HOGWARTS LEGACY

Post-launch Study

Chen Yang
Guillaume Gerony
Han Mingzhou
Hans Sebastian Mulyawan
Wang Liang Bing

TABLE OF CONTENTS

INTRODUCTION

Overview of the game and studio, as well as the problem statement, data sources, and data preprocessing

1

Overview

"Hogwarts Legacy" is a role-playing action game set in the Wizarding world of Harry Potter

- Published by **Avalanche Studio**
- First release **since acquisition** by Warner Bros.
- Most **anticipated** game of 2023

Avalanche Studios

Problem Statement

Marketing

Assess efficacy of the company’s **marketing campaign**

Quality of Life

Focus post-launch support on fixing **critical bugs**

Sentiments Analysis

Shift in sentiments pre and post-launch

Players’ Experience

Identify **significant challenges** encountered

Data Sources

Hogwarts Legacy	reddit	STEAM
Pre-Launch		
Post-Launch		
Demographics	Fans	Players
Characteristics	Conversational	Issue-specific
Measure	Engagement Level & Complaints	Complaints

Data Sources

Reddit

Offer valuable insights into how fans feel about a game well **before its release** and are typically more **conversational** in nature

reddit

Steam

Provide more **specific information** to the actual **post-launch reactions** from the players and challenges faced in the gameplay

STEAM

< T 1 >

Data Preprocessing

1ST	2ND	3RD	4TH	5TH	6TH	7TH
~~URLs or punctuations~~	**Standardize**	**Lemmatize**	**~~Stop words~~**	**~~None values~~**	**Join**	**Generate Tf-idf features**
Remove URLs and punctuation from the text	Convert all texts to lowercase and tokenize the text	Reduce words to their base or root form	Remove common words that do not add much meaning	Remove any None values from the tokens	Join the tokens together to form a string	*Only for sentiment analysis*

SENTIMENT ANALYSIS

Choice of Model & Performance

2

Methodology

Machine Learning Model

Made use of custom model to predict the sentiments behind each post and comment (positive, negative, neutral).

Labelling of Training/Testing Dataset

Manually labelled ~2,500 posts and comments; selected based on proportion of posts from each month, and the final sentiment of each datapoint is determined based on votes.

Choice of Model

Tuned four different models (XGBoost, Logistic Regression, Random Forest, MLP). Settled for XGBoost since it performed the best.

Performance Comparison

MLP

	precision	recall	f1-score	support
0	0.55	0.52	0.53	118
1	0.71	0.71	0.71	228
2	0.64	0.67	0.65	128
accuracy			0.65	474
macro avg	0.63	0.63	0.63	474
weighted avg	0.65	0.65	0.65	474

XGBoost

Classification report:

	precision	recall	f1-score	support
0	0.65	0.45	0.53	118
1	0.70	0.86	0.77	228
2	0.78	0.70	0.74	128
accuracy			0.71	474
macro avg	0.71	0.67	0.68	474
weighted avg	0.71	0.71	0.70	474

RandomForest

	precision	recall	f1-score	support
0	0.85	0.35	0.49	118
1	0.66	0.90	0.76	228
2	0.75	0.67	0.71	128
accuracy			0.70	474
macro avg	0.76	0.64	0.66	474
weighted avg	0.73	0.70	0.68	474

Logistic Regression

	precision	recall	f1-score	support
0	0.62	0.47	0.54	118
1	0.71	0.80	0.75	228
2	0.67	0.66	0.66	128
accuracy			0.68	474
macro avg	0.67	0.64	0.65	474
weighted avg	0.68	0.68	0.67	474

Performance Comparison

XGBoost

Classification report:

	precision	recall	f1-score	support
0	0.65	0.45	0.53	118
1	0.70	0.86	0.77	228
2	0.78	0.70	0.74	128
accuracy			0.71	474
macro avg	0.71	0.67	0.68	474
weighted avg	0.71	0.71	0.70	474

- 0 = neutral, 1 = negative, 2 = positive
- F1 score for negative and positive classifications are 0.77 and 0.74 respectively
- Model cannot reliably label neutral comments but it's alright because we are more focused on positive and negative posts/comments
- Overall weighted F1 score is 0.70

Performance Comparison

Author	Model	Data Source and Dataset	Accuracy (%)
Pang et al. (2002) [21]	NB, ME, SVM	Movie reviews (IMDb)- 700 (+) and 700 (-) reviews	77−82.9
Dave et al. (2003) [7]	NB, ME, SVM	Product reviews(Amazon)	88.9
Pang & Lee (2004) [18]	NB, SVM	Movie reviews (IMDb)- 1000 (+) and 1000 (-) reviews	86.4-87.2
Gamon (2004) [10]	SVM	Customer reviews (feedback)	69.5-77.5
Pang & Lee (2005) [20]	SVM, SVR, Regression, Metric Labeling	Movie reviews (IMDb)- 5006 reviews	54.6-66.3
Cui et al., (2006) [5]	Winnow, Discriminative ML classifier	Online electronic product reviews- 320k reviews	F1 Score-0.90
Kennedy & Inkpen (2006) [13]	SVM	Movie reviews (IMDb)- 1000 (+) and 1000 (-) reviews	80– 85.9
Chen et al. (2006) [4]	Decision Trees C4.5, SVM, NB	Books Reviews (Amazon)- 3,168 reviews	84.59
Boiy et al. (2007) [3]	SVM, Multinomial NB, ME	Movie reviews (IMDb)- 1000 (+) and 1000 (-) reviews, Car reviews- 550 (+) and 222 (-) reviews	90.25
Annett & Kondrak (2008) [1]	SVM, NB, Decision Tree	Movie reviews (IMDb)- 1000 (+) and 1000 (-) reviews	Greater than 75%
Shimada & Endo (2008) [23]	SVR, SVM OVA, ME	Product Reviews (video games)	NA
Dasgupta & Ng (2009) [6]	SVM and Clustering based	Movie reviews (IMDb) and product reviews (Amazon)- 1000 (+) and 1000 (-) reviews	69.5-93.7
Ye et al. (2009) [31]	NB, SVM and Character based N - gram model	Travel blogs from travel.yahoo.com- 591 (-) and 600 (+) reviews	80.71-85.14
Paltoglou & Thelwall (2010) [17]	SVM	Movie Reviews (IMDb)- 1000 (+) and 1000 (-) reviews, Multi-Domain Sentiment Dataset (MDSD)- 8000 reviews	MR-96.90, MDSD-96.40
Xia et al. (2011) [28]	NB, ME, SVM, meta-classifier combination	Movie Reviews (IMDB), product reviews (Amazon)- 1000 (+) and 1000 (-) reviews	88.65
Kang et al. (2011) [12]	Improved NB	Restaurant Reviews- 5700 (+) and 757 (-) reviews	83.6

- Excerpt from a paper titled *'Sentiment Analysis: A Comparative Study on Feature Selection and Classification Techniques'*
- Reported F1 scores and accuracies ranging from 60~90% for various sentiment analysis models

Performance Comparison

NLTK's VADER

Test F1 score: 0.30695827836035205

	predicted neutral	predicted negative	predicted positive
true neutral	16	49	53
true negative	91	44	93
true positive	14	12	102

Stanza's Sentiment Analyzer

Test F1 score: 0.30695827836035205

	predicted neutral	predicted negative	predicted positive
true neutral	16	49	53
true negative	91	44	93
true positive	14	12	102

VADER: rule-based sentiment analysis tool that uses a pre-constructed dictionary of sentiment-laden words to determine the sentiment of a given text.

Stanza: uses a deep learning-based approach, relying on a pre-trained neural network model to identify the sentiment of a given text.

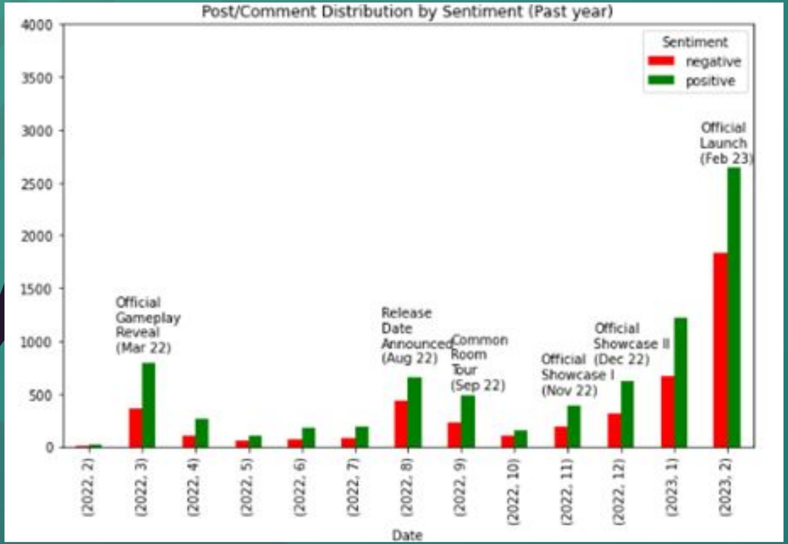
Conclusion: Relevance of the training data is a key determinant of model performance

SENTIMENT ANALYSIS

Insights from our model

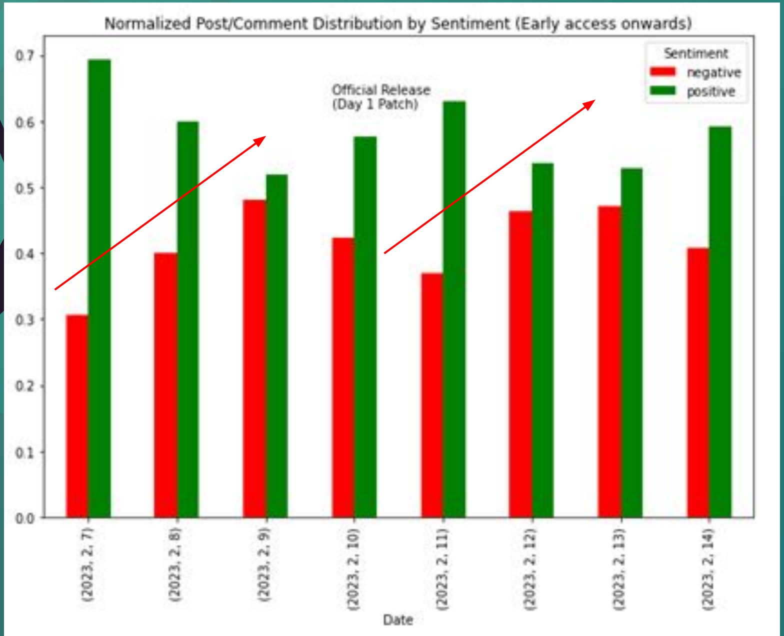
3

Engagement Levels Over Time



- Steady uptick in activity over the past year
- Noticeable spike in activity in months where marketing materials were released
- Exponential increase in engagement as the studio ramped up marketing efforts
- Sentiments were largely positive throughout

Post-launch Sentiments



- Similar patterns observed in the first three days of early-access and that of official launch.
- Exceptional initial response, followed by increase in proportion of negative discussions
- Hypothesis to be confirmed later:
  - Uptick mainly due to influx of excited new players
  - Game was plagued with issues which became apparent after the initial sense of excitement wore off
  - Day-1 patch was ineffective

Generating Insights

Extracting key complaints from negative reviews & discussions

4

Methods

Network analysis

Determine and isolate the **most significant** posts/comments/reviews

Topic modeling

Identify the **primary ideas and themes**

Network Analysis

We used a **bag of words** approach alongside the **page-rank algorithm** to determine the most significant posts and comments.

- Nodes: Individual posts/comments
- Undirected Edges: common unique tokens between two documents
- Edge Weights: common unique tokens divided by total number of unique tokens

Negative Steam Reviews

- Poor performance of game on PC, even on PCs which met or exceeded the minimum requirements
- Frame rate drops
- Stuttering

Reddit Discussions

Before Release

- Eager anticipation for Harry Potter theme and its exclusive activities
- Compliment passion and effort of developers
- Concerns over price and pre-ordering system
- Skepticism over combat system

After Release

- Developers exhibited passion
- Intricate design of Harry Potter universe
- Poor performance on PC
- Various bugs and gameplay features that could be optimized
- Lack of immersion - dearth of interactivity

Topic Modeling

We used Non-Negative Matrix Factorization (NMF). NMF is a matrix factorization technique that groups frequently occurring words together.

Negative steam reviews

- Poor performance of game on PC
- Problems with the combat system

Post-release Reddit Discussions

Positive

- Mainly focused on the game’s visuals
- Character design, rooms, castles

Negative

- Similar to negative steam reviews
- “annoying” and “weird” used to describe characters
- Character actions and interactions not well received

Insights Generated

Insights from Network Analysis &
Topic Modelling

Insights Generated

- Day 1 patch which the developers claim will solve performance issues likely ineffective
- High proportion of positive comments following game's release likely due to new players' excitement
- Problems mentioned previously caused negative sentiment to increase following the initial influx of positive sentiments

Conclusion

Summary and Limitations

5

Summary

Overall well-received

- Positive sentiments outweigh the negative
- Overwhelmingly positive ratings by professional reviewers

IGN 9

STEAM 92%

Marketing strategies were successful

- Significant spikes in engagement

Areas of improvement

Gameplay experience

- Generic storyline
- Lack of meaningful player choice

Technical issues

- Many technical issues persisted
    - Frequent crashes
    - Texture pop-in
    - Input lag
- Despite the Day-1 patch

Recommendations

While challenging to make significant changes to the storyline and interactivity after the launch,

- Avalanche should take note of these issues and use them as a **guide** for future titles or downloadable content

Limitations

- Only 2500 data points used for training
  - Label more data points for model training
  - May lead to better performance

- Models not able to account for linguistic complexities such as **double negatives and sarcasm**
  - Try neural network models with suitable architectures
  - Try other approaches for extracting features (e.g., Word2Vec)
  - Use large language models (LLMs) for sentiment analysis instead

Thank you!

Methodology

```python
params_xgbm = {
    'max_depth': Integer(3, 10),
    'learning_rate': Real(0.001, 1.0, prior='log-uniform'),
    'reg_alpha': Real(1e-9, 1.0, prior='log-uniform'),
}

gs_xgbm = BayesSearchCV(
    n_jobs = -1,
    estimator = xgbm,
    search_spaces = params_xgbm,
    cv = 5,
    scoring = 'f1_weighted',
    random_state= 42
).fit(X_train, y_train)
```

Hyperparameter Tuning with BayesSearchCV

- Bayesian Optimization
- Define range of values for each hyperparameter
  - Real - Specifies a continuous hyperparameter space
  - Integer - Specifies a discrete hyperparameter space
- Use log-uniform prior when we expect the optimal value to be found over several orders of magnitude
- Use uniform when we have a clear range of possible values