny* topic within the positive reviews, primarily associated with the RB of *Total War: ROME II* (CHALK, 2018; ARIF, 2018). In this case, players used positive reviews to express hostility toward an update that increased the frequency of female generals, employing positive ratings as a vehicle to disseminate misogynistic discourse:

> ***Total War: ROME II (Chinese):*** *"CA idiots, forcing political correctness. Why don't you just go all the way with being "correct" then—make all generals female, Black, and gay, plus HIV-positive, animal-rights activists, and disabled, preferably also Muslim and refugees."*

Finally, hate-speech-related topics further reinforce this pattern: RB can be mobilized in opposing ideological directions within the same platform. While in *Feed the Cups* negative reviews function as moral sanctions against sexist behavior by the developer, in *Total War: ROME II* positive reviews are strategically used to propagate misogynistic and anti–political correctness rhetoric in response to increased female representation. These cases demonstrate that RB constitutes a flexible protest mechanism, capable of being deployed both to contest hate speech and to actively propagate it. The same dynamic extends to political controversies, with the direction and framing of the protest shaped by the ideological orientation and motivations of the mobilized community. Another important aspect is the global nature of RB. The identified topics span multiple languages, regions, and geopolitical contexts, including disputes involving China, Russia, the United States, and broader debates. This suggests that RB should not be understood as a localized behavior, but rather as a transnational form of digitally mediated collective action shaped by global political, cultural, and ideological tensions.

This section also highlights an important distinction between RB and other forms of hateful or toxic expression commonly observed in online communities. Prior work iden-

tifies behaviors such as attacks motivated by camaraderie, in which harassment is performed for the approval of other aggressors, and actions driven by entertainment value, where users generate chaos "*for the lulz*" (STRINGHINI; BLACKBURN, 2024). While such dynamics may also appear within RB episodes, our findings suggest that RB is not primarily a form of trolling or performative toxicity. Instead, it more often reflects coordinated expressions of dissatisfaction directed at developers and their decisions, with reviews used as a mechanism to pressure or sanction them. In this sense, RB functions less as random hostile expression and more as a collective protest boosting tool through which communities attempt to influence/stress developers and the trajectory of their games.

7.2 Agency, Governance, and the Weaponization of Reviews

Our results indicate that RB is not merely a technical anomaly or isolated form of toxic behavior, but rather a large-scale form of coordinated collective action embedded within digital social networks. The identified topics reveal how users appropriate platform review systems to mobilize around shared political, ideological, and economic grievances, transforming review infrastructures into spaces of collective visibility and protest. At the same time, the diversity of topics and controversies demonstrates that RB emerges from decentralized and heterogeneous communities rather than from a single organized movement.

As referenced in Section 6.2 approximately 10% participated in more than one event, while just 1.38% engaged in more than two episodes. This suggests that RB is not primarily sustained by a small core of persistent attackers, but instead by recurring waves of temporary collective mobilization driven by specific controversies. In this sense, RB resembles episodic networked collective action, where different communities converge around shared grievances and use platform mechanisms to amplify their positions. The predominance of topics such as *Commercial Fraud* and *Refund Requests* further suggests the existence of a standardized "lexicon of protest." By framing dissatisfaction through refund demands and accusations of fraud, users strategically employ the commercial language of the platform to maximize visibility and reputational pressure. As argued by (CANTONE; TOMASELLI; MAZZEO, 2025), RB operates as a semiotic act in which review scores communicate broader messages beyond product evaluation. Rather than simply refusing consumption, users repurpose platform reputation systems as mechanisms of collective pressure directed toward developers and publishers.

Our findings also reveal an important tension between collective user agency and platform governance. Steam's moderation model treats RB primarily as "off-topic" activity to be filtered or hidden from visibility metrics. However, the presence of coordinated positive-review campaigns and politically coded discourse demonstrates that users adapt strategically to these moderation practices. Instead of abandoning protest, communities

modify the form and framing of reviews to bypass visibility reduction and automated detection. This aligns with (WIRZBURGER, 2025), who argues that Steam manages politically motivated protest through depoliticization strategies that reduce the public visibility of dissent while preserving the appearance of platform neutrality.

From a social network perspective, these dynamics illustrate how platform infrastructures shape the visibility, diffusion, and persistence of collective behavior. Steam's review system simultaneously functions as a reputation mechanism, a communication channel, and a site of political expression. By algorithmically suppressing certain forms of coordinated activity while leaving others visible, the platform influences which collective actions become publicly amplified and which are rendered socially invisible. Consequently, RB should not be understood solely as abusive reviewing behavior, but also as a form of digitally mediated collective participation emerging from the interaction between users, platform governance, and transnational online communities.

Finally, our results reinforce that RB possesses strong ideological flexibility. The same review infrastructure can be mobilized both to contest discriminatory behavior and to propagate hate speech, depending on the motivations of the participating communities. Cases such as *Feed the Cups* and *Total War: ROME II* demonstrate that RB is not inherently progressive or reactionary; rather, it constitutes a flexible mechanism of collective pressure shaped by the political and cultural orientation of the mobilized network. This highlights the limitations of moderation systems based exclusively on review volume or polarity, since the social meaning and impact of RB depend fundamentally on the underlying discourse, motivations, and patterns of collective mobilization.

Table 3 – BERTopic results: negative Reviews (left) and positive Reviews topics (right), including associated keywords, total reviews, and number of affected games.

Negative Reviews				**Positive Reviews**			
Topic	**Keywords**	**Reviews**	**Games**	**Topic**	**Keywords**	**Reviews**	**Games**
Commercial Fraud	fraud, fraud, game, game, game, game, fake, game, sales, premium	38356	87	Fun	fun, pretty fun, very fun, extremely fun, fun love to play, so fun, fun, funny, like, cool	32840	80
Refund Requests	rnm refund, refund, recharge to refund, refund template included, rnm, nato, bund, bvvd, trash profit, devs	26909	87	General Gameplay	gameplay, games, playing, game, played, good, review, better, fun, well	22054	82
Chinese Discrimination	Chinese Tekken player, discrimination, unfair treatment, Namco abandoning, unfairness, harm to players, tournament, competitive, forced forfeiture, retaliation	6512	39	Truck Simulation	simulator, simulator, simulation, simulation, truckersmp, truck, truck, truck, trucker, truck	1939	8
Game Economy	economy, economy, economy, economy, economy, economy, economy, economy, economy, economy	3648	6	Ukraine Conflict	in Ukraine, of Ukraine, in Ukraine, ukraini, to Ukraine, Ukraine, in Ukraine, Ukrainian, soviet, putin	1913	53
Dev Greed	snail, snails, snail, salt, saltthesnail, salted, salting, greedy, greed, dog never stops eating	2684	41	Updated Economy	they fixed, economy, economy, economy, economy, economy, economy, economy, economy	1411	38
Store Exclusivity	store, store16, customer, epic exclusive, consumer, epic, sells, purchasing, owned, google	1765	6	Military Misogyny	generals, generals, misogynist, women, armies, females, women, faction, reviews, factions	598	31
Racism	discrimination, discrimination, discrimination, racial, racial, racism, race, racial, racist, ethnic	1399	41	Snail Greed	I sold, soul, souls, sold, snail, snail, snails, snail, snailhail, sold	546	19
Political Speech	twitter, opinion, john, abortion, abortions, banning, reviews, political, ban, speech	1399	56	Map Design	map design, mapleri, maps, maps, map, map, big map part 3, merge 2, world map, map	257	26
Censorship	censorship, censorship, censorship, censorship, censoring, censored, censorship, censored, duckrights	1324	22	Culinary Content	recipe, flour, ingredients, cooked, cook, recipe, salad, flour, fried, dough	176	26
Chinese Identity	Chinese players identity, visibility of mainland China, self-identification, China reference, abandoned promise, refund request, money focus	952	15	Naval War	navy, fleet, marine, naval, marina, navy, fleet, navy, submarine, ships	169	6
Repair Costs	repair costs, repair, repair, repairs, repairing, repair, repair, repair, repair, repairing	661	10	Commercial Fraud	sales fraud, sales fraud, do not buy, do not buy, false advertising, scam, fraud, altered content	135	1
Malicious Review	game, fake, fraud, fraud, game, game, leave a bad review, maliciously, company, buy	539	6	Chinese Protest	torrent China, Beijing Spring, protests, demonstrations, counterrevolutionary, Republic of China, Taiwan	83	11
Pay-to-Win	p2w, p2w user, pw2, p2fun, p2f, p2work, p2s, p2p, p2e, p2win	436	4	Bizarre Speech	snow what, snow, snow, says, sky, he, his, bizarre, sky, believed	64	4
War Propaganda	Russia loses war, Russia bombed Ukraine, West stayed silent, US bombed Syria, Iraq, Afghanistan, Yugoslavia, Somalia	179	4	Verbal Attack	knees, knee, knees, broken glass, breathe, chyyyy, breath, muh, kinda, insulting China	64	2
LGBT Discourse	gays, gaym, gays, gays, gay, gay, gays, gays, lgbt, gayhttps	163	18	Audio Listening	listened, listened, listened, listening, listened, listen, heard, hear, heard, hearing	58	1

8 Machine Learning Results for Review Bomb Detection

This chapter addresses RQ4 and RQ5 through a supervised machine learning framework that models RB as a temporal phenomenon. RQ4 investigates whether RB episodes can be automatically distinguished from organic waves of negative feedback, while RQ5 identifies the semantic, temporal, and behavioral signals that are most informative for this distinction. Rather than proposing a definitive RB detection model, our objective is to evaluate the feasibility of automated RB detection and analyze which features best characterize the phenomenon.

The experimental framework comprises the temporal window formulation, strategies for handling class imbalance, the engineered feature space, the evaluated model configurations, and the classification protocol. Building upon this framework, this chapter presents the classification performance obtained by the evaluated models (Section 8.2) and the feature importance and interpretability analyses that address RQ5 (Section 8.1).

8.1 Feature Importance and Analysis

To provide an initial understanding of how the engineered features differ between RB and non-RB periods, Figure 15 presents the kernel density estimation (KDE) distributions for all ten features. In the figure, blue curves represent non-RB windows, while red curves correspond to RB windows. The 384 embedding dimensions were omitted from this visualization due to space constraints and limited interpretability.

A few features display different distributions across the two classes, suggesting that they may contribute useful discriminative information for classification. An example is `not_recommended_ratio`, non-RB windows exhibit a broader and more dispersed distribution, with density values generally remaining below 2. In contrast, RB windows present a pronounced concentration at lower values, with a density peak approaching 5 between approximately 0 and 0.2. Indicating that non-RB windows tend to have way more positive reviews than the RB windows.

Another distinction can be observed for `period_entries_percentage`. The distribution for non-RB windows is heavily concentrated near zero and reaches density values close to 40, whereas the RB distribution is substantially more dispersed, with its highest density occurring around 5. This behavior indicates that most non-RB windows correspond to periods containing only a very small fraction of a game's total reviews. RB windows, on the other hand, are more frequently associated with larger proportions of the

total review volume, reflecting the sudden bursts of activity commonly observed during bombing episodes.

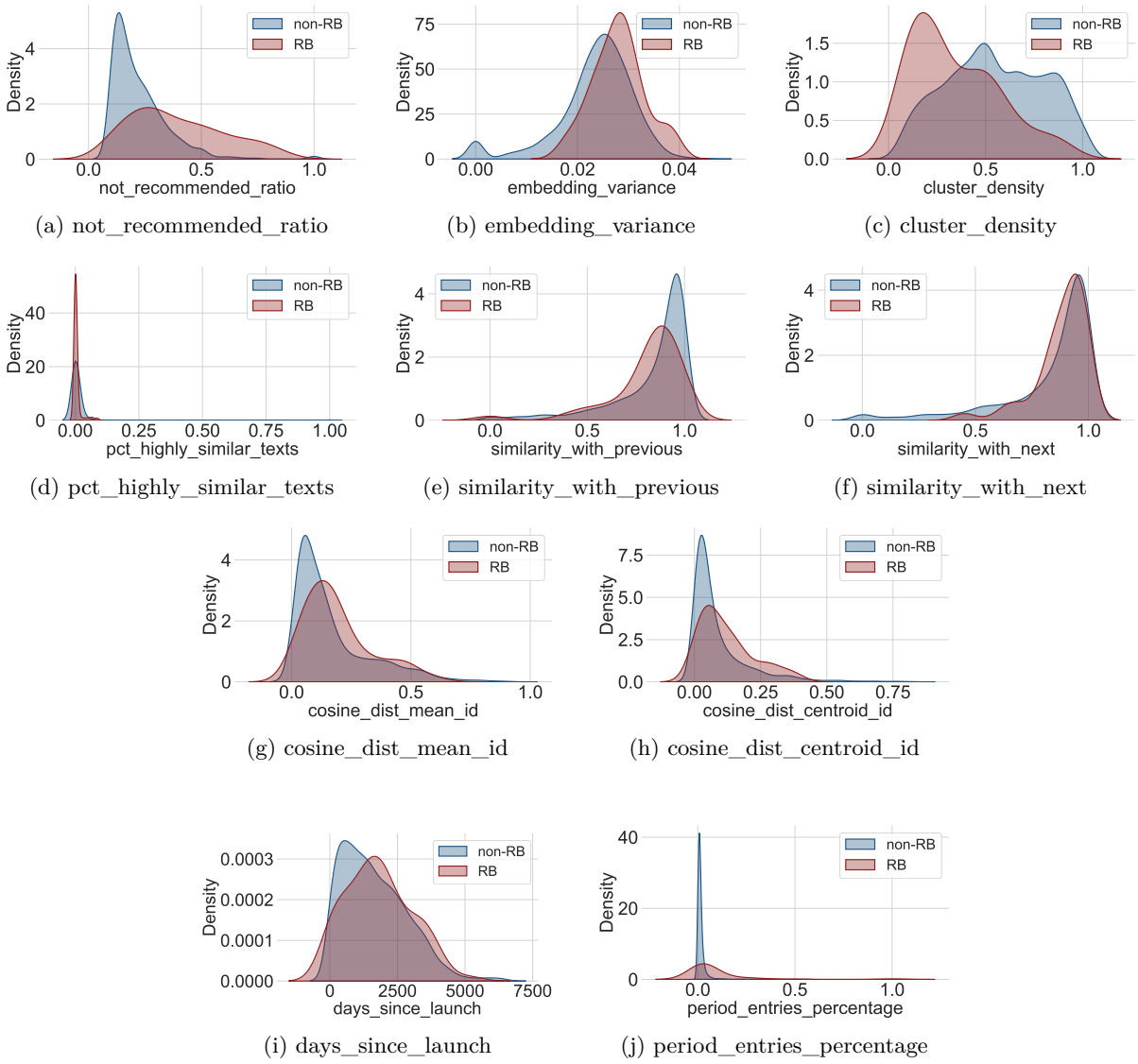
Another feature exhibiting noticeable separation is `cluster_density`. Non-RB windows are more concentrated at values below 0.5, while RB windows present greater density at higher values. This suggests that reviews posted during RB periods are more likely to concentrate around a dominant semantic cluster, whereas regular review activity tends to be distributed across a wider variety of discussion topics.

In contrast, features such as `embedding_variance`, `similarity_with_previous`, and `similarity_with_next` exhibit largely overlapping distributions for RB and non-RB windows. This overlap indicates that these features individually provide limited class separation. Nevertheless, even features with similar marginal distributions may remain valuable when combined with other variables, since machine learning models can exploit nonlinear interactions that are not directly observable through univariate distribution analysis alone. Moreover, since these features are directly derived from the chosen embedding model, exploring larger and more complex architectures (e.g., billion-parameter LLMs) may generate representations that naturally exhibit more distinct distributions across classes.

To better understand the separability of RB and non-RB periods in the proposed feature space, Figure 16 presents a two-dimensional t-SNE projection of all temporal windows. RB windows are represented in red, while non-RB windows are shown in blue. Although t-SNE is primarily a visualization technique and does not preserve global distances, it provides useful insights into the structure of the learned representation.

Several dense clusters can be observed in the projection, particularly around tSNE_1 values close to 60, around tSNE_1 near -60, and in the lower region of the projection where tSNE_2 approaches -60. Each cluster likely corresponds to groups of windows associated with a specific game whose review vocabulary, writing style, and discussion topics differ substantially from those of other games. Interestingly, most of these clusters remain spatially separated from RB windows, suggesting that RB periods often exhibit semantic characteristics distinct from the regular discussion patterns of their respective games. Only the cluster around tSNE_1 = -60 contains RB windows in proximity, indicating that in this game the semantic transition associated with RB may be less pronounced.

A few localized concentrations of RB windows can also be identified, such as those near (tSNE_1 = -5, tSNE_2 = 0) and (tSNE_1 = 10, tSNE_2 = 40). One possible explanation is that these windows correspond to RB episodes discussing similar controversies or motivations, which is consistent with the recurring thematic patterns identified in Chapter 7. Nevertheless, most RB windows remain dispersed throughout the projection rather than forming a single compact cluster. Consequently, RB periods do not appear as a geometrically isolated region in the low-dimensional representation. Although some



(a) not_recommended_ratio (b) embedding_variance (c) cluster_density

(d) pct_highly_similar_texts (e) similarity_with_previous (f) similarity_with_next

(g) cosine_dist_mean_id (h) cosine_dist_centroid_id

(i) days_since_launch (j) period_entries_percentage

Figure 15 – Kernel density estimation (KDE) distributions of the ten analyzed features. Red curves represent review bombing (RB) periods, while blue curves represent non review bombing (non RB) periods.

local structure can be observed, substantial overlap remains between RB and non-RB windows, highlighting the challenging nature of the classification task and indicating that successful detection requires a more complex set of features.

Figure 17 presents the Pearson correlation matrix computed for the ten engineered features and the target variable `review_bomb`. The 384 embedding dimensions were omitted from this visualization due to space constraints and limited interpretability.

The target variable exhibits relatively weak positive correlations with `period_entries_percentage` (0.26) and `not_recommended_ratio` (0.20), which are the highest among the engineered features. These relationships are consistent with previous findings discussed throughout Chapter 6. RB episodes are typically characterized by abrupt increases in review activity and negative reviews, making both features naturally associated with RB periods.

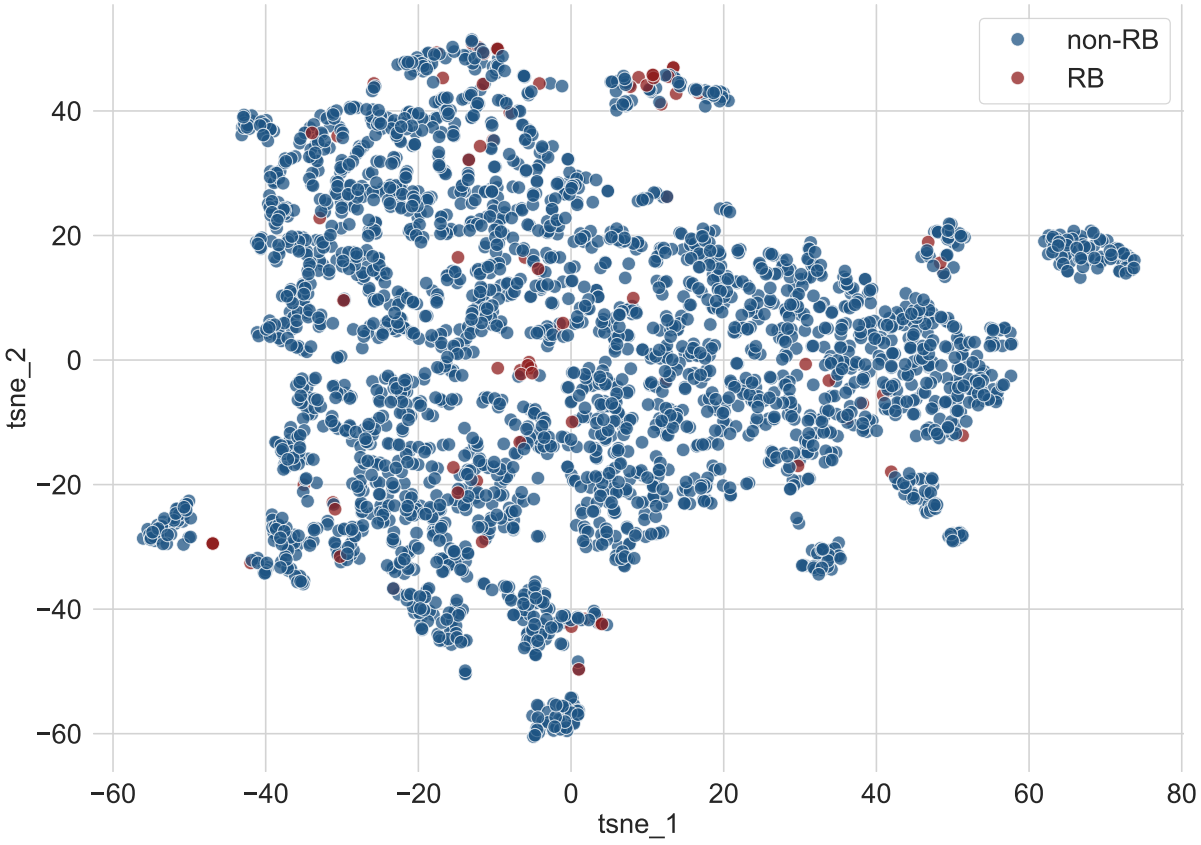


Figure 16 – t-SNE projection from RB and non-RB period features.

Several strong correlations can be observed among the engineered features themselves. For example, `cosine_dist_mean_id` and `similarity_with_previous` present a strong negative correlation of -0.82. This result is expected because windows that are highly similar to their preceding temporal context tend to be less semantically distant from the overall distribution of windows associated with the same game.

An even stronger relationship is observed between `cosine_dist_mean_id` and `cosine_dist_centroid_id`, which reach a correlation of 0.95. This near-perfect correlation indicates that both features capture closely related notions of semantic atypicality. While the first measures the average distance to all windows belonging to the same game, the second measures the distance to the global semantic centroid of the game. Consequently, both descriptors provide highly similar information regarding how unusual a temporal window is relative to the overall review history. Another interesting relationship is the positive correlation of 0.68 between `similarity_with_previous` and `similarity_with_next`. This suggests that temporal windows tend to exhibit local semantic continuity, meaning that periods similar to their preceding window are also likely to be similar to the following one.

To further investigate feature relevance, Figure 18 presents the average absolute SHAP values computed for the best-performing classifier, namely XGBoost. Figure 18a shows the ten most influential features considering the complete feature space, including

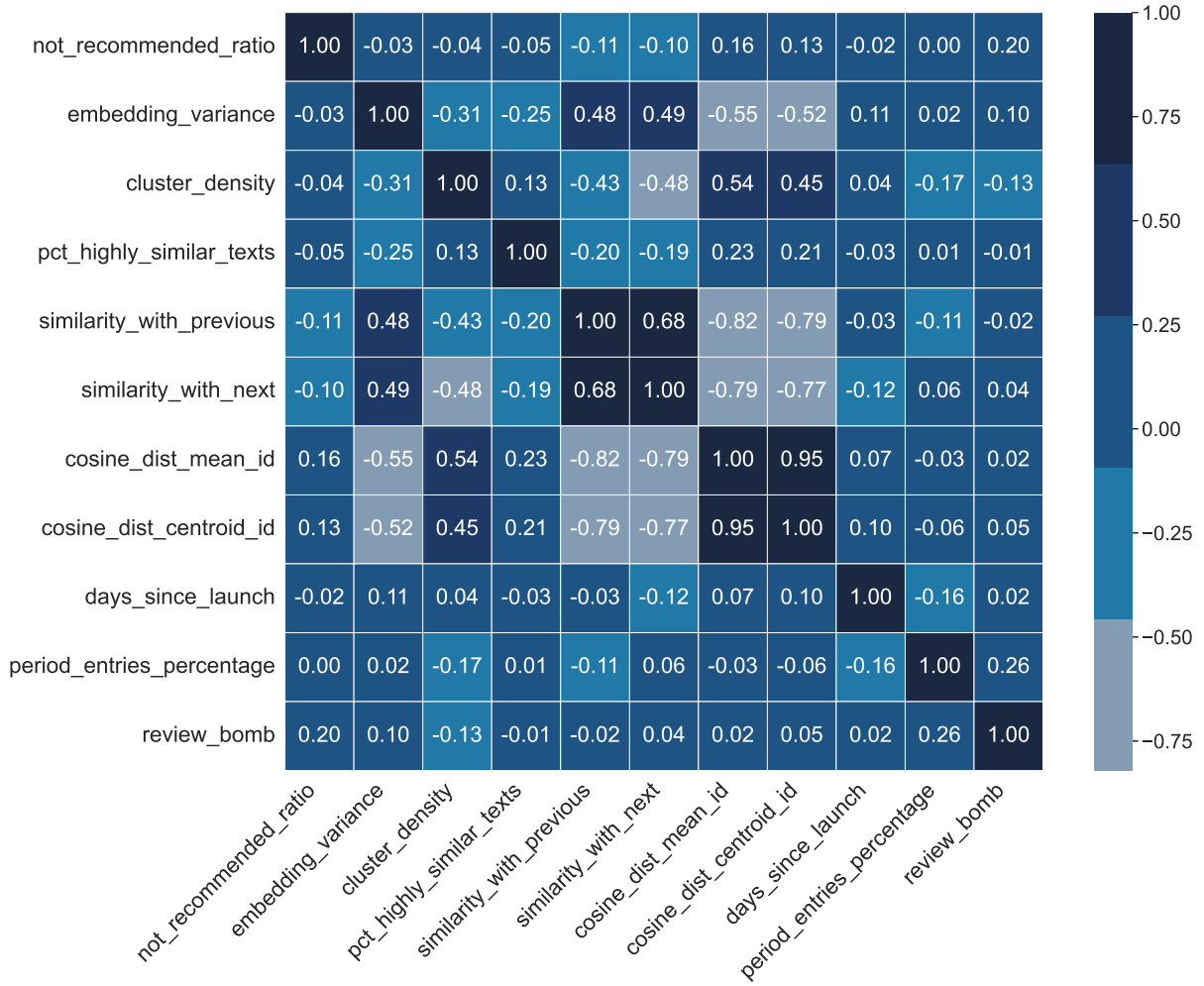


Figure 17 – Heatmap of the Pearson correlation between features.

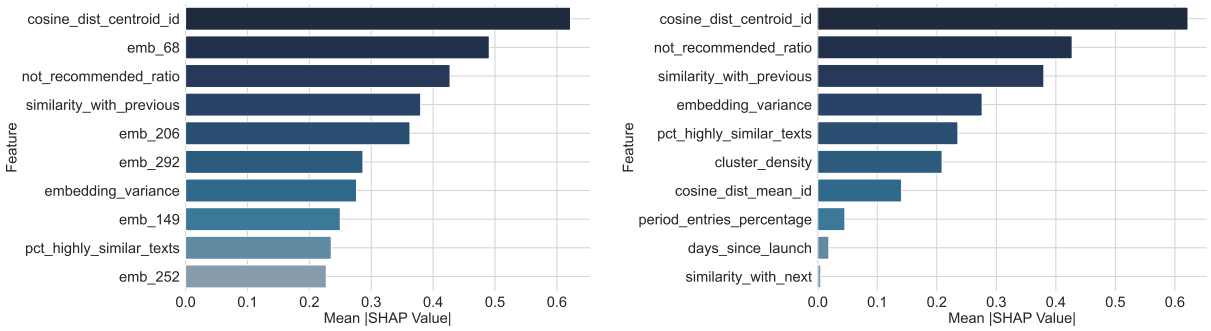
the embedding dimensions.

The most impactful feature is `cosine_dist_centroid_id`. This result aligns with the operational definition of RB adopted in this study, which characterizes the phenomenon as introducing a collective semantic shift in the review discourse of a game. During RB episodes, a substantial fraction of reviews becomes focused on a specific controversy or external event, causing the semantic representation of the corresponding temporal window to diverge from the overall review history of the game. As a result, distance-based measures of semantic atypicality become particularly informative for classification.

An additional observation is the significant contribution of embedding dimensions themselves. Among the ten most important features, five correspond to individual embedding dimensions. This indicates that the semantic representation generated by the language model contains relevant information that complements the engineered temporal and behavioral features.

To improve interpretability, embedding dimensions were removed from a second SHAP analysis, resulting in Figure 18b, which displays the relative importance of the ten

engineered features only. Several interesting patterns emerge from this analysis. Most notably, `similarity_with_previous` exhibits considerably greater importance than `similarity_with_next`. One possible explanation is that the onset of a RB event often produces an abrupt semantic disruption relative to the immediately preceding period. In contrast, once the bombing episode begins, subsequent windows may continue discussing the same controversy, reducing the discriminative value of similarity with future periods. As expected, not_recommended_ratio also appears among the most influential features, reinforcing the importance of negative review concentration as a signal of RB activity.



(a) All features. (b) Non-embedding features.

Figure 18 – Average SHAP importance. (a) Importance considering all features used by the classifier. (b) Importance considering only non-embedding features.

Finally, Figure 19 presents the SHAP beeswarm analysis for the engineered features. Since individual embedding dimensions are difficult to interpret, they were excluded from this analysis.

Examining the three most influential features reveals several informative trends. First, `cosine_dist_centroid_id` displays a clear monotonic behavior in which larger values contribute positively to RB predictions. This indicates that windows that are semantically distant from the overall review history of a game are more likely to be classified as RB periods. Such behavior is consistent with the expectation that RB episodes introduce discussions that differ substantially from the game's typical review topics.

Second, higher values of `not_recommended_ratio` are associated with positive SHAP values, increasing the probability of RB classification. This result directly reflects the operational definition of RB adopted in this study, which characterizes the phenomenon as a coordinated surge of negative reviews within a short period.

Finally, lower values of `similarity_with_previous` tend to increase the model's confidence toward the RB class. In other words, windows that differ substantially from the semantic content of the immediately preceding period are more likely to be identified as RB episodes. This behavior reinforces the hypothesis that RB often emerges as an abrupt disruption of the normal review discourse rather than as a gradual evolution of ongoing discussions.

Overall, the SHAP analyses indicate that the most relevant signals for RB detection, from the models perspective, are associated with semantic deviation, abrupt temporal changes, and concentrations of negative reviews. These findings are consistent with the behavioral patterns identified throughout the previous chapters and provide additional evidence that the proposed features capture meaningful characteristics of RB events.

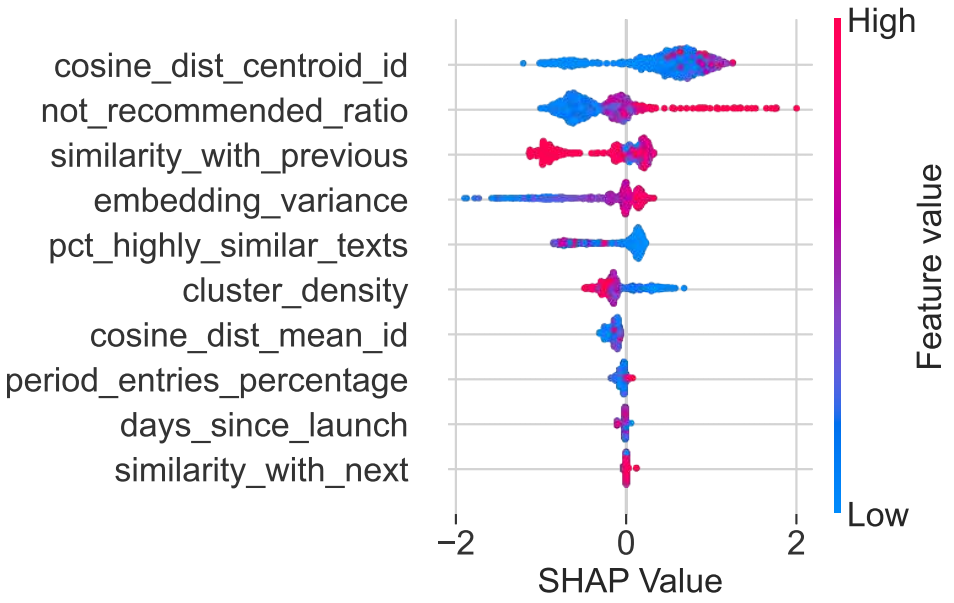


Figure 19 – SHAP beeswarm importance considering only non-embedding features.

Model	P_0	R_0	$F1_0$	P_1	R_1	$F1_1$	$\mathbf{F1}_{macro}$
XGBoost	0.974 ± 0.000	0.999 ± 0.000	0.986 ± 0.000	0.648 ± 0.006	0.062 ± 0.002	0.114 ± 0.003	0.550 ± 0.001
RF	0.975 ± 0.000	0.998 ± 0.000	0.986 ± 0.000	0.625 ± 0.006	0.096 ± 0.001	0.166 ± 0.002	0.576 ± 0.001
NN	0.976 ± 0.004	0.604 ± 0.424	0.639 ± 0.395	0.025 ± 0.005	0.403 ± 0.442	0.039 ± 0.017	0.339 ± 0.191

Table 4 – Baseline classification results using only aggregated semantic embeddings as input features. Results are reported as mean ± standard deviation across the 5-fold cross-validation procedure.

8.2 Classifier Models Results

Tables 4 and 5 summarize the performance obtained for the baseline representation and the proposed temporal window-based representation, respectively. In both tables, P_0, R_0, and $F1_0$ correspond to precision, recall, and F1-score for the non-RB class, while P_1, R_1, and F11 correspond to the same metrics for the RB class. F1*macro* denotes the macro-averaged F1-score across both classes. Since oversampling consistently produced the best results, only the oversampling configurations are reported. Throughout this analysis, $F1_1$ is considered the primary evaluation metric, as the main objective is to correctly identify RB periods.

Table 4 presents the results obtained using only semantic embeddings without temporal aggregation or engineered features. Overall, the baseline models achieved poor performance for RB detection. Although XGBoost and Random Forest reached moderate precision values for the RB class (0.6477 and 0.6252, respectively), both models exhibited extremely low recall, below 0.10. Consequently, their $F1_1$ scores remained low, reaching only 0.1139 and 0.1661. The neural network achieved substantially higher recall (0.4028), but at the cost of very low precision (0.0247), resulting in unstable performance across folds. These results indicate that semantic embeddings alone do not provide sufficient information to reliably distinguish RB periods from regular reviewing activity.

"Negative" Reviews Only								
Thr	**Model**	P_0	R_0	$F1_0$	P_1	R_1	$F1_1$	$\mathbf{F1}_{macro}$
0.10	XGBoost	0.991 ± 0.003	0.989 ± 0.006	0.990 ± 0.003	0.684 ± 0.110	0.694 ± 0.102	**0.678 ± 0.064**	0.834 ± 0.033
0.10	RF	0.985 ± 0.002	0.993 ± 0.004	0.989 ± 0.002	0.709 ± 0.126	0.501 ± 0.059	0.581 ± 0.058	0.785 ± 0.030
0.10	NN	0.989 ± 0.004	0.962 ± 0.015	0.975 ± 0.008	0.344 ± 0.117	0.634 ± 0.143	0.439 ± 0.117	0.707 ± 0.062
0.15	XGBoost	0.983 ± 0.004	0.980 ± 0.008	0.981 ± 0.004	0.570 ± 0.112	0.599 ± 0.086	0.579 ± 0.082	0.780 ± 0.043
0.15	RF	0.976 ± 0.006	0.994 ± 0.005	0.984 ± 0.002	0.789 ± 0.134	0.411 ± 0.137	0.511 ± 0.112	0.748 ± 0.057
0.15	NN	0.987 ± 0.007	0.953 ± 0.026	0.969 ± 0.012	0.437 ± 0.137	0.712 ± 0.158	0.511 ± 0.090	0.740 ± 0.050
0.20	XGBoost	0.975 ± 0.010	0.964 ± 0.023	0.969 ± 0.010	0.588 ± 0.228	0.615 ± 0.124	0.565 ± 0.081	0.767 ± 0.045
0.20	RF	0.982 ± 0.009	0.948 ± 0.035	0.964 ± 0.015	0.523 ± 0.183	0.725 ± 0.151	0.571 ± 0.094	0.768 ± 0.053
0.20	NN	0.982 ± 0.006	0.944 ± 0.036	0.962 ± 0.016	0.525 ± 0.226	0.714 ± 0.091	0.566 ± 0.112	0.764 ± 0.064
All Reviews								
Thr	**Model**	P_0	R_0	$F1_0$	P_1	R_1	$F1_1$	$\mathbf{F1}_{macro}$
0.10	XGBoost	0.984 ± 0.003	0.985 ± 0.012	0.985 ± 0.005	0.567 ± 0.178	0.470 ± 0.111	0.482 ± 0.044	0.733 ± 0.023
0.10	RF	0.978 ± 0.004	0.999 ± 0.002	0.988 ± 0.002	0.867 ± 0.125	0.255 ± 0.046	0.389 ± 0.052	0.689 ± 0.027
0.10	NN	0.989 ± 0.005	0.907 ± 0.084	0.944 ± 0.046	0.245 ± 0.104	0.648 ± 0.184	0.323 ± 0.089	0.634 ± 0.066
0.15	XGBoost	0.976 ± 0.004	0.989 ± 0.006	0.982 ± 0.003	0.626 ± 0.083	0.427 ± 0.120	0.497 ± 0.093	0.740 ± 0.047
0.15	RF	0.974 ± 0.004	0.994 ± 0.005	0.984 ± 0.002	0.788 ± 0.175	0.378 ± 0.114	0.489 ± 0.082	0.737 ± 0.041
0.15	NN	0.984 ± 0.006	0.956 ± 0.028	0.969 ± 0.014	0.424 ± 0.148	0.621 ± 0.146	0.480 ± 0.126	0.725 ± 0.069
0.20	XGBoost	0.974 ± 0.010	0.967 ± 0.027	0.970 ± 0.013	0.602 ± 0.193	0.590 ± 0.150	**0.558 ± 0.074**	0.764 ± 0.042
0.20	RF	0.969 ± 0.008	0.983 ± 0.012	0.976 ± 0.004	0.704 ± 0.182	0.493 ± 0.121	0.553 ± 0.060	0.765 ± 0.030
0.20	NN	0.984 ± 0.009	0.911 ± 0.057	0.945 ± 0.028	0.408 ± 0.162	0.759 ± 0.146	0.496 ± 0.111	0.720 ± 0.068

Table 5 – Mean ± standard deviation for the *OVER* strategy. Best $F1_1$ per threshold and review set is highlighted in bold.

The proposed temporal representation substantially improved performance across all evaluated classifiers. As shown in Table 5, every configuration achieved higher $F1_1$ values than the baseline, demonstrating the importance of incorporating temporal and structural information into the representation. The best overall result was obtained by XGBoost using only negative reviews and a negativity threshold of 0.10, achieving an $F1_1$ score of 0.6779.

Across all the classifiers, we can see that restricting feature extraction to negative reviews improved RB detection. Since RB episodes are primarily characterized by coordinated negative feedback, removing positive reviews appears to strengthen the signal

Model	Best Hyperparameters (Optuna)
XGBoost	n_estimators=255
	max_depth=5
	learning_rate=0.0796
	subsample=0.7824
	colsample_bytree=0.9141
	min_child_weight=2
	gamma=2.5712
RF	n_estimators=425
	max_depth=20
	min_samples_split=15
	min_samples_leaf=6
	max_features=sqrt
NN	hidden_layer_sizes=(64,32)
	dropout=0.3993
	learning_rate=2.94e-05
	alpha=4.21e-06
	final_epochs=10

Table 6 – Best hyperparameters identified by Optuna for the baseline representation using only aggregated semantic embeddings.

Model	Best Hyperparameters (Optuna)
XGBoost	n_estimators=477
	max_depth=8
	learning_rate=0.1976
	subsample=0.8000
	colsample_bytree=0.6025
	min_child_weight=5
	gamma=2.7793
RF	n_estimators=693
	max_depth=6
	min_samples_split=9
	min_samples_leaf=8
	max_features=log2
NN	hidden_layer_sizes=(64,32)
	dropout=0.4200
	learning_rate=0.0034
	alpha=7.46e-04
	final_epochs=15

Table 7 – Best hyperparameters identified by Optuna for the negative reviews only feature-enhanced configuration corresponding to the best-performing classifier setting.

associated with RB behavior while reducing noise introduced by unrelated reviews. Under this configuration, the best results for all three classifiers were obtained with the lowest negativity threshold (0.10). This suggests that more aggressive filtering removes not only non-RB windows but also potentially useful contextual information and minority-class examples. Although higher thresholds increase the concentration of negative reviews, they simultaneously reduce the number of available RB instances, making generalization more difficult.

By comparing the classifiers, we can see that XGBoost consistently achieved the best balance between precision and recall for the RB class. Under the best configuration, it obtained $P_1 = 0.6840$ and $R_1 = 0.6944$, indicating a balanced ability to identify RB periods while maintaining a relatively low false-positive rate. In contrast, Random Forest and the neural network generally favored recall over precision, identifying a larger number of RB windows at the cost of increased false positives. Indicating that XGBoost was more effective at learning decision boundaries capable of separating RB and non-RB periods within the proposed feature space.

Tables 6 and 7 report the best hyperparameter configurations identified by Optuna for the baseline and proposed representations, respectively. The best-performing XGBoost model relied on relatively low learning rates combined with deeper trees and regularization

mechanisms, suggesting that RB detection requires modeling complex nonlinear interactions between semantic and temporal features while controlling overfitting.

Overall, the results demonstrate that the proposed temporal representation provides a substantial improvement over semantic embeddings alone. The increase in $F1_1$ from 0.1661 in the best baseline model to 0.6779 in the best proposed configuration highlights the contribution of the engineered semantic, structural, and temporal features. Nevertheless, RB detection remains a challenging task due to severe class imbalance, the temporal sparsity of RB events, and the diversity of behaviors associated with different episodes. Even under the best-performing configuration, distinguishing RB periods from legitimate waves of negative feedback remains a difficult classification problem.

9 Conclusion

This study provides a large-scale, longitudinal characterization of review bombing on Steam, analyzing over 103 million reviews to understand the sociotechnical dynamics of coordinated digital protest. By moving beyond the prevailing view of RB as a mere statistical anomaly, our findings establish the phenomenon as a sophisticated form of user agency that blends protest, platform affordances, and collective identity. Across 89 validated episodes, RB consistently emerges as a short-lived, temporally bounded burst of negative reviews that follows a recognizable structure of external coordination, on-platform mobilization, and concentrated execution (**RQ1**), unified by a shared behavioral signature: sudden spikes in review volume, sharp increases in negativity, and a recurring "protest lexicon" built around refund requests and fraud accusations.

RB episodes also display temporal, behavioral, and game-level signatures that distinguish them from organic feedback (**RQ2**). Temporally, they are predominantly post-release phenomena unrelated to launch quality, with 87.6% occurring more than a year after release, and they sharply deviate from otherwise stable review trajectories. Behaviorally, participants resemble engaged players in playtime and activity but show more polarized recommendation histories than typical reviewers, and rarely appear across multiple episodes, suggesting situational mobilization rather than a persistent core of attackers. At the game level, RB clusters around specific developers, genres—particularly "Action" and "Indie" titles—and external controversies, frequently targeting well-received titles and producing intense yet localized surges of negativity. These patterns indicate that RB is neither random noise nor purely automated activity, but a coordinated and temporally bounded form of collective action with identifiable signatures.

Our topic modeling analysis further shows that players strategically weaponize the review system to pressure developers, addressing **RQ3**. Beneath a dominant surface layer of commercial complaints—refund demands and fraud accusations, which together account for roughly 75% of mapped negative reviews—episodes are frequently driven by geopolitical disputes, national identity grievances, opposition to monetization or platform-exclusivity decisions, and ideological conflicts involving gender, diversity, and representation. Many episodes function as vehicles for coordinated harassment and hate speech, particularly when targeting inclusion initiatives, turning the review system into a tool for enforcing exclusionary norms. Critically, the coexistence of cases such as *Feed the Cups*, where RB sanctioned sexist remarks by a developer, and *Total War: ROME II*, where positive-review RB propagated misogynistic backlash against female representation, demonstrates that RB is ideologically flexible and cannot be characterized as inherently progressive or reactionary.

Our machine learning experiments address **RQ4** and **RQ5** while highlighting the difficulty of automated RB detection and contributing a formal framework to computationally represent the phenomenon. Semantic embeddings alone proved insufficient, yielding an F1-score below 0.17 for the RB class in the baseline configuration. Once temporal, behavioral, and semantic-deviation features were incorporated, and feature extraction was restricted to negative reviews, XGBoost reached an F1-score of 0.677—a substantial improvement, though still far from a solved problem, confirming that automated detection is feasible but remains genuinely difficult under real-world class imbalance and discourse heterogeneity. SHAP analyses consistently identify semantic distance from a game's historical review centroid, the concentration of negative reviews within a window, and abrupt semantic dissimilarity relative to the immediately preceding window as the strongest predictive signals, indicating that RB is best captured not by any single indicator, but by the joint occurrence of a topical rupture and a polarity spike within a short time frame. Rather than classifying isolated reviews or games, our approach models RB as coordinated collective action unfolding over time, offering a foundation for moderation strategies that move beyond reactive, volume-based filtering toward a deeper understanding of the temporal and semantic signatures of these events.

9.1 Limitations

A few limitations, already noted throughout this work, should be considered when interpreting our results. First, our ground truth depends on Steam's proprietary "off-topic" moderation criteria, which may both miss coordinated manipulation that evades detection and include episodes that diverge from the operational definition of RB adopted in this study; although we mitigated this through manual qualitative validation, this process inevitably introduced researcher bias. Second, all quantitative findings are derived exclusively from Steam, whose purchase-gated review system may produce dynamics that do not generalize to open-reviewing platforms such as Metacritic, Amazon, or Yelp. Finally, although all topic extraction and clustering were performed automatically by BERTopic, we relied on ChatGPT (GPT-5.5) to generate topic labels and translate non-English reviews into English for interpretation. While this substantially improved readability and consistency, LLM-assisted labeling and translation may introduce additional linguistic and interpretative biases.

9.2 Future Work

We identify three main directions for future research. The first is **cross-platform data collection**, replicating our pipeline on other review ecosystems—Metacritic, the Epic Games Store, Amazon, and, beyond gaming, Rotten Tomatoes, IMDb, Yelp, and

Google Maps—to test whether the signatures identified here generalize beyond Steam's purchase-gated review model. The second is developing **labeling strategies independent of platform moderation**, for instance by cross-referencing statistical anomaly detection in review streams with independent evidence of external coordination from Reddit, Discord, or X/Twitter. The third is the development of **more robust and temporally-aware classifiers**, including temporal/sequential variants of XGBoost that incorporate rolling and lagged window features, recurrent andattention-based deep learning architectures (LSTMs, Temporal Convolutional Networks, andTransformer-based sequence models) capable of learning long-range dependencies directly from review streams, and graph-based approaches connecting users, reviews, and games, which prior work in adjacent domains has shown to be effective but which remain largely unexplored for RB specifically. Finally, we plan to publicly release an anonymized version of the dataset used in this study, enabling reproducibility and supporting future research on review bombing while preserving user privacy.

Bibliography

ADAMS, R. N. **Book of Demons Review Bombed Over Pro-Ukrainian Actions**. 2022. Disponível em: <https://techraptor.net/gaming/news/book-of-demons-review-bombed-over-ukraine-support/>.

ADAMS, R. N. **The Witcher 3, Cyberpunk 2077 Review Bomb Underway**. 2022. Disponível em: <https://techraptor.net/gaming/news/witcher-3-cyberpunk-2077-review-bomb-underway/>.

AKIBA, T. et al. Optuna: A next-generation hyperparameter optimization framework. In: **Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining**. [S.l.: s.n.], 2019. p. 2623–2631.

ARIF, S. **Total War: Rome 2 Is Being Review Bombed on Steam over Imagined Controversy**. 2018. Disponível em: <https://www.ign.com/articles/2018/09/25/total-war-rome-2-is-being-review-bombed-on-steam-over-imagined-controversy/>.

BAEZA-YATES, R.; RIBEIRO-NETO, B. et al. **Modern information retrieval**. [S.l.]: ACM press New York, 1999. v. 463.

BAOWALY, M. K.; TU, Y.-P.; CHEN, K.-T. Predicting the helpfulness of game reviews: A case study on the steam store. **Journal of Intelligent & Fuzzy Systems**, SAGE Publications Sage UK: London, England, v. 36, n. 5, p. 4731–4742, 2019.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, n. 1, p. 5–32, 2001.

CAI, Y. et al. Detecting spam movie review under coordinated attack with multi-view explicit and implicit relations semantics fusion. **IEEE Transactions on Information Forensics and Security**, IEEE, v. 19, p. 7588–7603, 2024.

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). **Quality & Quantity**, Springer, Dordrecht, Netherlands, v. 59, n. 1, p. 315–341, fev. 2025. ISSN 1573-7845. Disponível em: <https://doi.org/10.1007/s11135-024-01981-z/>.

CARTER, G. **Metacritic Brings Down The Hammer On "Review Bombers"**. 2011. Disponível em: <https://www.escapistmagazine.com/metacritic-brings-down-the-hammer-on-review-bombers/>.

CARTER, J. **Overwatch 2 Enters Steam as a Review-Bombed Top Seller**. 2023. Disponível em: <https://www.gamedeveloper.com/business/-i-overwatch-2-i-enters-steam-as-a-review-bombed-top-seller>.

CHALK, A. **Total War: Rome 2 is getting review-bombed on Steam because of women generals (Updated)**. 2018. Disponível em: <https://www.pcgamer.com/total-war-rome-2-is-getting-review-bombed-on-steam-because-of-women-generals/>.

CHALK, A. **Players protest Epic's Metro Exodus exclusive by review-bombing the series on Steam**. 2019. Disponível em: <https://www.pcgamer.com/metro-review-bomb-steam/>.

CHALK, A. **Former Tripwire CEO tells Tucker Carlson that cancel culture 'destroyed' him**. 2022. Disponível em: <https://www.pcgamer.com/former-tripwire-ceo-tells-tucker-carlson-that-cancel-culture-destroyed-him/>.

CHATZAKOU, D. et al. Hate is not binary: Studying abusive behavior of #gamergate on twitter. In: **Proceedings of the 28th ACM Conference on Hypertext and Social Media**. NY: ACM, 2017. (HT '17), p. 65–74. ISBN 9781450347082.

CHATZAKOU, D. et al. Measuring #gamergate: A tale of hate, sexism, and bullying. In: **Proceedings of the 26th International Conference on World Wide Web**. [S.l.: s.n.], 2017. (WWW '17), p. 1285–1290. ISBN 9781450349147.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: Association for Computing Machinery, 2016. (KDD '16), p. 785–794. ISBN 9781450342322. Disponível em: <https://doi.org/10.1145/2939672.2939785>.

CHESS, S.; SHAW, A. A conspiracy of fishes, or, how we learned to stop worrying about# gamergate and embrace hegemonic masculinity. **Journal of Broadcasting & Electronic Media**, Taylor & Francis, v. 59, n. 1, p. 208–220, 2015.

CORONADO-BLÁZQUEZ, J. A nlp approach to" review bombing" in metacritic pc videogames user ratings. **arXiv preprint arXiv:2405.06306**, 2024.

DAMMANN, L. **Tekken 8 is Being Review Bombed Yet Again**. 2024. Disponível em: <https://gamerant.com/tekken-8-review-bombed-why/>.

FENLON, W. **Steam Review Bombing Is Working, and Chinese Players Are a Powerful New Voice**. 2017. Disponível em: <https://www.pcgamer.com/steam-review-bombing-is-working-and-chinese-players-are-a-powerful-new-voice/>.

FERGUSON, C. J.; GLASGOW, B. Who are gamergate? a descriptive study of individuals involved in the gamergate controversy. **Psychology of Popular Media**, Educational Publishing Foundation, v. 10, n. 2, p. 243, 2021.

FINE, T. L. **Feedforward Neural Network Methodology**. New York, NY, USA: Springer, 1999. ISBN 0-387-98745-2.

GRAMUGLIA, A. **Why The CW's Batwoman Is Being Bombed With Bad Reviews**. 2019. Disponível em: <https://www.cbr.com/batwoman-swarmed-with-review-bombs/>.

GROOTENDORST, M. **BERTopic: Neural topic modeling with a class-based TF-IDF procedure**. 2022. Disponível em: <https://arxiv.org/abs/2203.05794>.

HALL, C. **Playerunknown's Battlegrounds latest victim of review bomb on Steam**. 2017. Disponível em: <https://www.polygon.com/2017/10/2/16402414/playerunknowns-battlegrounds-review-bomb-chinese-ads-vpn/>.

HALL, C. **Call of Duty: Modern Warfare review bombed over Russian portrayal**. 2019. Polygon. Online; Publicado em 28 de Outubro de 2019. Disponível em: <https://www.polygon.com/2019/10/28/20936496/call-of-duty-modern-warfare-review-bombing-russian-military-highway-of-death>. Acesso em: 27 de maio de 2024.

HERNANDEZ, P. **Popular YouTuber Says He Won't Cover Game After Twitter Spat**. 2015. Disponível em: <https://kotaku.com/popular-youtuber-says-he-wont-cover-game-after-twitter-1698242976>.

HOWE, W. T.; LIVINGSTON, D. J.; LEE, S. K. Is #NotMyBattlefield Rooted in Gamer Identity? An Examination of Demographic Factors, Genre Preference, and Technology Use of Gamers. In: **Proceedings of the 52nd Hawaii International Conference on System Sciences**. Hawaii, USA: IEEE, 2019. (HICSS'19).

IGGY. **Pokemon Sword And Shield Review Bombed On Metacritic**. 2019. Disponível em: <https://nintendosoup.com/pokemon-sword-and-shield-review-bombed-on-metacritic/>.

Imbalanced Learn Developers. **TomekLinks**. 2025. Disponível em: <https://imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.TomekLinks.html>.

Imbalanced Learn Developers. **SMOTE**. 2026. Disponível em: <https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html>.

JAMES, K. **Grand Theft Auto: The Trilogy – The Definitive Edition Review – As Controversial as Ever**. 2021. Disponível em: <https://checkpointgaming.net/reviews/2021/11/grand-theft-auto-the-trilogy-the-definitive-edition-review-as-controversial-as-ever/>.

JONES, K. S. A statistical interpretation of term specificity and its application in retrieval. **Journal of documentation**, MCB UP Ltd, v. 28, n. 1, p. 11–21, 1972.

JUNIOR, S. M. d. S. et al. **Estudo do fenômeno de review bombing com processamento de linguagem natural**. Dissertação (Mestrado) — Universidade Tecnológica Federal do Paraná, 2025.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: **Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995. (IJCAI'95), p. 1137–1143. Disponível em: <https://dl.acm.org/doi/10.5555/1643031.1643047>.

KRAEMER, T.; WEIGER, W. H.; HEIDENREICH, S. Do all stars shine the same? investigating the nonlinear effects of user and critic reviews on video game sales. **Journal of Business Research**, Elsevier, v. 188, p. 115034, 2025.

KREMERS, R. **How are Steam review scores calculated?** 2026. Disponível em: <https://www.omni-labs.com/blog/how-are-steam-review-scores-calculated>.

KUCHERA, B. **Gamers fight back against lackluster Spore gameplay, bad DRM**. 2008. Disponível em: <https://arstechnica.com/gaming/2008/09/gamers-fight-back-against-lackluster-spore-gameplay-bad-drm/>.

KUCHERA, B. **The anatomy of a review bombing campaign**. 2017. Polygon. Disponível em: <https://www.polygon.com/2017/10/4/16418832/pubg-firewatch-steam-review-bomb>.

LETIZI, R.; NORMAN, C. "you took that from me": Conspiracism and online harassment in the alt-fandom of the last of us part ii. **Games and culture**, SAGE Publications Sage CA: Los Angeles, CA, v. 19, n. 4, p. 513–534, 2024.

LOUREIRO, J. **O review bombing de The Last of Us: Part 2 já começou no Metacritic**. 2020. Disponível em: <https://www.eurogamer.pt/o-review-bombing-de-the-last-of-us-part-2-ja-comecou-no-metacritic>.

LOWRY, B. **Helldivers 2 Players Review Bomb the Game Down to 'Mostly Negative' on Steam as Frustrations Mount**. 2026. Disponível em: <https://tech.yahoo.com/gaming/articles/helldivers-2-players-review-bomb-204817060.html>.

LU, C. et al. Patches and player community perceptions: Analysis of no man's sky steam reviews. In: **Proceedings of DiGRA 2020 conference: Play everywhere**. [S.l.: s.n.], 2020.

LUNA, A. **Review Bombing: A Look at the All-Female Ghostbusters**. 2021. Disponível em: <https://insight.balancenow.co/review-bombing-a-look-at-the-all-female-ghostbusters/>.

LUNDBERG, S. M. et al. From local explanations to global understanding with explainable ai for trees. **Nature machine intelligence**, Nature Publishing Group UK London, v. 2, n. 1, p. 56–67, 2020.

MAATEN, L. Van der; HINTON, G. Visualizing data using t-sne. **Journal of machine learning research**, v. 9, n. 11, 2008.

MARTES, D. O. et al. Transformer models for paraphrase detection: A comprehensive semantic similarity study. **Computers**, MDPI, v. 14, n. 9, p. 385, 2025.

MCINNES, L. **UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction**. 2018. Disponível em: <https://umap-learn.readthedocs.io/en/latest/>.

MCINNES, L.; HEALY, J.; ASTELS, S. **The hdbscan Clustering Library**. 2016. Disponível em: <https://hdbscan.readthedocs.io/en/latest/>.

MCKEAND, K. **A Brief History of How Steam Review Bombing Damages Developers**. 2017. Disponível em: <https://www.pcgamesn.com/history-of-steam-review-bombing>.

MCMAHON, C. **No, the Majority of Star Wars "Fans" Did Not Hate the Last Jedi**. 2020. Disponível em: <https://colinmcmahonauthor.com/2020/01/18/no-the-majority-of-star-wars-fans-did-not-hate-the-last-jedi/>.

MELO, P.; PIMENTEL, C. A campanha de Ódio contra the last of us part ii. In: **Anais Estendidos do XXI Simpósio Brasileiro de Jogos e Entretenimento Digital**. Porto Alegre, RS, BRA: Sociedade Brasileira de Computação, 2022. (SBGames '22), p. 428–437. Disponível em: <https://doi.org/10.5753/sbgames_estendido.2022.226031>.

MORTENSEN, T. E. Anger, fear, and games: The long event of# gamergate. **Games and Culture**, SAGE Publications Sage CA: Los Angeles, CA, v. 13, n. 8, p. 787–806, 2018.

MUNNA, M. H.; RIFAT, M. R. I.; BADRUDDUZA, A. S. M. Sentiment analysis and product review classification in e-commerce platform. In: **23rd International Conference on Computer and Information Technology**. New York, NY, USA: IEEE, 2020. (ICCIT '20), p. 1–6. Disponível em: <https://doi.org/10.1109/ICCIT51783.2020.9392710>.

PARZEN, E. On estimation of a probability density function and mode. **The annals of mathematical statistics**, JSTOR, v. 33, n. 3, p. 1065–1076, 1962.

PAYNE, W. B. Review bombing the platformed city: Contested political speech in online local reviews. **Big Data & Society**, SAGE Publications Sage UK: London, England, v. 11, n. 3, p. 20539517241275879, 2024.

PEARSON, K. Vii. note on regression and inheritance in the case of two parents. **Proceedings of the Royal Society of London**, v. 58, n. 347-352, p. 240–242, 12 1895. ISSN 0370-1662. Disponível em: <https://doi.org/10.1098/rspl.1895.0041>.

PINHEIRO, P. **Assim como o jogo, 2º ano de The Last of Us está recebendo "review bombing"**. 2025. Disponível em: <https://jovemnerd.com.br/noticias/series-e-tv/2o-ano-the-last-of-us-recebendo-review-bombing>.

PLUNKETT, L. **Steam Is 10 Today. Remember When It Sucked?** 2013. Disponível em: <https://kotaku.com/steam-is-10-today-remember-when-it-sucked-1297594444/>.

PLUNKETT, L. **War Thunder 'Revises' Economy, Fans Review-Bomb Game To Hell**. 2023. Disponível em: <https://kotaku.com/war-thunder-f2p-pc-protest-steam-pc-review-economy-p2w-1850468060/>.

PONISZEWSKA-MARAŃDA, A.; HYNASIŃSKI, P.; BOROWSKA, B. Detection techniques of review bombing through nlp and machine learning–literature review. 2026.

RACINOWSKA, O. **Paradox games under fire. Review bombing from Chinese players hits Stellaris, Crusader Kings and Europa Universalis**. 2025. Disponível em: <https://www.gamepressure.com/newsroom/paradox-games-under-fire-review-bombing-from-chinese-players-hits/zd799d/>.

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 and the 9th International Joint Conference on Natural Language Processing**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019. (EMNLP–IJCNLP '19), p. 3982–3992. Disponível em: <https://doi.org/10.18653/v1/D19-1410>.

RIBEIRO, M. V. G.; PIMENTEL, C. A.; MELO, P. de F. Quando as avaliações viram bombas: Explorando a dinâmica do review bombing nos jogos no metacritic. In: SBC. **Brazilian Symposium on Multimedia and the Web (WebMedia)**. [S.l.], 2024. p. 249–256.

SALDANHA, L.; SILVA, S. M. da; FERREIRA, P. D. "community" in video game communities. **Games and Culture**, SAGE Publications Sage CA: Los Angeles, CA, v. 18, n. 8, p. 1004–1022, 2023.

SANDERS, G. **Review Bombing: The Toxic Tactic That Doesn't Work**. 2022. Disponível em: <https://medium.com/itsfullofstars/review-bombing-the-toxic-tactic-that-doesnt-work-4c75fd269004>.

SCHUFF, D.; MUDAMBI, S.; WANG, M. Understanding the review bombing phenomenon in movies and television. In: **Proceedings of the 57th Hawaii International Conference on System Sciences**. Honolulu, Hawaii: Hawaii International Conference on System Sciences, 2024. (HICSS '24), p. 299–307. Disponível em: <https://doi.org/10.24251/HICSS.2024.036>.

SCULLEY, D. Web-scale k-means clustering. In: **Proceedings of the 19th international conference on World wide web**. [S.l.: s.n.], 2010. p. 1177–1178.

SHEN, R.-P.; LIU, D.; SHEN, H.-S. Detecting review manipulation from behavior deviation: A deep learning approach. **Electronic Commerce Research and Applications**, Elsevier, v. 60, p. 101283, 2023.

SIMS, D. **A Change for Rotten Tomatoes Ahead of Captain Marvel**. 2019. Disponível em: <https://www.theatlantic.com/entertainment/archive/2019/03/rotten-tomatoes-captain-marvel-review-ratings-system-online-trolls/584032/>.

SIUDA, P. et al. Broken promises marketing. relations, communication strategies, and ethics of video game journalists and developers: The case of cyberpunk 2077. **Games and Culture**, SAGE Publications Sage CA: Los Angeles, CA, v. 19, n. 5, p. 650–669, 2024.

STEAM. **User Reviews Revisited**. 2019. Disponível em: <https://steamcommunity.com/games/593110/announcements/detail/1808664240333155775/>.

Steam. **Steam Charts**. 2025. Disponível em: <https://steamdb.info/app/753/charts/>.

STEAM. **Steam User Reviews FAQ**. 2026. Disponível em: <https://help.steampowered.com/faqs/view/2DA6-9CB3-F84A-643E>.

STRINGHINI, G.; BLACKBURN, J. Understanding the phases and themes of coordinated online aggression attacks. In: **Social Processes of Online Hate**. Routledge, 2024. cap. 11. Disponível em: <https://doi.org/10.4324/9781003472148-11>.

TASSI, P. **The 'Warcraft 3: Reforged' Metacritic Review Bombing Campaign Has Escalated**. 2020. Disponível em: <https://www.forbes.com/sites/paultassi/2020/02/02/the-warcraft-3-reforged-metacritic-review-bombing-campaign-has-escalated/>.

TASSI, P. **'Rings Of Power' Is Getting Review Bombed So Hard Amazon Suspended Reviews Entirely**. 2022. Disponível em: <https://www.forbes.com/sites/paultassi/2022/09/05/rings-of-power-is-getting-review-bombed-so-hard-amazon-suspended-reviews-entirely/>.

THORHAUGE, A. M. The steam platform economy: From retail to player-driven economies. **New Media & Society**, SAGE Publications Sage UK: London, England, v. 26, n. 4, p. 1963–1983, 2024.

U/RANDOMDRAGONSLAYER. **Help | I need to get refund for a game because the devs are sexist as hell**. 2024. Disponível em: <https://www.reddit.com/r/GirlGamers/comments/1d0pqn1/help_i_need_to_get_refund_for_a_game_because_the/>.

VÄLISALO, T.; RUOTSALAINEN, M. “sexuality does not belong to the game”: Discourses in overwatch community and the privilege of belonging. **Game Studies: The international journal of computer game research**, IT-Universitetet i København, n. 3, 2022.

WALKER, I. **Trolls Review Bomb Cyberpunk 2077 After Devs Speak Out Against Russia’s Invasion Of Ukraine**. 2022. Disponível em: <https://www.theguardian.com/world/2020/dec/07/monster-hunter-film-pulled-china-scene-sparks-backlash/>.

WANG, J.; WU, C. Camouflage is not easy: Uncovering adversarial fraudsters in large online app review platform. **Measurement and Control**, SAGE Publications Sage UK: London, England, v. 53, n. 9-10, p. 2137–2145, 2020.

WATERCUTTER, A. **The Acolyte and the Long-Awaited Death of Review-Bombing**. 2024. Disponível em: <https://www.wired.com/story/the-acolyte-and-the-long-awaited-death-of-review-bombing/>.

WIKIPEDIA. **Review bomb**. 2026. Disponível em: <https://en.wikipedia.org/wiki/Review_bomb/>.

Wikipedia Contributors. **List of Review-Bombing Incidents**. 2026. Disponível em: <https://en.wikipedia.org/wiki/List_of_review-bombing_incidents>.

WILLIAMS, H. **Gran Turismo 7 Continues To Be Review Bombed On Metacritic**. 2022. Disponível em: <https://www.gamespot.com/articles/gran-turismo-7-continues-to-be-review-bombed-on-metacritic/1100-6501763/>.

WIRZBURGER, A. Platforms and the governmentality of political consumerism: A critical analysis of review bombing and de/politicization on steam. **Platforms & Society**, SAGE Publications, Thousand Oaks, CA, USA, v. 2, p. 1–16, maio 2025. ISSN 2976-8624. Disponível em: <https://doi.org/10.1177/29768624251341217>.

ZHANG, L. et al. Gaim: graph-aware feature interactional model for spam movie review detection. In: IEEE. **2022 26th international conference on pattern recognition (ICPR)**. [S.l.], 2022. p. 621–628.