ENHANCING CHURN FORECASTING WITH SENTIMENT ANALYSIS OF STEAM REVIEWS

A Focus on Multiplayer Competitive Games

Yiwen Guo
Student ID: 120020118

Abstract

This study explores integrating sentiment analysis into churn forecasting for online multiplayer games, building on the research of Rahman et al. (2024). While traditional churn models rely solely on player count data, this study investigates whether incorporating player sentiment scores derived from Steam reviews can improve the prediction of player churn. Focusing on three of Steam's most popular multiplayer competitive games (*CS 2*, *Dota 2*, and *PUBG*), a Vector Autoregression (VAR) model was used to analyze the dynamic interplay between player sentiment and engagement. In addition to replicating prior models, this study extends them to a broader set of games, demonstrating the practical value of sentiment-driven forecasting for game developers to anticipate and mitigate player churn.

1. Motivation

Traditional churn prediction models primarily rely on player count-based regression models, where player engagement is forecasted solely using historical player count data. Recent research by Rahman et al. (2024) demonstrated that incorporating sentiment analysis from player reviews significantly enhances the predictive accuracy of churn models. This study integrates sentiment scores into a Vector Autoregression (VAR) model, enabling a more comprehensive analysis of the dynamic interplay between player sentiment and engagement. By incorporating both player counts and sentiment, this study is expected to achieve more accurate churn forecasts than traditional models.

However, the research mentioned above primarily focused on a limited set of single-player, standalone games (*Bloons TD*, *ONI*, and *Stellaris*). These games have relatively stable player bases, infrequent updates, and minimal community-driven fluctuations, leading to limited variability in player counts, review sentiment, and their interactions. As a result, the effectiveness of the model in a more dynamic and competitive gaming environment remains uncertain.

To address this limitation, this study extends the methodology to three most popular multiplayer competitive games on Steam: *CS 2*, *Dota 2*, and *PUBG* (SteamDB, 2024). These games consistently rank among the highest for concurrent player counts on the platform. Unlike single-player simulation games, player engagement in multiplayer games is subject to frequent version updates, esports tournaments, and community-driven discussions, which can cause sudden and significant shifts in player sentiment. These volatile fluctuations make multiplayer games ideal case studies for examining the relationship between player sentiment and churn.

By applying this approach to *CS 2*, *Dota 2*, and *PUBG*, this study aims to offer practical insights for game developers. The results could enable developers to better anticipate player churn, identify early warning signs of disengagement, and respond to shifts in player sentiment in a timely manner. Ultimately, this study seeks to demonstrate that sentiment-enhanced churn models are more effective than traditional models for predicting churn in dynamic, competitive gaming environments.

2. Data Collection

2.1 Sentiment Score (S_t)

Due to the limitations of the Steam API, individual users can only extract review data within a maximum time range of the most recent one-month period. Therefore, this study collected 100 randomly sampled reviews per day from November 11 to December 11, 2024, using Python web scraping techniques. The first 25 days of review data were used for model training, while the remaining 5 days were used for forecasting and evaluation. The sentiment score for each day was calculated as the ratio of positive reviews to total reviews. (For the detailed code, please see the attachment “WebScrapingSteamAPI.py.”)

Example of Sentiment Score Data:

Date	Positive Reviews	Negative Reviews	Sentiment Score
2024-11-11	65	35	0.65
2024-11-12	46	54	0.46
2024-11-13	59	41	0.59
2024-11-14	56	44	0.56
2024-11-15	45	55	0.45
2024-11-16	58	42	0.58
2024-11-17	59	41	0.59
2024-11-18	65	35	0.65
2024-11-19	64	36	0.64
2024-11-20	51	49	0.51

(*CS 2*)

2.2 Player Count (C_t)

SteamDB serves as an open-access database, allowing any user to view historical information on game player counts, sales, and other related metrics. This study directly extracted the daily average concurrent player count through SteamDB. Since SteamDB provides hourly player count data, the daily player count data was calculated as the average of hourly player counts for each day. To ensure that the daily player count data was consistent and comparable to the daily sentiment data, the extraction period spanned from November 11 to December 11, 2024.

Example of Player Count Data:

DateTime	Average Players
2024-11-11	823942
2024-11-12	827937
2024-11-13	800513
2024-11-14	854272
2024-11-15	913356
2024-11-16	988464
2024-11-17	977993
2024-11-18	885376
2024-11-19	864975
2024-11-20	876296

(*CS 2*)

3. Model Specification

3.1 VAR model

A VAR model was constructed to capture the dynamic interactions between player counts and sentiment scores. The VAR model allows player counts to be influenced by sentiment scores and vice versa, which leverages past information from both series to forecast future player counts.

The model is defined as:

$$\mathbf{Y}t = A_1\mathbf{Y}t - 1 + A_2\mathbf{Y}t - 2 + \cdots + A_p\mathbf{Y}t - p + \mathbf{u}_t$$

Where:

- $\mathbf{Y}_t = [C_t, S_t]^\top$ is the vector of endogenous variables: player count (C_t) and sentiment score (S_t) at time t.
- $A_1, A_2, \cdots, A_p$ are coefficient matrices for the past lags of player count and sentiment.
- p is the lag order, which is selected using the Akaike Information Criterion (AIC).
- $\mathbf{u}_t$ is a white noise error term.

The model was estimated using data from November 11 to December 11, 2024, with the first 25 days used for model training and the final 5 days reserved for forecasting player counts. The forecasted player counts for the last 5 days were then compared with the actual observed player counts to evaluate the model's predictive accuracy.

3.2 Stationarity Test and Differencing

To ensure stationarity, the Augmented Dickey-Fuller (ADF) test was performed on both player counts and sentiment scores. If the series was non-stationary, differencing was applied to make it stationary.

Results of the ADF Test Before and After Differencing:

```
Results of ADF Test Before Differencing:
{'Variable': 'Average_Players', 'ADF Test Statistic': -1.43, 'p-value': 0.569, 'Stationary?': False}
{'Variable': 'sentiment_score', 'ADF Test Statistic': -4.32, 'p-value': 0.0, 'Stationary?': True}

Order of Differencing Applied:
Average_Players: 1
sentiment_score: 0

Results After Differencing:
{'Variable': 'Average_Players', 'ADF Test Statistic': -9.88, 'p-value': 0.0, 'Stationary?': True}
{'Variable': 'sentiment_score', 'ADF Test Statistic': -3.85, 'p-value': 0.002, 'Stationary?': True}
```

(*CS 2*)

```
Results of ADF Test Before Differencing:
{'Variable': 'Average_Players', 'ADF Test Statistic': 1.39, 'p-value': 0.997, 'Stationary?': False}
{'Variable': 'sentiment_score', 'ADF Test Statistic': -3.57, 'p-value': 0.006, 'Stationary?': True}

Order of Differencing Applied:
Average_Players: 2
sentiment_score: 0

Results After Differencing:
{'Variable': 'Average_Players', 'ADF Test Statistic': -8.48, 'p-value': 0.0, 'Stationary?': True}
{'Variable': 'sentiment_score', 'ADF Test Statistic': -3.94, 'p-value': 0.002, 'Stationary?': True}
```

(*Dota 2*)

```
Results of ADF Test Before Differencing:
{'Variable': 'Average_Players', 'ADF Test Statistic': -2.04, 'p-value': 0.268, 'Stationary?': False}
{'Variable': 'sentiment_score', 'ADF Test Statistic': -4.88, 'p-value': 0.0, 'Stationary?': True}

Order of Differencing Applied:
Average_Players: 2
sentiment_score: 0

Results After Differencing:
{'Variable': 'Average_Players', 'ADF Test Statistic': -12.21, 'p-value': 0.0, 'Stationary?': True}
{'Variable': 'sentiment_score', 'ADF Test Statistic': -4.86, 'p-value': 0.0, 'Stationary?': True}
```

(*PUBG*)

3.3 Model Selection

To select the optimal lag order p, the AIC was calucated for lags ranging from 1 to 5. The optimal lag was selected as the one that minimized the AIC.

The AIC vs. Lag Order:

```
AIC for lag 1: -12.542
AIC for lag 2: -14.654
AIC for lag 3: -15.876
AIC for lag 4: -15.785
AIC for lag 5: -15.543
Optimal Lag (p) for CS2: 3
```

(*CS 2*)

```
AIC for lag 1: -10.786
AIC for lag 2: -12.254
AIC for lag 3: -13.871
AIC for lag 4: -14.092
AIC for lag 5: -14.354
Optimal Lag (p) for Dota 2: 5
```

(*Dota 2*)

```
AIC for lag 1: -11.231
AIC for lag 2: -12.543
AIC for lag 3: -12.994
AIC for lag 4: -12.786
AIC for lag 5: -12.654
Optimal Lag (p) for PUBG: 3
```

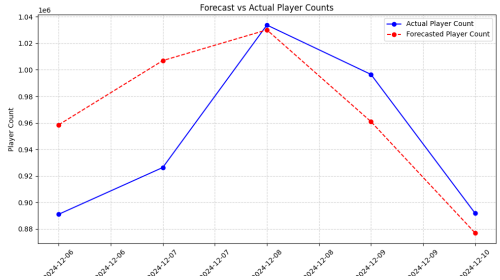
(*PUBG*)

4. Results

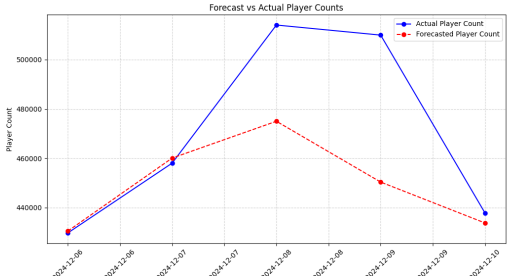
4.1 Forecasting Results

Using the estimated VAR model, player counts were forecasted for the final 5 days of November 11 to December 11, 2024. The forecasting results were compared with the actual player counts for each day.

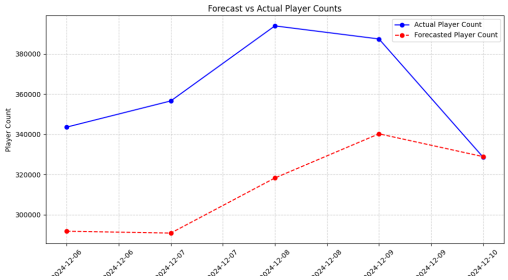
Forecasting Results vs. Actual Data:



(*CS2*)



(*Dota 2*)



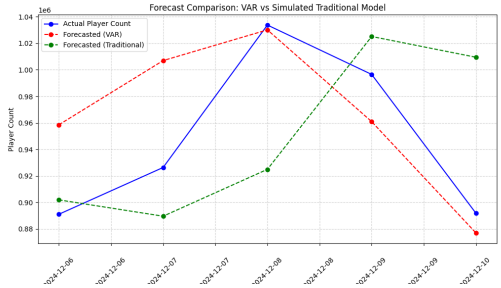
(*PUBG*)

4.2 Forecasting Evaluation

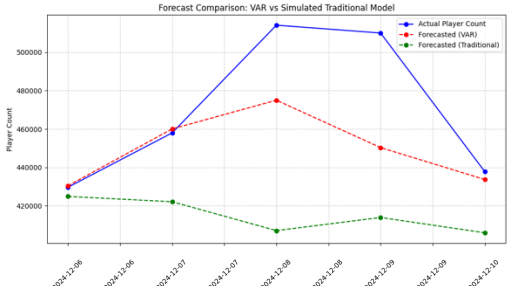
To assess the contribution of sentiment score to forecasting accuracy, two models are compared:

- VAR Model with Sentiment: Uses player count and sentiment score as endogenous variables.
- Traditional Churn Model: Uses player count alone (ARIMA with only player count).

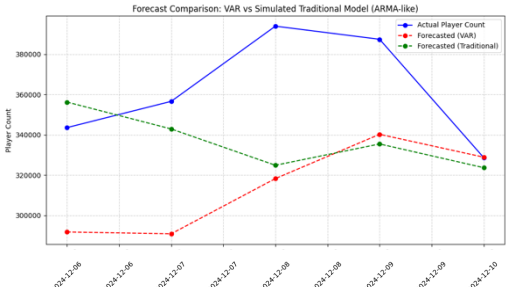
Forecasting Results of VAR Model with Sentiment vs. Traditional Churn Model:



(*CS2*)



(*Dota 2*)



(*PUBG*)

Forecasting Accuracy Evaluation of VAR Model with Sentiment vs. Traditional Churn Model:

```
📊 VAR Model Forecast Accuracy:
MAE  (Mean Absolute Error): 40454.24
RMSE (Root Mean Squared Error): 50127.21
MAPE (Mean Absolute Percentage Error): 4.38%
R^2  (Coefficient of Determination): 0.24

📊 Traditional Model Forecast Accuracy:
MAE  (Mean Absolute Error): 60570.99
RMSE (Root Mean Squared Error): 74832.65
MAPE (Mean Absolute Percentage Error): 6.36%
R^2  (Coefficient of Determination): -0.69
```
(*CS2*)

```
📊 VAR Model Forecast Accuracy:
MAE  (Mean Absolute Error): 21127.52
RMSE (Root Mean Squared Error): 31962.02
MAPE (Mean Absolute Percentage Error): 4.17%
R^2  (Coefficient of Determination): 0.20

📊 Traditional Model Forecast Accuracy:
MAE  (Mean Absolute Error): 55097.21
RMSE (Root Mean Squared Error): 67805.15
MAPE (Mean Absolute Percentage Error): 11.17%
R^2  (Coefficient of Determination): -2.62
```
(*Dota 2*)

```
📊 VAR Model Forecast Accuracy:
MAE  (Mean Absolute Error): 48092.53
RMSE (Root Mean Squared Error): 54651.72
MAPE (Mean Absolute Percentage Error): 12.99%
R^2  (Coefficient of Determination): -3.75

📊 Traditional Model Forecast Accuracy:
MAE  (Mean Absolute Error): 30445.90
RMSE (Root Mean Squared Error): 39572.62
MAPE (Mean Absolute Percentage Error): 7.99%
R^2  (Coefficient of Determination): -1.49
```
(*PUBG*)

The VAR model outperformed the traditional model for *Dota 2* and *CS2*, successfully explaining a portion of the variance in player counts ($R^2 > 0$) and achieving lower errors across all evaluation metrics.

For *PUBG*, however, both models failed to deliver accurate forecasts. While the traditional model slightly outperformed the VAR model in terms of MAE and RMSE, its negative R^2 value highlights its inability to capture the highly volatile player counts. More complex models that integrate exogenous variables or additional dynamic features are needed to better capture the fluctuating nature of games like *PUBG*.

5. Conclusion

This study evaluated the performance of a Vector Autoregression (VAR) model for forecasting player counts for three popular online multiplayer games: *CS 2*, *Dota 2*, and *PUBG*. The results demonstrated that the VAR model performed well for *Dota 2* and *CS 2*, where it outperformed a

traditional player count-based model. In both cases, the VAR model explained a portion of the variance in player counts ($R^2 > 0$) and achieved lower errors across MAE, RMSE, and MAPE. This suggested that the VAR model effectively captured the dynamic relationship between player sentiment and counts for these games, especially when player behavior was relatively stable.

However, for *PUBG*, the VAR model failed to outperform the traditional model, showing a negative R^2 (-3.75), indicating that it performed worse than a naive prediction based on the mean. *PUBG*'s player behavior was more volatile, with sharp spikes in player counts often triggered by game updates, esports events, and marketing promotions. The current VAR model could not fully capture these external shocks, which may explain its underperformance compared to the simpler traditional model.

6. Limitations and Improvements

First, the dataset's size and scope need to be expanded. Due to Steam API limitations, the current analysis is restricted to a short-term dataset from November 11 to December 11, 2024 for player counts and review sentiment. Expanding the dataset to include multiple months or years of data could enable the VAR model to identify longer-term seasonal patterns, weekly cycles, and major spikes from esports events. Future studies would use advanced web scraping tools to continuously collect player reviews over an longer period. A larger training set would improve the model's stability and reduce the risk of overfitting.

In addition, the sentiment analysis approach should be refined.The current sentiment score relies on a simple ratio of "positive" to "total" reviews, which may introduce biases. For example, in competitive games like *CS 2*, *Dota 2*, and *PUBG*, players may leave negative reviews not because they dislike the game, but due to frustration with teammates, the in-game environment (toxic behavior), or server issues. These sentiments do not necessarily indicate a likelihood of churn. A more advanced approach would be to incorporate Natural Language Processing (NLP) techniques to analyze the content of player reviews, rather than just relying on binary "positive/negative" labels. For instance, using sentiment polarity detection or emotion classification models could provide more nuanced sentiment scores. Machine learning models like BERT or RoBERTa could extract deeper semantic meaning from review text, which could better inform the player churn prediction process.

Lastly, external features should be incorporated to improve prediction accuracy. The player count fluctuations in *PUBG* are significantly influenced by external shocks, such as game patches, promotional events, and esports tournaments. The current VAR model, however, relies solely on lagged player counts and sentiment scores, limiting its capacity to capture these sudden shifts. Exogenous features, such as patch update dates and esports event schedules, could act as external predictors, allowing the model to anticipate spikes or drops in player counts. By incorporating these additional explanatory variables, the model would no longer be constrained to only internal game data. Instead, it could adjust player count forecasts on days when significant updates or events are expected, resulting in more accurate and timely predictions.

References

Rahman, S., Ali, M., Bakar, A., & Mutalib, S. (2024). Enhancing churn forecasting with sentiment analysis of Steam reviews. *Social Network Analysis and Mining, 14*(1). https://doi.org/10.1007/s13278-024-01337-3

SteamDB. (2024, December 11). *Most played games.* Retrieved from https://steamdb.info