Exploring Gamer Sentiments on Steam using NLP

Christian Anthony C. Stewart
Department of Computer, Information Sciences and Mathematics
University of San Carlos
Cebu City, Philippines
christianstewart5111@gmail.com

***Abstract*—Emotion theory proposes that human beings have a limited number of emotions [1]. This paper aims to uncover the underlying sentiments expressed by gamers through the application of Natural Language Processing (NLP) techniques. By examining the intricate language patterns present in user reviews, the objective is to utilize TF-IDF (Term Frequency-Inverse Document Frequency) values to construct a predictive model. This model is designed to predict whether a game will be recommended by a player based on the textual content of the review. Moreover, the study utilizes TextBlob for sentiment analysis to explore potential correlations between the sentiment polarity of reviews and the likelihood of game recommendation. Leveraging a comprehensive training dataset, the model will undergo rigorous training to capture the subtleties of gamer feedback. Subsequently, the model's performance will be evaluated using a separate dataset to ensure its efficacy in generating predictions for new and unseen reviews. Embark on this analytical journey to uncover the motivations behind gamers' recommendations—or lack thereof—for their preferred games.**

***Index Terms*—Natural Language Processing, Sentiment Analysis**

I. Introduction

Understanding the sentiments of gamers is crucial in deciphering their preferences and behaviors within the gaming community. Emotion theory suggests that human beings experience a limited range of emotions, which serve as fundamental drivers behind their actions and interactions. Within the gaming landscape, these emotions play a significant role in shaping user experiences and influencing gaming choices. By delving into the realm of Natural Language Processing (NLP), researchers can unlock valuable insights hidden within user reviews and discussions on gaming platforms like Steam.

The ability to understand human emotions from written text, out of the various modalities, has been a significant challenge in the field of natural language processing. Despite such predicament, sentiment analysis is critical in capturing the nuances of human emotion especially in text and researchers rely on machine learning to develop models to tackle such [2]. In the field of NLP, the challenge lies in extracting and analyzing sentiments from the vast trove of user-generated content on gaming platforms.

By harnessing the power of machine learning algorithms and linguistic analysis, researchers can gain deeper insights into the motivations and preferences driving gamers' recommendations and critiques. Through this exploration of gamer sentiments on Steam using NLP analysis, the study aims to unveil the underlying emotions shaping the gaming community and pave the way for enhanced user experiences and game development strategies.

The core of the study is aimed to develop a predictive sentiment model for analyzing user reviews of games on the Steam platform.

Specifically, the objectives were the following:

1) Preprocess a Steam Reviews Dataset [3].
2) Develop a predictive model for sentiment analysis.
3) Evaluate the performance of the model.
4) Provide a sanity check of predicted results.

II. Initial Attempt

A. A Peek at the Data

The training dataset contains a game title column, a year column, a user review column that contains the textual gamer feedback and a user suggestion (recommendation) column. User suggestions are marked as "1" for recommended and "0" otherwise in the "user suggestion" column.

TABLE I
First 5 Rows of Steam Game Review Dataset

	title	developer	publisher	tags	overview
0	Spooky's Jump Scare Mansion	Lag Studios	Lag Studios	['Horror', 'Free to Play', 'Cute', 'First-Pers...	Can you survive 1000 rooms of cute terror? Or ...
1	Sakura Clicker	Winged Cloud	Winged Cloud	['Nudity', 'Anime', 'Free to Play', 'Mature', ...	The latest entry in the Sakura series is more ...
2	WARMODE	WARTEAM	WARTEAM	['Early Access', 'Free to Play', 'FPS', 'Multi...	Free to play shooter about the confrontation o...
3	Fractured Space	Edge Case Games Ltd.	Edge Case Games Ltd.	['Space', 'Multiplayer', 'Free to Play', 'PvP'...	Take the helm of a gigantic capital ship and g...
4	Counter-Strike: Global Offensive	Valve, Hidden Path Entertainment	Valve	['FPS', 'Multiplayer', 'Shooter', 'Action', 'T...	Counter-Strike: Global Offensive (CS: GO) expa...

We strip the unnecessary information from the dataset as part of the preprocessing and it would look like how it is in Table. II.

TABLE II
Data Instance from Steam Reviews Dataset

ID	Title	Year	User Review	Suggestion
1	Spooky's Jump...	2016.0	I'm scared and ...	1
2	Spooky's Jump...	2016.0	Best game, better...	1
3	Spooky's Jump...	2016.0	A littly iffy on the...	1
4	Spooky's Jump...	2015.0	Great game, fun and...	1
5	Spooky's Jump...	2015.0	Not many games have...	1

Here, we group the reviews by title and investigate the difference in recommendation numbers (recommended or not) to get an intuitive feel for the most and least popular

games in the dataset. The resulting percentage is similar to a common gaming measure of how many gamers "like the game", as on Google, etc.

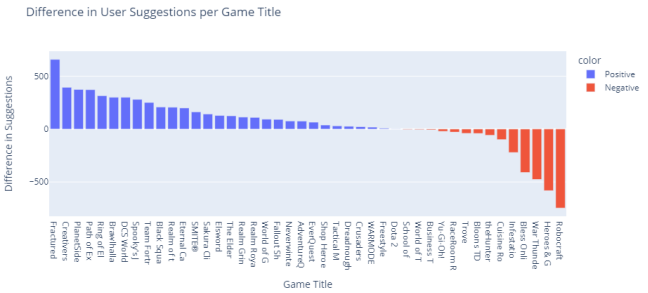

Fig. 1. Difference in User Suggestions per Game Title

From the Fig. 1, we can obtain the 5 Best and Worst titles based on user suggestions. This is visible in Table III and Table IV. In the tables, user suggestions are marked as "1" for recommended and "0" otherwise in the "user suggestion" column.

TABLE III
5 BEST TITLES

Title	0	1	Difference	Percentage	color
Fractured Space	30	688	658	95.82	Positive
Creativerse	49	443	394	90.04	Positive
PlanetSide 2	49	423	374	89.62	Positive
Path of Exile	43	415	372	90.61	Positive
Ring of Elysium	52	367	315	87.59	Positive

From the game overview dataset, there are about 64 unique titles which is a quite limited number but it is a good start for this term paper.

TABLE IV
5 WORST TITLES

Title	0	1	Difference	Percentage	color
Infestation	350	129	-221	26.93	Negative
Bless Online	561	151	-410	21.21	Negative
War Thunder	598	122	-476	16.94	Negative
Heroes Generals	663	82	-581	11.01	Negative
Robocraft	794	48	-746	5.70	Negative

A good way to visualize the most common words in a corpus of text is to generate a word cloud. A word cloud is a visual representation of text data where the size of each word indicates its frequency or importance in the document. Fig. 3 is a word cloud that is generated from words across all reviews which displays the top 10 most frequent words, and then generates and plots a word cloud with the 100 most common words set against a white background.

B. Binary Classifier Model: TD-IDF and Logistic Regression

A Baseline Binary Classifier Model using TF-IDF (Term Frequency-Inverse Document Frequency) and Logistic Regression is a foundational approach in natural language processing (NLP) for text classification tasks [4].

Fig. 2. Difference in User Suggestions per Game Title

In the context of sentiment analysis or text classification, the Bmodel serves as a benchmark or starting point for more sophisticated models. It involves preprocessing the text data, such as tokenization and removing stop words, followed by transforming the text into numerical vectors using TF-IDF. Then, Logistic Regression is applied to these TF-IDF vectors to train a model that can predict the sentiment or classify the text into predefined categories.

While simple, this baseline model provides a straightforward yet effective approach for text classification tasks. It is easy to implement and interpret, making it a suitable choice for initial exploratory analysis or as a reference point for evaluating more complex models.

C. Result

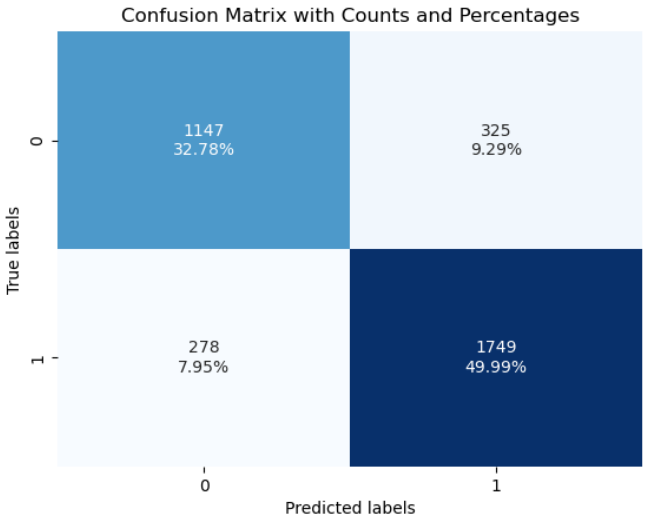

Fig. 3. Confusion Matrix of Binary Classifier Model

The baseline model produces an accuracy score of 82.7% in predicting recommendations based on the textual content of user reviews. The Fig. 2 shows a visualization of the predicted labels for the test set and Table.V is the breakdown of the evaluation metrics.

TABLE V
METRIC RESULTS

	Precision	Recall	F1-score
not recommended	0.80	0.78	0.79
recommended	0.84	0.86	0.85
Accuracy	0.83	0.83	0.83
Macro Avg	0.82	0.82	0.82
Weighted Avg	0.83	0.83	0.83

III. OPTIMIZING THE MODEL

A. GridSearchCV

The baseline model produces an accuracy score of 82.7% in predicting recommendations based on the textual content of user reviews. This is already a well-performing model but higher accuracy can be achieved by fine-tuning the parameters of the TF-IDF model.

GridSearchCV is a technique used for hyperparameter tuning in machine learning models. It is part of the scikit-learn library in Python and is used to systematically search for the optimal combination of hyperparameters for a given estimator (machine learning model) using cross-validation [5].

You define a set of hyperparameters and their possible values, and GridSearchCV will exhaustively search through all possible combinations of these hyperparameters. For each combination, it evaluates the performance of the model using cross-validation, typically with k-fold cross-validation.

B. Result

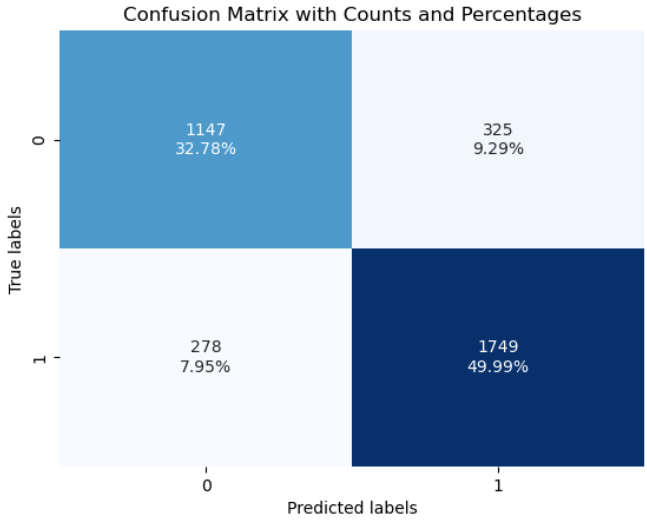

Fig. 4. Confusion Matrix of the Optimized Model

Having established that the best-performing model in our grid search is the model with the parameters defined above, the data can be refit with this optimized model. The Fig. 4 shows a visualization of the predicted labels for the test set. The improved performance model obtained an accuracy of 85%. This is a high rate of success in predicting recommendations based solely on NLP methods. Table.VI is the breakdown of the evaluation metrics.

TABLE VI
METRIC RESULTS

	Precision	Recall	F1-score
not recommended	0.84	0.80	0.82
recommended	0.86	0.89	0.87
Accuracy	0.85	0.85	0.85
Macro Avg	0.85	0.84	0.85
Weighted Avg	0.85	0.85	0.85

C. Sentiment Analysis of Reviews

Until now, the analysis of reviews and resulting model have relied upon the frequency of terms across the pool of reviews. However, it is also interesting to look into any possible correlation between the sentiment polarity of a review and a user's willingness to recommend a game. Sentiment polarity is a measure of the emotional tone of a piece of text, indicating whether the sentiment is positive, negative, or neutral.

With the usage of TextBlob, a common NLP library that provides a simple API for common natural language processing (NLP) tasks, such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more. TextBlob is built on top of NLTK (Natural Language Toolkit) and Pattern, two popular NLP libraries in Python [5].

In this case, it is being used for sentiment analysis and applied to the user review.

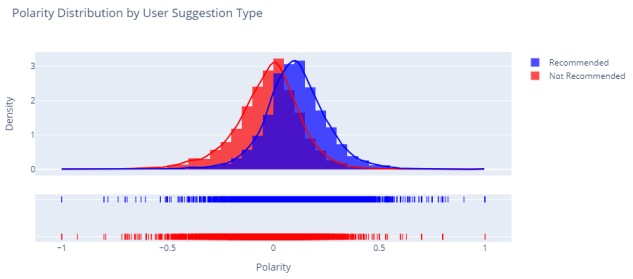

Fig. 5. Polarity Distribution of User Suggestions

Here, a histogram of the sentiment polarity of reviews is plotted with kernel density estimation (KDE) for sets of recommended and not recommended reviews to visualize the correlation between positive sentiment in recommended reviews and neutral/negative sentiment in non-recommended reviews. Sentiment polarity is categorized as follows: negative (< 0), neutral ($= 0$) and positive (> 0).

This histogram in Fig. 5 basically visualizes potential relationship between the sentiment expressed in a review and the likelihood of a game being recommended or not by the reviewer. Finally, it filters and returns the subset of reviews

that have a negative sentiment polarity but still resulted in a user suggestion, which could be considered as misalignments between sentiment and user recommendations.

TABLE VII
POLARITY TABLE OF USER SUGGESTION

Sentiment Polarity	Reviews	Total User Suggestion	Percentage
Positive	7849	9968	78.74%
Negative/Neutral	4162	7526	55.3%

Table VII shows this relationship. Positive sentiment polarity reviews yielding suggestions (similar to an accuracy score) are at 78.74% which means more than half of positive reviews align with recommendations for the game while 55.3% of Negative (also neutral reviews without suggestion so skewed in terms of proportion) polarity reviews that did not yield a suggestion. This means that about half of the negative/neutral reviews align with non-recommendations for the games of the entire dataset.

IV. TEST SET ANALYSIS

In this section, a different Steam user review dataset is used to see how the developed model performs on further unseen data. The dataset is recorded in 'test.csv' and consists of user reviews of completely different games. The NLP-based ML pipeline defined above will then be applied to the dataset to predict user suggestion labels.

Without having any information regarding the actual user suggestion data for this second dataset, an additional sanity check step is taken to establish the legitimacy of the predicted results. This extra step consists of pulling recommendation data from Steam's user review API to establish an intuitive feel for expected recommendation rates for the games in the set.

Fig. 6. Word Cloud of Test and Train Set

A word cloud is generated for the test set (as before) as means of comparison with the training set and for an informative glance at the common terms in this dataset as seen in Fig. 6.

A. Predicting User Suggestion

Using the TD-IDF model established previously for the training set. The transformed data is then passed through the optimised logistic regression model to predict whether or not a gamer would suggest the game to others. (1 or 0, respectively, in the dataset). This can be seen in Table VIII.

TABLE VIII
DATA INSTANCE FROM PREDICTION MODEL DATASET

ID	Title	Year	User Review	Suggestion Prediction
1	CSGO	2015.0	Nice graphics...	0
2	CSGO	2018.0	I would not...	0
3	CSGO	2018.0	I have tried...	0
4	CSGO	2015.0	The game is great...	0
5	CSGO	2015.0	I thank TrulyRazor...	1

B. Sentiment Analysis

As with the training set, sentiment polarity analysis is performed on the test dataset to investigate the possible correlation between emotional tone in the textual content of the user reviews and the potential recommendation of a game as seen in Fig.7.

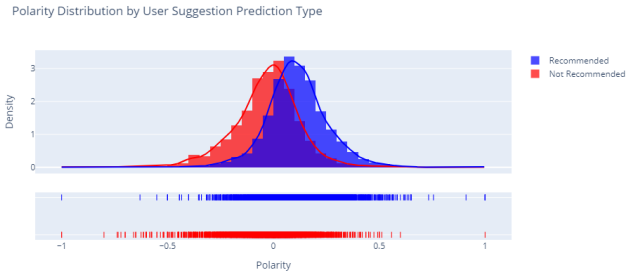

Fig. 7. Polarity Distribution of User Suggestion Prediction

Table IX shows this relationship. Positive sentiment polarity reviews yielding suggestions (similar to an accuracy score) are at 80.52% predicted by the model. This is an improvement over the test dataset while 57.05% of Negative (also neutral reviews without suggestion so skewed in terms of proportion) polarity reviews that did not yield a suggestion. This means that about half of the negative/neutral predicted suggestions align with non-recommendations for the games of the training dataset which is also an improvement from the test dataset.

TABLE IX
POLARITY TABLE OF USER SUGGESTION

Sentiment Polarity	Reviews	Total User Suggestion	Percentage
Positive	3812	4734	80.52%
Negative/Neutral	1889	3311	57.05%

C. Sanity Check of Predicted Results

To ensure the validity of the predicted outcomes without access to the actual user suggestion data for the test dataset, a further verification step is performed by retrieving recommendation data directly from Steam's user review API [6], providing a baseline understanding of anticipated

recommendation levels for the games under consideration.

By extracting the Steam game IDs for the games in the dataset, we then use this information to obtain review data for a specific date range (in this case, relevant to the dataset before December 2018). We then calculate the total number of positive and negative recommendations, and compute the ratio of positive to negative recommendations.The extracted positive suggestion rate can then be compared to the predicted suggestion rate of 58.84% achieved using the model.

The ratio of positive to negative recommendations provides an insight into the proportion of positive recommendations expected for a set of reviews for a given game, much like a sample compared to a population (this being the population and the dataset, a sample). In the standard mean calculation above, the spread of user reviews per game is not taken into account. This could lead to skewed results for a larger number of games that are either very positively or negatively rated.

TABLE X
COMPARISON OF BASELINE USER REVIEW AND PREDICTED RESULTS

	Game Title	Game ID	Up Total	Down Total	Ratio Up/Down	Weighted Ratio Up/Down	Multiplied Ratios
0	Counter-Strike: Global Offensive	730	2344547	309459	0.883399	0.050841	0.044913
1	World of Warships	552990	6809	2417	0.738023	0.033576	0.024780
2	Star Trek Online	9900	7142	2114	0.771608	0.054286	0.041887
3	Paladins®	444090	147387	26313	0.848515	0.055794	0.047342
4	Shadowverse CCG	453480	4979	1708	0.744579	0.041648	0.031010
5	Tree of Savior (English Ver.)	372000	8845	5516	0.615904	0.048614	0.029942
6	VEGA Conflict	339600	1361	1232	0.524875	0.011287	0.005924
7	Minion Masters	489520	5538	1052	0.840364	0.041888	0.035201
8	The Lord of the Rings Online™	212500	6385	1427	0.817332	0.044600	0.036453
9	Fishing Planet	380600	9582	2966	0.763628	0.047270	0.036097
10	Crush Crush	459820	6137	801	0.884549	0.036613	0.032386
11	Dungeon Defenders II	236110	9691	3415	0.739432	0.050919	0.037651
12	Governor of Poker 3	436150	1953	718	0.731187	0.008089	0.005915
13	Digimon Masters Online	537180	2541	1211	0.677239	0.026096	0.017673
14	Shakes and Fidget	438040	5683	879	0.866047	0.014748	0.012773
15	Champions Online	9880	2019	796	0.717229	0.034145	0.024490
16	Magic Duels	316010	1214	204	0.856135	0.095032	0.081360
17	Aura Kingdom	268420	102	29	0.778626	0.031358	0.024416
18	H1Z1	433850	103386	82468	0.556275	0.029387	0.016347
19	GUNS UP!	446150	0	0	NaN	NaN	NaN

To fix this issue, the calculated positive suggestion ratio should be adjusted to include the proportion of games that make up the dataset; the Up/Down ratios for each game should be weighted depending on the number of reviews for that game in the dataset as seen in Table X. These weighted ratios can then be used to calculate a weighted average of expected positive recommendations which better represents the test dataset in question.

The overall, weighted suggestion/recommendation rate for the games in the test set is thus: 58.66%. This value is much more in line with the rate calculated from our predictive model: 58.84%.

V. CONCLUSION

Indeed, the resulting weighted positive suggestion rate is calculated to be 58.66%, which is much closer to the value (58.84%) based on the predictions of the model. Understanding the sentiments of gamers is paramount in deciphering their preferences and behaviors within the gaming community. Emotion theory suggests that human beings experience a limited range of emotions, which serve as fundamental drivers behind their actions and interactions. These emotions play a significant role in shaping user experiences and influencing gaming choices within the dynamic landscape of gaming platforms like Steam.

Despite the challenges inherent in extracting sentiments from written text, sentiment analysis is critical in capturing the nuances of human emotion, especially in text-based interactions. Through the utilization of natural language processing (NLP) techniques, researchers can delve into the vast trove of user-generated content on gaming platforms to unearth valuable insights hidden within user reviews and discussions.

The core objective of the study is to develop a predictive sentiment model for analyzing user reviews of games on the Steam platform. This involves preprocessing the data, developing machine learning models, and evaluating their performance using standard evaluation metrics. The study also aims to establish a correlation between sentiment polarity in reviews and the likelihood of a game being recommended by the user. Clearly, the objectives of the paper was met.

Through the rigorous analysis and interpretation of the findings, the study contributes to the advancement of natural language processing techniques in understanding and predicting user sentiments in the gaming domain. Ultimately, this research serves as a foundation for future studies aimed at enhancing user experiences and informing game development strategies in the ever-evolving landscape of digital gaming.

REFERENCES

[1] C. D. Wilson-Mendenhall, L. F. Barrett, and L. W. Barsalou, “Neural evidence that human emotions share core affective properties,” Psychological Science, vol. 24(6), pp. 947-956, 2013.

[2] M. Z. Ashgar et.al., “A deep neural network model for the detection and classification of emotions from textual content,” Complexity, pp. 1-12, 2022.

[3] Steam Game Review Dataset. (2020, December 24). Kaggle. https://www.kaggle.com/datasets/arashnic/game-review-dataset

[4] Karabiber, F. (n.d.). Binary classification. https://www.learndatasci.com/glossary/binary-classification/

[5] PYPI · The Python Package Index. (n.d.). PyPI. https://pypi.org/

[6] Steam Store. (n.d.). https://store.steampowered.com/