Steam Games Insight Engine

Khang Nguyen, Dylan Shah, Blair Lee, Jun Ze He, Mehwish Ahmed

May 2025

Abstract

While Steam offers both data-driven statistics on customer satisfaction and community-based game reviews, both can lack the nuanced details and comprehensive analysis that would help interested gamers make their decisions. To this end, we present the Steam Games Insight Engine (SGIE), the first unified framework combining credibility detection, multi-class emotion analysis, and aspect-based sentiment analysis for gaming reviews. Our approach uses user data about playtime and up votes to filtering reviews followed by an ensemble of fine-tuned GPT-2 Medium, RoBERTa, and DeBERTa models for comprehensive analysis. We achieved 93.68% accuracy in emotion classification with our ensemble approach, 99.85% accuracy in credibility detection using KNN with PCA, and comprehensive aspect-based analysis revealing predominantly positive sentiment across all gaming aspects. The ensemble method with disagreement penalty provides improved confidence calibration, with models agreeing on 60.1% of cases while appropriately reducing confidence during disagreement. This multi-dimensional approach provides the first comprehensive solution for authentic, emotionally-aware game review analysis at scale.

1 Introduction

Steam, Valve Corporation's digital distribution platform and the world's largest PC gaming platform [19], hosts over 132 million monthly active users [19] and more than 100,000 games [18], generating millions of user reviews that create an overwhelming information landscape for consumers [12, 16]. While Steam provides basic filtering mechanisms such as binary recommendation indicators ("X% find this helpful") or temporal sorting, these mechanisms fail to capture the nuanced emotional responses and specific game aspects that influence purchasing decisions [12, 25]. This poses a central question for our paper: *How can we automatically extract comprehensive, credible insights from gaming reviews that capture both emotional nuance and aspect-specific feedback to support informed player decision-making?*

To answer that question, we present the Steam Games Insight Engine (SGIE), a unified framework that combines credibility detection, multi-class emotion analysis, and aspect-based sentiment analysis to provide comprehensive gaming review insights. Our approach first employs isolation forest techniques on reviewer behavioral patterns to filter suspicious content [13], then applies an ensemble of fine-tuned transformer models (GPT-2, RoBERTa, and DeBERTa) to extract six-dimensional emotional profiles and aspect-specific sentiment from authenticated reviews. This multi-stage pipeline addresses the interconnected challenges of review authenticity, emotional nuance, and feature-specific feedback that existing single-dimension approaches cannot handle effectively [21, 6]. Our evaluation demonstrates significant improvements over baseline approaches, with implications for both consumer decision-making and developer feedback analysis.

2 Background

Limitations of Current Sentiment Analysis Traditional sentiment analysis approaches face fundamental limitations when applied to gaming contexts. Sentiment analysis research has predominantly focused on binary positive/negative classification, with systematic reviews showing 81.3% accuracy for binary approaches versus only 60.2% for multi-class emotion detection [1]. This binary focus proves inadequate for gaming contexts, however, as traditional models achieve only 0.70 AUC on game review datasets due to gaming-specific challenges including sarcasm, contextual references, and comparative statements [4].

Missing Credibility Integration Current sentiment analysis systems fundamentally lack integrated fake review detection mechanisms [6]. Recent research has explicitly documented that "existing sentiment-based detection systems fail to capture consumer feelings" and notes "significant impact on sentiment scores used to identify fake feedback" when credibility filtering is applied post-hoc [24]. This sequential approach creates systematic bias, as sentiment analysis operates on potentially manipulated data before credibility assessment occurs. The problem becomes severe for gaming platforms, where coordinated manipulation campaigns can skew sentiment distributions, sometimes without detection systems activating at all [3].

Gaming Platform Manipulation Gaming platforms face widespread review manipulation beyond individual fake reviews. Academic analysis documents major review bombing incidents involving thousands of coordinated suspicious reviews [3], while platform operators such as Valve have implemented specialized detection systems to combat organized campaigns. These manipulation patterns create systematic bias that traditional sentiment analysis cannot detect, as the coordinated nature of attacks differs qualitatively from random fake reviews.

Absence of Gaming-Specific Aspect-Based Sentiment Analysis (ABSA) A systematic review of 727 ABSA studies (2008-2024) reveals gaming as notably absent from domain coverage, with 53 of 62 public ABSA datasets focused on restaurants, hotels, and digital products instead [21]. Further research has shown that gaming-specific aspects - graphics quality, gameplay mechanics, narrative design, and community dynamics - require specialized handling that existing ABSA frameworks do not provide [25], especially as terminology like "frame drops," "boss mechanics," or "skill trees" creates a fundamental domain adaptation challenge that conventional ABSAs cannot provide.

Research Gap and Motivation The convergence of limitations stated above creates a comprehensive gap requiring integrated solutions. Existing approaches cannot provide the multi-dimensional insights necessary for informed gaming decisions or actionable developer feedback. This gap motivates our unified framework approach, where credibility detection, multi-class emotion analysis, and gaming-specific aspect analysis operate as integrated components rather than isolated systems.

3 Methodology

3.1 Web Scraping

We created the web scraper with the bs4, langdetect, and request libraries [11, 5, 15]. The web scraper collects game reviews from Steam based on `gameID` filtering for English content and capturing user metadata. In particular, for every given comment, we collect total playtime for the given game, number of games owned, comment content and votes, as well as timestamp created for that comment. Comment and player ID are removed for privacy and ethical purposes,

and timestamps were collected at the review level but later removed from any analysis due to time constraints. If utilized, the timestamps would be subsequently aggregated into weekly bins to ensure user anonymization.

3.2 Emotion-based Sentiment Analysis

3.2.1 Training Framework Architecture

To ensure consistent training procedures in all three transformer models while accommodating their architectural differences, we implement a modular training framework centered on a base trainer class. The `base_trainer.py` module provides a unified interface that encapsulates common training functionality and implements critical components shared across all models.

Model-specific trainers (`gpt2_trainer.py`, `roberta_trainer.py`, and `deberta_trainer.py`) then inherit from this base class and override only the components that require architecture-specific handling. This inheritance-based design ensures that training hyperparameters, evaluation protocols, and optimization strategies remain consistent across all models, which alleviates the difficulties of the ensemble learning setup while maintaining the architectural diversity necessary for effective ensemble learning.

3.2.2 Transformer Architecture Selection

Building upon this unified training framework, we carefully selected three complementary transformer architectures: GPT-2 Medium (355M parameters), RoBERTa Base (125M parameters), and DeBERTa-v3 Base (184M parameters). These architectures represent distinct paradigms in transformer design that offer complementary strengths for emotion classification. GPT-2 brings autoregressive understanding that excels at capturing sequential narrative flow, RoBERTa provides optimized bidirectional encoding for robust contextual representations, and DeBERTa contributes enhanced positional understanding through sophisticated attention mechanisms. This architectural diversity enables our ensemble to capture different aspects of emotional expression while maintaining computational efficiency.

The transformer architecture underlying all three models leverages self-attention mechanisms to capture contextual relationships between tokens [20]. This design allows transformers to process sequences in parallel while building rich representations that capture long-range dependencies and complex contextual relationships. For emotion classification, this capability proves particularly valuable as emotional expression rarely depends on individual words in isolation, but emerges instead from complex interactions between words, phrases, and contextual cues that may be distributed throughout the text.

3.2.3 Low-Rank Adaptation (LoRA) Strategy

We employ Low-Rank Adaptation (LoRA) for efficient fine-tuning across all models in our ensemble [8]. LoRA addresses several critical challenges in training large transformer ensembles: (1) memory constraints that make full fine-tuning of 355M+ parameter models prohibitively expensive, (2) computational costs that would make training three large models simultaneously infeasible, (3) catastrophic forgetting that can degrade pre-trained representations during aggressive fine-tuning, and (4) overfitting risks associated with updating millions of parameters on relatively small task-specific datasets.

LoRA introduces trainable low-rank matrices to approximate weight updates, significantly reducing the number of parameters requiring fine-tuning. For each target weight matrix $W \in \mathbb{R}^{d\times d}$, LoRA introduces two smaller matrices $A \in \mathbb{R}^{d\times r}$ and $B \in \mathbb{R}^{r\times d}$, where $r \ll d$, along with a scaling factor α. The modified weight becomes: $\tilde{W} = W + \frac{\alpha}{r}AB$. This formulation provides

substantial computational advantages, reducing memory overhead from $O(d^2)$ to $O(rd)$ which is crucial for consumer-grade GPUs where VRAM is scarce.

We apply LoRA to both attention mechanisms (query, key, and value projections) and feed-forward layers, using rank $r = 16$ and scaling factor $\alpha = 16$ across all models for optimal performance-efficiency trade-off. This configuration reduces trainable parameters while maintaining sufficient capacity for effective task learning: GPT-2 Medium trains 16.31% of its parameters (58.9M of 361M), RoBERTa Base trains 33.22% (42.3M of 127M), and DeBERTa-v3 Base trains 54.56% (102M of 187M). By preserving the original pre-trained weights frozen and only updating the low-rank adaptation matrices, LoRA maintains the knowledge acquired during pre-training while enabling efficient task-specific adaptation, ensuring feasible training times and negating any possible memory overflow.

3.2.4 GPT-2 Implementation

The first component of our ensemble is GPT-2 Medium, a decoder-only transformer model with 355M parameters developed by OpenAI that focuses on next-word prediction through causal attention [17]. Unlike bidirectional models, GPT-2 only considers previous and current inputs when calculating attention scores, making it particularly effective for understanding sequential dependencies and narrative flow in text—crucial aspects for interpreting emotional progression in reviews.

To adapt GPT-2 for emotion classification, we replace the original language modeling head with a classification head that projects the final hidden state to six emotion categories. This adaptation leverages GPT-2's autoregressive understanding while redirecting its output toward discrete emotional classification rather than text generation.

3.2.5 RoBERTa and DeBERTa Integration

Complementing GPT-2's autoregressive strengths, we incorporate two bidirectional transformer models that provide different perspectives on contextual understanding. RoBERTa Base (125M parameters) represents an optimized approach to BERT's bidirectional encoding [14]. RoBERTa improves upon the original BERT architecture through refined training procedures, including removal of the Next Sentence Prediction task, training with larger batches, and implementation of dynamic masking. These optimizations result in more robust contextual representations while maintaining computational efficiency suitable for ensemble deployment.

The third component, DeBERTa-v3 Base (184M parameters), introduces sophisticated positional understanding through disentangled attention mechanisms [7]. DeBERTa separately encodes content and position information, allowing the model to better capture relative positional relationships between tokens—particularly valuable for understanding contextual nuances and subtle emotional cues that depend on word positioning and syntactic structure.

3.2.6 Ensemble Training Configuration

With our architectural framework and model selection established, we implement a comprehensive training protocol designed to optimize both the performance of the individual model and the effectiveness of the ensemble. We trained our transformer-based models using the Emotion Dataset from Kaggle [10]. All models are trained with carefully tuned hyperparameters that balance training efficiency with model performance, ensuring fair comparison and optimal ensemble synergy.

We employ a learning rate of 3×10^{-5} with the AdamW optimizer across all architectures, selected through systematic hyperparameter search to optimize convergence while preventing overfitting. Training proceeds for a maximum of 10 epochs with early stopping based on validation loss, allowing models to achieve optimal performance while avoiding overtraining.

Due to memory constraints and architecture-specific computational requirements, we implement differentiated batch sizes: GPT-2 Medium uses batch size 4, while RoBERTa Base and DeBERTa-v3 Base use batch size 6. Gradient accumulation maintains effective batch sizes while accommodating GPU memory limitations, and we implement mixed precision training (FP16) with gradient checkpointing to optimize memory usage during training.

Learning rate scheduling using `ReduceLROnPlateau` (patience = 5 evals, reduction factor = 0.1) allows models to fine-tune more precisely as training progresses. This comprehensive training setup ensures optimal performance for each individual model, while maintaining the computational efficiency and architectural diversity necessary for effective ensemble deployment.

3.2.7 Novel Ensemble Methodology with Disagreement Penalty

Our methodology culminates with a novel ensemble approach that incorporates a disagreement penalty mechanism to improve prediction reliability and confidence calibration. This innovation addresses a critical limitation of conventional ensemble methods: the tendency to maintain artificially high confidence even when individual models disagree significantly, which would indicate underlying prediction uncertainty.

Our disagreement penalty mechanism operates on the fundamental principle that inter-model disagreement signals prediction uncertainty and should appropriately reduce ensemble confidence. Rather than simply averaging predictions, we explicitly quantify and penalize disagreement to produce more reliable and better-calibrated ensemble predictions.

For mathematical clarity, we denote our three models by G (GPT-2), R (RoBERTa), and D (DeBERTa). For each input, we first compute the mean probability distribution in our three complementary models: $p_{\text{mean}} = \frac{1}{3}(P_G + P_R + P_D)$ where $P_i \in \mathbb{R}^6$ represents the probability distribution over the six emotion classes for model i.

We then calculate the disagreement penalty by measuring the pairwise squared differences between the prediction of the model: $\text{penalty} = w_{\text{penalty}} \cdot \frac{1}{3} \sum_{i,j \in \{G,R,D\}, i<j} ||P_i - P_j||_2^2$ where w_{penalty} controls the penalty strength. This formulation captures the extent to which the models disagree, with higher values indicating greater uncertainty that should be reflected in the reduced final confidence.

The final ensemble prediction applies this penalty to appropriately modulate confidence based on the consensus of the model: $P_{\text{final}} = \text{normalize}(\max(p_{\text{mean}} - \text{penalty}, 0))$. This ensures that all probabilities remain non-negative, while normalization maintains a valid probability distribution. This elegant formulation allows for leveraging collective model knowledge while appropriately reducing confidence when consensus is lacking.

Through empirical validation, we determined the optimal penalty weight $w_{\text{penalty}} = 0.5$, which effectively balances accuracy improvements with appropriate confidence calibration. This value penalizes disagreement sufficiently to encourage prediction reliability, while also allowing for legitimate uncertainty in genuinely ambiguous cases.

3.3 Credibility Detection

To identify suspicious or potentially fake reviews, we employed a heuristic labeling approach in the absence of true labels. We developed multiple label sets (y_1 through y_6) representing different levels of strictness in detecting suspicious behavior. y_2 is the most conservative, applying narrow conditions that capture only overtly suspicious behavior (fake rate $\approx 0.2\%$), while y_1 is slightly less conservative ($\approx 1.1\%$ fake rate). Labels y_3 to y_6 incrementally relax thresholds and stack additional rules, labeling up to $\approx 15\%$ of reviews as potentially fake. This multi-threshold approach allows evaluation of model sensitivity to different definitions of suspicious behavior.

To construct these heuristic rules, we examined histograms and percentile-based thresholds of key features including `playtime_forever`, `num_reviews`, `votes_up`, and `weighted_vote_score`. Several suspicious patterns emerged: (1) extremely short playtime (< 10 minutes) paired with

unusually high review scores (> 0.96), indicative of bot behavior, (2) prolific reviewers with hundreds of reviews but minimal community engagement (low upvotes), suggesting spam accounts, (3) users owning hundreds of games but playing very few, associated with collection hoarding rather than genuine gameplay, and (4) overly positive reviews with no community corroboration (high scores but zero upvotes or comments). Each rule was designed to isolate unusual behavioral patterns backed by data distribution analysis, with percentile thresholds formalizing cutoff points for identifying outlier behavior.

After generating heuristic labels, we trained supervised classifiers using K-Nearest Neighbors (KNN) algorithm. Each dataset was split into training (60%), validation (20%), and test (20%) sets, with grid search over $k \in [1, 20]$ to determine optimal neighbors for each label set. Despite the large dataset size, KNN's non-parametric nature and effectiveness with moderate dimensionality (19 features) made it an appropriate baseline for capturing patterns in reviewer behavior without complex model fitting.

To address potential dimensionality issues, we applied Principal Component Analysis (PCA) for dimensionality reduction prior to classification. PCA transforms the feature space into orthogonal components, removing redundant correlations and potentially denoising irrelevant variation while mitigating KNN's degradation in high-dimensional spaces. We retained the smallest number of principal components capturing $> 85\%$ of total variance (consistently 6 components across all label sets) and performed grid search over both PCA components and KNN neighbors to optimize performance.

3.4 Aspect-Based Sentiment Analysis

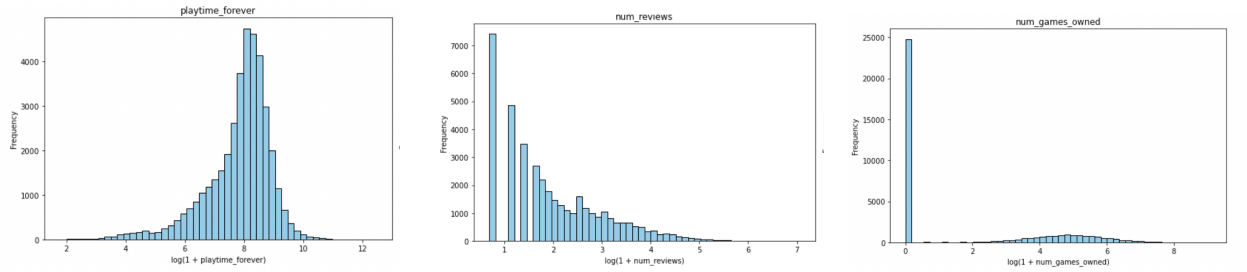
To gain insight into specific aspects of video games, we implemented two complementary approaches for aspect-based sentiment analysis (ABSA): a guided non-negative matrix factorization (NMF) method for unsupervised topic discovery, and a keyword-based pipeline for tracking predefined gaming aspects. This dual approach balances exploratory analysis of emergent themes with targeted monitoring of known game features, providing comprehensive coverage of aspect-specific sentiment patterns in gaming reviews.

The guided NMF approach employs matrix factorization to extract latent topics from review text while incorporating domain knowledge through seed words. In particular, guided NMF reconstructs the document-term matrix A by approximating it as the product of two lower-dimensional matrices W and H [23], iteratively minimizing the reconstruction error $||A - WH||$ while incorporating seed word supervision to guide topic formation toward gaming-relevant aspects. This semi-supervised approach enables discovery of nuanced aspects that may not be captured by predefined categories, while ensuring topical coherence through strategic seed word selection. The extracted topics are then processed through pre-built ABSA models: `deberta-v3-base-abas-v1.1` for aspect extraction and `twitter-xlm-roberta-base-sentiment` for sentiment classification [2, 22].

The keyword-based ABSA pipeline provides a more direct and interpretable approach for tracking sentiment toward specific game features such as graphics, combat, performance, and story. This method leverages curated keyword lists tailored to gaming terminology (e.g., "lag," "fps," "crash" for performance; "graphics," "visuals," "art" for graphics) to identify aspect-relevant sentences, then applies sentiment classification using the DistilBERT model fine-tuned on SST-2 model [9]. To ensure reliability, we filter predictions with confidence scores below 0.6, retaining only high-confidence sentiment assignments. This approach is computationally efficient and easily interpretable, making it well-suited for tracking specific aspects across large datasets and comparing sentiment patterns between different games.

4 Results and Discussion

4.1 Credibility Detection Results



(a) (b) (c)

Figure 1: Histograms of Certain Features

Label	Best k	Validation Accuracy	Test Accuracy
y_1	5	99.62%	99.41%
y_2	1	99.91%	99.85%
y_3	5	98.56%	98.60%
y_4	7	97.58%	97.28%
y_5	3	97.35%	97.30%
y_6	5	97.13%	97.14%

Label	Best k	PCA Components	Val Accuracy	Test Accuracy
y_1	7	6	99.67%	99.49%
y_2	5	6	99.94%	99.85%
y_3	9	6	98.73%	98.68%
y_4	7	6	98.00%	97.73%
y_5	7	6	97.78%	97.59%
y_6	5	6	97.57%	97.64%

Table 1: Validation and test accuracy for KNN after PCA-based dimensionality reduction.

Our heuristic-based credibility detection achieved excellent performance across all suspicious behavior definitions, demonstrating effective behavioral pattern analysis for identifying potentially fake gaming reviews.

Dataset Characteristics: Figure 1 illustrates key feature distributions used in credibility assessment. The histograms reveal distinct behavioral signatures: playtime distributions show moderate gaming hours with clear bot account outliers, review frequency patterns expose prolific reviewers deviating from typical engagement, and community metrics highlight reviews with disproportionately low social validation. These distributional analyses informed our percentile-based threshold selection for suspicious behavior identification.

Classification Performance: Our K-Nearest Neighbors approach with grid search optimization demonstrated robust performance across all heuristic labels. Baseline KNN results (Table 1) show exceptional accuracy ranging from 97.13% to 99.91%, with the most conservative definition (y_2) achieving 99.85% test accuracy. Progressive threshold relaxation (y_1 through

y_6) shows expected performance degradation, yet even the most permissive label set maintains 97.14% accuracy.

Dimensionality Reduction Impact: Principal Component Analysis preprocessing also improved generalization performance, reducing the 19-dimensional feature space to 6 components while also capturing ¿85% of total variance consistently improved test accuracy, particularly for challenging inclusive definitions (y_4 to y_6). The consistent selection of 6 components indicates stable underlying structure in reviewer behavioral data.

Validation: The consistently high performance ($> 97\%$) across all label sets validates our heuristic labeling strategy. KNN's effectiveness using only behavioral metadata suggests our rule-based approach captures genuine manipulation patterns rather than noise, providing confidence for subsequent pipeline stages operating on cleaned data with reduced manipulation artifacts.

4.2 Emotion-based Sentiment Analysis

We trained our transformer-based models using the Emotion Dataset from Kaggle. This dataset includes emotion categories: *joy*, *sadness*, *anger*, *fear*, *surprise*, and *love* [10]. All transformer-based models achieve a great generalization performance, as shown in Table 2. Next, we used our fine-tuned transformer-based models to classify the emotion labels of Black-Myth: WuKong's reviews.

Based on Figures 2 and 3, all transformer-based models demonstrate high confidence and comparable performance in classifying reviews of Black Myth: WuKong. Although each individual model maintains a high confidence level, the ensemble model exhibits a lower confidence, likely due to disagreement among the models, as illustrated in Figures 4 and 5. In particular, there is a notable disagreement between joy and anger classifications across all models (Figure 4), which can be attributed to ambiguous player reviews containing both aggressive and positive language. As shown in Figure 5, the models agree on the same emotion label for 60.1% of the reviews, while they completely disagree in 3.6% of the cases. Additionally, GPT-2 agrees with RoBERTa and DeBERTa at rates of 73.1% and 71.1%, respectively, and RoBERTa agrees with DeBERTa at 72.3%.

Model	**Training Accuracy**	**Testing Accuracy**
DeBERTa-v3 (355M)	90.69 %	93.65%
RoBERTa Base (125M)	90.13%	93.35%
GPT-2 (355M)	90.22%	93.68%

Table 2: Accuracy after fine-tuning and testing accuracy for different transformer models.

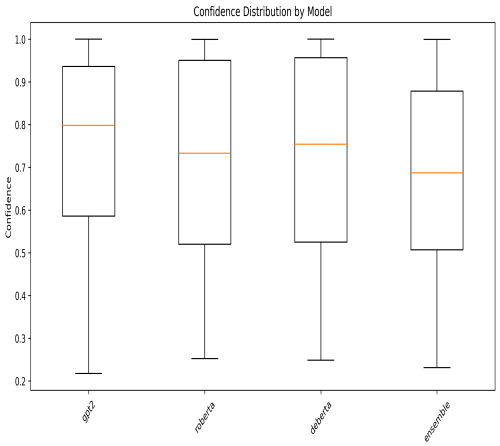


Figure 2: Model Confidence Boxplot

Figure 3: Emotion Distributions

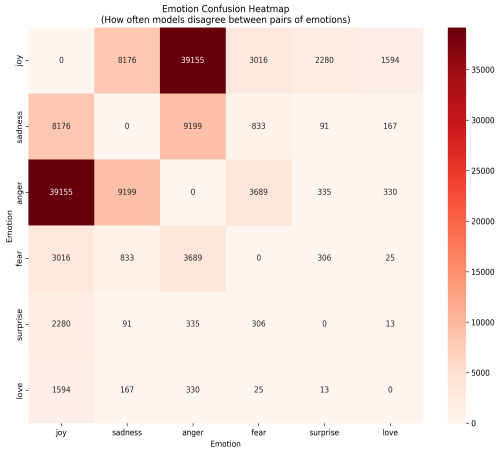


Figure 4: Confusion Matrix of Model Disagreement

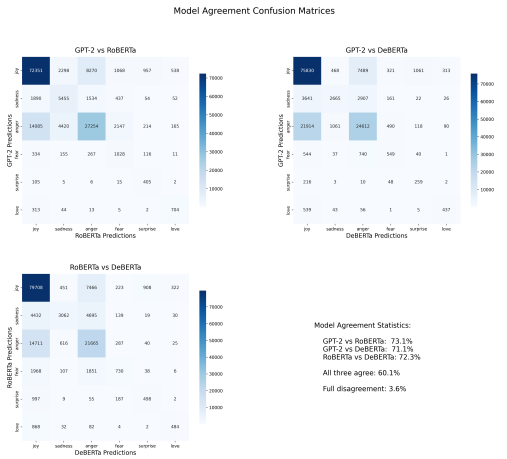


Figure 5: Model Comparison

4.2.1 Aspect-Based Sentiment Analysis Results

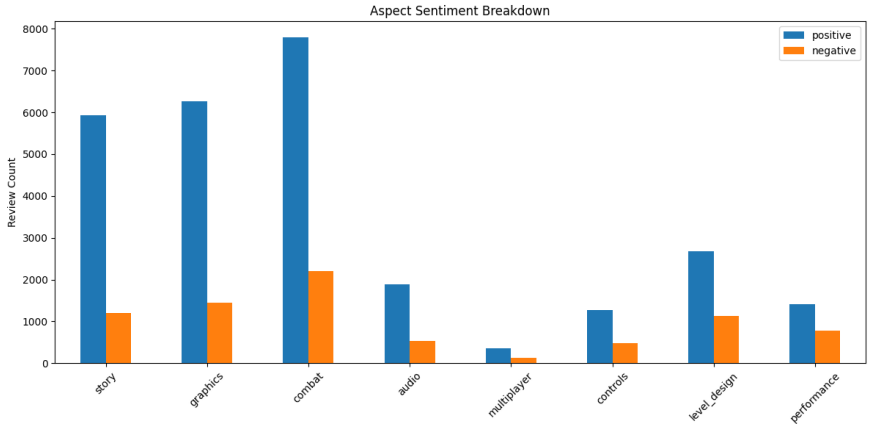


Figure 6: Aspect-level sentiment distribution for BlackMyth: WuKong

This dual-method approach demonstrates complementary strengths: the guided NMF method discovers emergent aspects and nuanced themes within gaming discourse, while the keyword-based pipeline provides reliable, interpretable tracking of established gaming aspects. Together, these methods enable comprehensive aspect-based analysis suitable for both exploratory research and targeted monitoring applications. Both approaches demonstrated significant evidence that people had very positive reviews for many different aspects of BlackMyth: WuKong including gameplay, graphics, etc. as shown in Figure 6.

5 Conclusion

This project set out to explore the process in creating a data pipeline, and also to build a tool that will provide value to the user. Taking in an input of a video game ID, our pipeline first scrapes Steam library for reviews on the inputted video game, and then cleans it using anomaly detection, KNN, and PCA to determine the credibility of reviews. The pipeline then takes the cleaned data in and begins an ensemble of emotion-based sentiment analysis and aspect-based sentiment analysis. Our specific report focused on the massively popular BlackMyth: WuKong game as an example. For the emotion-based sentiment analysis, we found it interesting that the models differed quite significantly in sentiment analysis, and that joy and anger were often confused with each other. For video game reviews, emotions is likely very tricky to distinguish as people can be upset at the boss, but still they love the game. For aspect-based sentiment analysis, the game had extremely positive sentiment overall and both models tended to agree.

6 Acknowledgments

Our group would like to thank Medium author Fabio Chiusano, ML researcher Sebastian Raschka, FacebookAI, Microsoft, and HuggingFace for invaluable resources in writing the code for emotion-based and aspect-based sentiment analysis.

7 Author Contributions

- Jun and Steve worked on web scrapping and the ensemble of emotion-based sentiment analysis.
- Dylan worked on cleaning the data and developing the credibility assessment of reviews.
- Blair and Mehwish worked on creating the ensemble of aspect-based sentiment analysis.
- Everyone contributed to the final report.

8 Verification Statement

Overall, our pipeline currently takes a video game ID found on Steam as the input for scrapping Steam, however the pipeline would have greater utility if the user could just input the game name. This would have been ideal if our group had more time. Also a part of the pipeline, it would've been very helpful to combine all the code and have a concise output for the user at the end. This would have taken a lot more time in making sure the code works together and would take a lot of time for all of the code to run at once.

For emotion-based sentiment analysis, our group would like to dive deeper into the results of the emotion-based sentiment analysis and why the models don't necessarily agree with each other. Looking into why reviews were confused between joy and anger would also provide insight into improving the model.

For aspect-based sentiment analysis, something that our group would have explored with more time is to improve topic modeling integration and fine-tuning the pre-built sentiment analysis models. Currently, our guided NMF just takes the top 5 terms in each aspect category and feeds those into the aspect-based sentiment analysis model. Being able to reduce words down to their root definitions, speeding up efficiency by not checking every review for every term, and better synonym detection are things that we would have explored with more time. Using pre-built models also greatly limits their effectiveness for video game specific reviews, and with more time we would have liked to explore more about developing our own model.

Finally, while we were unable to build a full pipeline, we have enough pieces of the puzzle to ensure that the pipeline is almost at its full completion. We believe that if given more time, we will be able to achieve a full functioning pipeline.

GitHub Repo: https://github.com/Steam-Games-Insight-Engine/Steam-Games-Insight-Engine

References

[1] Osamah Alsharif, Dalal Alshamaa, and Nadim Ghneim. Emotion classification in arabic text using deep learning. *Computers and Electrical Engineering*, 101:108038, 2022.

[2] Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France, June 2022. European Language Resources Association.

[3] Giovanni Gaetano Cantone, Giovanni Improta, Maria Triassi, and Stefano Castiglione. Review bombing: ideology-driven polarisation in online ratings: The case study of the last of us (part II). *Quality & Quantity*, 58(4):3569–3588, 2024.

[4] Zixuan Chen, Yang Liu, and Wei Zhang. Multi-task learning for game review classification with emotion and sarcasm detection. In *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 1095–1102. IEEE, 2023.

[5] nateprewitt graffatcolmingov, Lukasa. requests.

[6] Petr Hajek, Lyudmila Hikkerova, and Jean-Michel Sahut. Fake review detection in e-commerce platforms using aspect-based sentiment analysis. *Journal of Business Research*, 150:534–547, 2023.

[7] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*, 2021.

[8] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[9] HuggingFace. distilbert-base-uncased-finetuned-sst-2-english. `https://huggingface.co/distilbert-base-uncased-finetuned-sst-2-english`, 2020. Accessed: 2025-05-30.

[10] Bhavik Jikadara. Emotions dataset.

[11] leonard. Beautifulsoup.

[12] Dayi Lin, Cor-Paul Bezemer, and Ahmed E. Hassan. An empirical study of game reviews on the steam platform. *Empirical Software Engineering*, 24(1):170–207, 2019.

[13] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE, 2008.

[14] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[15] Michal.Danilk. langdetect.

[16] Simone Petrosino, Elia Loria, Alexander Kainz, and Johanna Pirker. The panorama of steam multiplayer games (2018-2020): A player reviews analysis. In *Proceedings of the 17th International Conference on the Foundations of Digital Games*, FDG '22. ACM, 2022.

[17] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[18] Statista. Number of games released on steam worldwide from 2004 to 2023, 2024. Accessed: 2025-05-30.

[19] Valve Corporation. Steam 2021 year in review. `https://steamcommunity.com/groups/steamworks/announcements/detail/3133946090937137590`, 2021. Accessed: 2025-05-30.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5998–6008, 2017.

[21] Hao Wang, Yan Lu, and Chengqing Zhang. A systematic review of aspect-based sentiment analysis: domains, methods, and trends. *Artificial Intelligence Review*, 57(4):1–42, 2024.

[22] Heng Yang, Biqing Zeng, Mayi Xu, and Tianxing Wang. Back to reality: Leveraging pattern-driven modeling to enable affordable sentiment dependency learning. *CoRR*, abs/2110.08604, 2021.

[23] Berna Yilmaz. Uncovering patterns and simplifying complexity: A guide to nmf. *Medium*, 2024.

[24] Akram Yousif and Jim Buckley. Impact of sentiment analysis in fake review detection. *arXiv preprint arXiv:2212.08995*, 2022.

[25] Yizhou Yu, Duc Thang Dinh, Binh Huy Nguyen, Faxing Yu, and Van-Nam Huynh. Mining insights from esports game reviews with an aspect-based sentiment analysis framework. *IEEE Access*, 11:54653–54669, 2023.