Unveiling Steam Reviews Discrepancies: A Scientific Analysis through Text Mining on Rating Scores and Recommendations

Pei Cheng Wang
pcw3@illinois.edu
University of Illinois,
Urbana-Champaign
Champaign, Illinois, USA

Yifan Zhong
yzhong32@illinois.edu
University of Illinois,
Urbana-Champaign
Champaign, Illinois, USA

Dayou Wu
dayouwu2@illinois.edu
University of Illinois,
Urbana-Champaign
Champaign, Illinois, USA

Jiajun Zhang
jiajunz6@illinois.edu
University of Illinois,
Urbana-Champaign
Champaign, Illinois, USA

Zhaoyang Zhu
zzhu62@illinois.edu
University of Illinois,
Urbana-Champaign
Champaign, Illinois, USA

ABSTRACT

To evaluate reviews of games on Steam based on the recommendation index and determine whether the comments are positive, a text mining approach is used. This methodology aims to develop a model capable of differentiating genuine recommendations from other comment types.

1 INTRODUCTION

As the largest digital distribution platform for PC gaming, Steam boasts an immensely large user base. For players in search of new games, Steam reviews serve as the most direct and significant influence on their decision-making. However, the majority of comments offer little value. The comment sections for numerous games are regularly inundated with irrelevant spam and remarks unrelated to the games themselves[7]. Moreover, a significant portion of comments may appear commendatory at first glance but are actually intended to be derogatory.

Building on this, the challenge of assessing genuine sentiment from user comments and making accurate content recommendations to users emerges as an intriguing subject. Prior studies on video game recommendation systems have shown that user sentiment data have a minimal or negligible effect on the performance of these systems.[5]. Our goal is to compare new LLM models' performance against the best-performing DeepNN models established in previous experiments as a baseline[12], to further assess whether our new model can outperform previous models in sentimental analysis and recommendations.

Sentiment analysis is a well-established field within Natural Language Processing (NLP), characterized by the application of various methodologies[11], including CNNs, LSTMs[1], as well as techniques based on Naive Bayes models, Linear Support Vector Machines, and Decision Trees[13]. These approaches sometimes miss the nuanced semantic elements within the text. The introduction of large models presents an opportunity to attain a more nuanced understanding of text semantics, helping us dive into the depth and critical information of sentiments. This could lead to more precise game content summaries and targeted recommendations for players with specific interests.

2 DESCRIPTION

evaluation.py is used to evaluate the performance of the sentiment analysis model. It loads the ground-truth.csv file containing the true labels and the sentiment-result.csv file containing the model's predicted results. Then, it preprocesses both datasets, filtering out empty labels and empty text. Next, it calculates the precision, recall, and F1 score of the predicted results, and prints out these evaluation metrics.

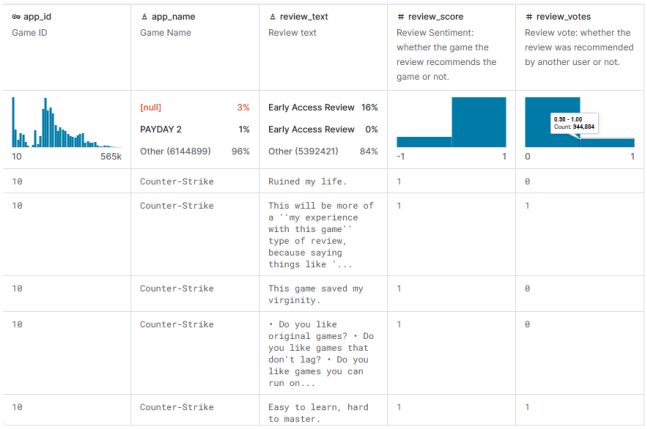
preprocessing.py is used to preprocess the raw dataset. It loads the dataset.csv file containing the original review texts and performs a series of processing steps on each review text: removing punctuation, converting text to lowercase, and removing stop words. Then, it saves the processed review texts to the review-text.json file for later use by the sentiment analysis model.

sentiment-marker.py uses a pre-trained zero-shot classification model to perform sentiment analysis on the review texts. It loads the review-text.json file containing the review texts and uses the zero-shot classification model to label each review with either positive or negative sentiment. Then, it saves the sentiment labels along with corresponding confidence scores to the sentiment-result.csv file for further evaluation.

3 DATA

This project utilize the "Steam Reviews" dataset accessible on Kaggle[9], which comprises more than 6.4 million public reviews in English sourced from the Steam Reviews section of the Steam store operated by Valve. This dataset characterizes each review by its text, the game ID it's associated with, the review's sentiment (either positive or negative), and the count of users who found the review useful. The dataset is stored in a `.csv` file format. Each record in the dataset represents a single review and includes the following fields:

- **game_id**: A unique identifier for the game being reviewed.
- **game_name**: The name of the game.
- **reviews_text**: The text of the review, which is used for sentiment analysis.
- **review_score**: Review Sentiment, indicating whether the review recommends the game (positive or negative).



app_id Game ID	app_name Game Name	review_text Review text	# review_score Review Sentiment: whether the game the review recommends the game or not.	# review_votes Review vote: whether the review was recommended by another user or not.
10 565k	[null] 3% PAYDAY 2 1% Other (6144899) 96%	Early Access Review 16% Early Access Review 0% Other (5392421) 84%	-1 1	0 1 0.00 - 1.00 Count: 344,894
10	Counter-Strike	Ruined my life.	1	0
10	Counter-Strike	This will be more of a ''my experience with this game'' type of review, because saying things like '...	1	1
10	Counter-Strike	This game saved my virginity.	1	0
10	Counter-Strike	• Do you like original games? • Do you like games that don't lag? • Do you like games you can run on...	1	0
10	Counter-Strike	Easy to learn, hard to master.	1	1

Figure 1: Dataset Overview and Samples

- **reviews_votes**: A count of how many other users found the review useful, providing insight into the impact and relevance of a review within the community.

This dataset allows for a detailed analysis of consumer sentiments and the effectiveness of reviews in influencing community engagement and preferences.

4 DATA PREPROSSESSING

The dataset employed in this study exhibited a high degree of integrity and cleanliness. This circumstance obviated the necessity for extensive data cleansing efforts, often requisite in similar studies to rectify issues such as missing values, duplications, or anomalous entries. Accordingly, this allowed preprocessing endeavors to be directed with greater specificity toward textual normalization procedures.

Data Cleaning and Preprossessing involves several steps to prepare the dataset for analysis. Stopwords, frequently occurring yet semantically sparse, were excised to mitigate their disproportionate influence on frequency-based analyses. Case normalization was executed by converting all textual data to lowercase, ensuring lexical consistency and precluding discrepancies arising from case variations. Furthermore, punctuation and numerical characters were expurgated, given their irrelevance to the sentiment analysis algorithm's requirements. The elimination of superfluous whitespace was the final step in the preprocessing sequence, serving to standardize data formatting and optimize computational efficiency. These preprocessing steps are crucial for cleaning the data and ensuring the accuracy of subsequent analyses.

Though the dataset is clean and organized, it lacks ground truth for review sentiments. Our project, aimed at providing sentiment analysis, required ground truth labels for training and testing the model. Therefore, we manually labeled reviews to ensure the robustness and effectiveness of the model. Reviews from different games were labeled to maintain diversity in the training set. Specifically, reviews from 20 games were labeled, with 30 labels per game, encompassing a wide range of sentiments to enhance model accuracy and generalizability.

These hand-labeled reviews form the core vocabulary for our project. To further refine the model's accuracy, we conducted fine-tuning based on this curated data. The dataset was strategically split into training and testing sets, with 450 reviews used for training and 150 for testing. This distribution ensures that the model is robustly trained on a substantial dataset and rigorously tested on a separate subset, effectively preventing overfitting and promoting generalization to new, unseen data.

5 MODEL

Transformer-based language models like GPT[8] and BERT[6] have revolutionized the field of natural language processing, surpassing earlier linear models and neural network approaches that relied on LSTMs or CNNs. These models have demonstrated superior performance in a variety of tasks, notably in sentence and text classification[10].

The computational model adopted for sentiment analysis in this project is the 'Twitter-roBERTa-base for Sentiment Analysis[3].' This model represents a domain-specific instantiation of the RoBERTa-base framework, having been subjected to rigorous training on a corpus of approximately 124 million tweets, curated over the span from January 2018 to December 2021 [3]. The model's fine-tuning was conducted utilizing the TweetEval benchmark [2], which is recognized for its robustness in evaluating natural language understanding and sentiment analysis models.

The 'twitter-roBERTa-base for Sentiment Analysis' model has been specifically optimized for the sentiment analysis task, which, within this context, is delineated into a tripartite classification system: positive, neutral, and negative sentiments. The underpinning for the model's linguistic proficiency in English is derived from its training on the unified TweetEval adaptation of the Semeval-2017 dataset [2], which is a seminal benchmark for sentiment analysis within Twitter.

The performance of the 'Twitter-roBERTa-base for Sentiment Analysis' model on the TweetEval leaderboard demonstrates its effectiveness in sentiment analysis compared to traditional models. With an accuracy of 66.7% in the sentiment category, it notably surpasses traditional machine learning models like LSTM and SVM, which achieved accuracies of 58.3% and 62.9% respectively[2]. This indicates that the advanced capabilities of transformer-based models, such as roBERTa, provide a significant improvement in understanding the nuances of language used in social media, compared to older techniques based on LSTM or SVM which lack the same depth of contextual understanding and adaptability.

6 TASK

Sentiment Analysis. The goal of the sentiment analysis task is to accurately classify each review in the dataset as positive, negative, or neutral. For this purpose, we utilize the "Steam Reviews" dataset sourced from Kaggle[9], which includes a diverse array of game reviews. The initial dataset is already preprocessed for basic textual normalization, ensuring the removal of irrelevant characters and standardizing the text format.

Fine-tuning. To further enhance the model's performance, we split the dataset into training and testing sets, employing fine-tuning techniques on the training data. This approach allows the

Model	Emoji	Emotion	Hate	Irony	Offensive	Sentiment	Stance	ALL(TE)
BERTweet	33.4	79.3	**56.4**	**82.1**	79.5	73.4	71.2	**67.9**
TimeLMs-2021	**34.0**	**80.2**	55.1	64.5	**82.2**	**73.7**	**72.9**	66.2
RoBERTa-Retrained	31.4	78.5	52.3	61.7	80.5	72.8	69.3	65.2
RoBERTa-Base	30.9	76.1	46.6	59.7	79.5	71.3	68	61.3
RoBERTa-Twitter	29.3	72.0	49.9	65.4	77.1	69.1	66.7	61.4
FastText	25.8	65.2	50.6	63.1	73.4	62.9	65.4	58.1
LSTM	24.7	66.0	52.6	62.8	71.7	58.3	59.4	56.5
SVM	29.3	64.7	36.7	61.7	52.3	62.9	67.3	53.5

Figure 2: TweetEval: Leaderboard[2]

model to learn from a broad spectrum of sentiments expressed in the reviews, thus improving its ability to generalize and accurately classify new, unseen data. The fine-tuning process involves adjusting model parameters, re-evaluating the training strategy, and optimizing the learning rate to better adapt to the nuances of sentiment expressed in gaming reviews.

7 EVALUATION

7.1 Model Performance

To assess the efficacy of the 'witter-roBERTa-base' model in sentiment analysis on the Steam review dataset, we compared the output of the model to our hand label results to provide a ground truth against which the model's predictions were compared.

Table 1: Model Performance Metrics

Metric	Value
Precision	0.98
Recall	0.80
F1 Score	0.88

Our model achieves a Precision of 0.98, Recall of 0.80, and F1 score of 0.88 on the test set.

7.2 Analysis

In the initial phase, our model demonstrated significant effectiveness in identifying inherently positive sentiments within Steam reviews, providing reasonable predictive outcomes. However, it encountered substantial challenges when dealing with negative sentiments. We observed that in numerous instances, reviews employed negative language to convey positive sentiment, exemplified by comments like "This game ruined my life," which, contrary to its negative facade, actually praises the game for being engaging and time-consuming.

This paradoxical use of negative language to express positive feedback represents a nuanced challenge that large language models often struggle with. The inherent complexity of detecting the actual sentiment behind such statements indicates a potential area for further model refinement.

To address this, we fine-tune the model to enhance its capability to interpret the true semantic content of sentiments expressed in reviews, particularly those that use negative language to convey positive meanings. Since these types of comments are not prevalent, we plan to employ the SMOTE oversampling technique[4] to augment our dataset, thereby providing a more balanced foundation for continuous fine-tuning of the model. This approach aims to improve the model's accuracy in classifying sentiments, especially in distinguishing between genuinely negative sentiments and positive sentiments expressed through negative language. To improve the robustness of our model, we will employ cross-validation strategies for sample selection.

7.3 Fine-tuning

By enhancing the model's ability to discern the intricate nuances of sentiment expression in user reviews, we aim to achieve a more accurate and nuanced understanding of user feedback, which is essential for developing a comprehensive sentiment analysis tool.

For the fine-tuning process, we selected 20 games and extracted 30 reviews for each game from our dataset. This ensured a diverse and representative sample for our sentiment analysis model. Recognizing the challenge of ambiguous sentiments in the dataset, 25% of the reviews were specifically chosen to address cases of "ambiguity". These reviews were manually labeled to serve as Ground Truth, providing a reliable basis for model training.

The fine-tuning of the Twitter-roBERTa-base model was specifically aimed at enhancing its ability to discern such nuanced expressions of sentiment. This step involved adjusting model parameters, retraining the model on these specially selected subsets of data, and iteratively testing the model's performance to ensure significant improvements in handling ambiguous cases.

Table 2: Model Performance Metrics After Fine-tuning

Metric	Value
Precision	0.93
Recall	0.84
F1 Score	0.883

Our model achieves a Precision of 0.93, Recall of 0.84, and F1 score of 0.883 on the test set.

8 DISCUSSION

Our results indicate that while the precision of the model decreased after fine-tuning on ambiguous data, the overall recall improved, leading to a higher F1-score. This outcome suggests that the model became better at identifying relevant cases (higher recall) but at the cost of incorrectly labeling more cases as positive (lower precision).

Possible reasons for this shift include:

- **Low Precision:** The presence of ambiguous data appears to have had a detrimental effect on the model's precision. Ambiguous comments, which often blend positive and negative sentiments, likely caused the model to misclassify cases that it might have previously judged correctly. This led to

an increased rate of false positives, where non-positive sentiments were incorrectly classified as positive, thus reducing precision.

- **Higher Recall:** The introduction of ambiguous data, especially negative comments labeled as positive, may have conditioned the model to classify a broader array of inputs as positive. This adjustment in the model's behavior increased the true positive rate, thereby enhancing recall. As the model marked more data as positive, it became less likely to miss actual positive cases, even though this came at the expense of precision.

These findings underscore the challenges and trade-offs involved in fine-tuning sentiment analysis models on ambiguous data. While such adjustments can enhance the model's ability to detect nuanced expressions of sentiment (as reflected by the increased recall and F1-score), they also highlight the potential for decreased precision, which can lead to less accurate classifications. Understanding these dynamics is crucial for further refining the model to balance precision and recall effectively.

9 CONCLUSION & FUTURE WORK

This study investigated the application of a machine learning-based model to perform sentiment analysis on a dataset replete with ambiguous reviews. This allowed us to balance sentiment representation and significantly enhance the accuracy of our model in distinguishing genuine negative sentiments from ironically positive expressions. Like "This game ruined my life." etc.. Additionally, we fine-tune our model through rigorous cross-validation methods to ensure its reliability and robustness in analyzing diverse user reviews.

For future work, enhancing our sentiment analysis model could greatly benefit from incorporating advanced contextual embeddings like those used in BERT/GPT models. These embeddings excel at interpreting complex language patterns, which are crucial for accurately understanding nuanced sentiments. Expanding the model to support multiple languages would not only improve its accuracy but also its global applicability, given the diverse ways sentiments are expressed across different cultures.

Adopting a continuous learning framework would allow the model to dynamically update and adapt to new data and evolving language trends. Moreover, integrating hybrid approaches that combine rule-based methods with machine learning could lead to finer sentiment detection, especially for sentiments that are subtly expressed or contain contradictions.

Furthermore, incorporating user feedback directly into the training process could enhance the model's accuracy and adaptability, ensuring it remains relevant to user needs. These improvements aim to develop a sophisticated and precise tool that would enable game developers and marketers to more effectively understand and respond to user sentiments.

REFERENCES

[1] Rohan Bais, Pasal Odek, and Seyla Ou. 2017. Sentiment Classification on Steam Reviews.

[2] Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa-Anke, and Leonardo Neves. 2020. TweetEval:Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Proceedings of Findings of EMNLP.*

[3] CardiffNLP. 2023. twitter-roBERTa-base-sentiment-latest. https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest/commits/main. Accessed: 2024-03-29.

[4] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[5] Germán Cheuque, José Guzmán, and Denis Parra. 2019. Recommender systems for Online video game platforms: The case of STEAM. In *Companion Proceedings of The 2019 World Wide Web Conference*. 763–771.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186.

[7] Lukas Eberhard, Patrick Kasper, Philipp Koncar, and Christian Gütl. 2018. Investigating helpfulness of video game reviews on the steam platform. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. IEEE, 43–50.

[8] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. *Improving Language Understanding with Unsupervised Learning.* Technical Report. OpenAI.

[9] Antoni Sobkowicz. 2017. *Steam Review Dataset.* https://doi.org/10.5281/zenodo.1000885

[10] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A Multi-task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of ICLR.*

[11] Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review* 55, 7 (2022), 5731–5780.

[12] Huang Yi, Sun Shiyu, Duan Xiusheng, and Chen Zhigang. 2016. A study on Deep Neural Networks framework. In *2016 IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*. 1519–1522. https://doi.org/10.1109/IMCEC.2016.7867471

[13] Zhen Zuo. 2018. Sentiment analysis of steam review datasets using naive bayes and decision tree classifier. (2018).