Study of Gender-based Playing Style Stereotype in Overwatch using Machine Learning Analysis

Ralph Jinghao Zhou¹, Yumeng Gao² and Qian Han³

¹J.N.Burnett, Richmond, British Columbia, Canada
²School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
³Department of Computer Science, Dartmouth College, Hanover, NH, United States

*Corresponding author’s e-mail: ralphzhou030821@gmail.com

Abstract. Since its release in 2016, Overwatch has become a game with a high amount of daily users (over 50 million users Worldwide) and made a great amount of profit for the company that produced it — Blizzard. However, it is common that the stereotypes and biases hide under the popularity of the games. In past researches, it is shown that gender could affect the designs, abilities, and player’s view of in-game characters. In this research, a survey regarding gender stereotypes in Overwatch is conducted, and statistical and machine learning methods are applied to analyze the results. The result suggests that many of the players in the game had given biased or stereotypical actions or thoughts toward female players. Simultaneously, we propose an Ensemble Late Fusion (ELF) method that unifies the weighted predicted probabilities generated from various state-of-art classifiers to classify the player behaviors with or without gender-based stereotypes in Overwatch. The experimental results show that when using ELF, the optimal classification result reaches AUC to 0.9044, which beats several well-known traditional machine learning and deep learning classifiers. Moreover, the connection between certain biased ideas and rather a user in sexist is graphed and discussed in the paper.

1. Introduction

MOBA is the abbreviation of Multiplayer Online Battle Arena Games [2]. The originator of MOBA games was “Star-Craft” which was born in 1998. In MOBA games, players are usually divided into two teams. The two teams compete with each other on a scattered game map. Each player controls a single game character through an interface. Several players with different roles in each team cooperate until the opponent’s base is demolished to achieve the goal of victory. In the MOBA game, the necessary elements are: a fixed base and defensive towers, regular dispatch of troops at a certain frequency, control of multiple heroes by both sides to fight against each other, and the demolition of the opponent’s base is victory.

Overwatch [3], [4] is a very popular MOBA team shooting game with a background of the near-future earth, which is globally listed in 2016. Each game will be joined by many different styles of heroes, mercenaries, scientists, adventurers, and strange people, fighting each other in an epic global battle. Overwatch is a team first-person shooter game. The heroes in the game are in an era full of disputes. Game tasks are divided into two types, one is “escort”, that is, the attacker must transport the designated target to the designated location, and the defender must stop the attacker until the end of the time. The other is “occupation”, where the two sides will fight for the control of the map. One
team will attack and the other will defend. The main purpose of the attacker is to occupy the key markers, and the defender must maintain control until the end of time. The positioning of game characters can be roughly divided into four types, namely assault, defence, reloading, and auxiliary. Players must cooperate to win the game.

However, there are various direct or indirect gender stereotypes in MOBA games, among which is the gender discrimination of male players against female players [7]. Many male players use excessively exaggerated ways to show their love or dislike for women. For example, when female players initiate any game-related topics, they will only focus on her gender, ignoring the true theme of her topic. Also, it is quite often those male players who play MOBA games most, and the MOBA game is also the patent of male players. But in fact, with the development of the MOBA game market, female players have gradually entered the consumer market of MOBA games. But the traditional patriarchal concept strongly restricts the development of female players in the game. This is not only reflected in the gender discrimination and stereotypes in the game, at the same time, but the male players have also expanded the scope of discrimination to female game practitioners and female e-sports players, etc.

Besides, female players are more likely to be influenced by social relationships and circles. More than 60% of female players started playing Overwatch driven by friends around or driven by curiosity, and the proportion of female players joining due to friend recommendations and being influenced by social networks is also higher than that of male players. Of course, there are more intuitive data. More than 70% of male players are more concerned about same-gender game recommendations. This can be understood as they are more like-minded. The driving force for the growth of female players is more driven by social communication and other users, and there is no obvious gender-driven tendency.

In this paper, we report on the study of the gender stereotype in Overwatch, one of the most popular MOBA games in the world since 2016. We focus on studying whether and how gender stereotypes exist, from male players to female players in the game. In the next section, we review the literature on stereotypes, gender stereotypes, and machine learning analysis in video games. We then describe the research methods and report the results of our approach to mixed methods, which includes a quantitative analysis of the most likely cause of gender stereotypes through machine learning and qualitative analysis of the collected answers from the conducted online survey. Finally, we provide a discussion of the consequences and ways to alleviate gender stereotypes in competitive MOBA games.

2. Literature Review

2.1. General Stereotypes

The survey [10] reviews the psychological literature on stereotypes, with special emphasis on cognitive and motivational factors that contribute to the formation, maintenance, application, and modification of stereotypes. In addition, the context-sensitive functions of stereotypes are emphasized and involving the representative issues of various stereotype models. This research [20] mainly focuses on gender representation in children’s cartoons that was conducted mainly in the 1970s. The results show that there is a significant difference between the salience and representation of male and female roles. Both male and female characters are stereotypes. Compared with female characters, male characters are more important, appear more frequently, participate in almost all the mentioned behaviors, and speak more. When the male or female behavior and communication variables are divided by the number of male or female characters or the total talking time, the results show that they are consistent with gender role stereotypes. In [14], the authors examine the ideological meaning of racial stereotypes in comedy through the text and audience analysis of “Rush Hour 2”. Most participants, regardless of race, believed that the racial jokes in the film were harmless. Many Asian and black participants found a source of positive pleasure in the negative representations of their race and did not make opposition statements. The result shows that universal conventions and textual methods in comedy encourage viewers to naturalize racial differences rather than challenge racial stereotypes.
2.2 Gender Stereotype in Video Games
Stereotypes in video games are very common. This article [12] begins with a comprehensive review of previous research on gender and ethnic character portrayal in video games. Then, a small-scale content analysis was conducted on the official trailer samples, introduction sequences, and covers of 19 most popular video games, and then the authors discussed the meaning of stereotypes in video games and the possible social and psychological effects on players. This paper [13] analyzed multiple billions of chat messages from Twitch, a social game streaming platform. It explored how streamer gender is related to the nature of conversation using a combination of computational text analysis methods (e.g. neural Embedding methods in vector spaces (paragraph vectors)). The result showed that conversation and gender objectification are common in chat, and female streamers get more comments about the game. The survey [11] focused on 11 online games to see how chat, gesture, the form of behavior, gender, and men with and without female avatars differ by “gender switcher”. Based on the results of the analysis, men may not want to hide their offline gender when using female avatars, but there is evidence that they reinforce the idealized concept of women’s appearance and communication. Nonetheless, choosing the gender of the avatar is not an expression of identity, but rather a strategic choice for players than the multi-modal code available in this exploration of digital space. This study [16] explores the experience of female players in one of the most popular games in the world, League of Legends, and there is a shortage of female players in a community where female players are as skilled as men. In addition, women who play with male partners are confident in their skills and often focus on supporting their partners rather than their own development. This study shows that one way to close the gender gap in games is to better understand and improve the social dynamics of popular games. In this paper [9], we present data from the World of Warcraft and Rift multi-site survey. This confirms this stereotype existing between novice and expert players. Experts are more likely than novices to adopt rules for avatar selection.

2.3 Machine Learning analysis in online activity
In this survey [15], the author classifies the main sentiments and technical features taken from the game plot using machine learning techniques. The author defined video game success rates such as low, medium, high, and very high. Its implementation includes the use of word2vec-based technology for natural language processing (NLP). The classification algorithms in the proposed research are Support Vector Machine (SVM), Artificial Neural Network (ANN), K-Nearest Neighbor Method (KNN), and Random Forest (RF). This model is evaluated using a confusion matrix with accuracy, precision, and an F score. This study [19] explored the possibility of creating an item recommendation system that adapts to specific situations and providing recommendations for items that have the greatest impact on the chances of winning the game League of Legends. The authors build a prediction algorithm based on artificial intelligence with a neural network that predicts the winning team in the game League of Legends. Then, the study navigates through some of the potential measures of the importance of functionality and use LIME as a way to estimate sorter locally to explain predictions. This paper [18] first systematically reviews and compares the performance of the machine learning algorithms most often used to predict match winners in DotA 2 computer game team drafts and aims to consider interactions with heroes in the draft. For this reason, the authors tested machine learning algorithms such as Naive Bayes Classifier, Logistic Regression, and Gradient Boosted Decision Trees. Also, the accuracy of the model prediction depends on the skill level of the player. The researchers have prepared a publicly available dataset that can be used by developing, testing and comparing algorithms, taking into account the differences in the data used in previous studies. This research [21] aims to achieve higher accuracy by adding play length as an input function on DotA 2 data. Additionally, the author applied a Multi-layer feed-forward neural network to predict game results with GPU available. The results showed that increasing the duration of the game did not improve the neural network’s performance and also did not exceed the performance of logistic regression. This work [8] examines the relationship between early learning speed and long-term performance using a longitudinal dataset of 400,000 players generated by new player’s observation of the widely played
MOBA League of Legends. The learning rate of new players in the competitive season explains the drastic fluctuations in the performance of the year-end. Two multivariate classifiers (logistic regression random forest) are used to predict who will be by the end of the season are considered master (top 0.05%) based on its performance. Two classifiers have similar prediction results, which extend the technology prediction period based on relatively sparse samples of initial data. In this paper [17], the author uses an artificial neural network (ANN) as a classifier to detect the gender of an email author, and a whale optimization algorithm (WOA) to find the optimal like weight and bias to improve the accuracy of ANN classification.

For our research, we mainly focus on the study of the gender stereotype in the popular online video game Overwatch and study whether and how gender stereotypes exist using both quantitative and qualitative analysis. In a word, our research is the first paper to apply machine learning analysis in gender stereotypes in Overwatch, which differs from all the previous research work conducted on the general stereotype, stereotype in video games, and machine learning analysis in video games.

3. Evaluation Pipeline

Figure 1. Overwatch [1] (First-person shooter MOBA game)

Team Roles are the roles each player plays in a team, and in Overwatch, the team role is mainly composed of the following three roles: DPS, Tank, and Support. The DPS roles are usually filled by heroes that deal the most damage. The Tank role on a team is filled by the player that is playing the Main Tank or Off-tank in the composition. The Support role is played by people playing characters that can heal other players (or support them in other ways). Historically, the Support role has been split into Main Support and Flex Support. The definition of these two roles has changed over time and is hard to define indefinitely. Generally, however, the Main Support is a character who focuses on less mechanically demanding jobs, and will likely have the most healing done [6].

Figure 2. Overwatch Roles [6]
In this paper, we adopt a mixed-method approach by using both quantitative analysis and qualitative analysis to find out if gender stereotypes are existing in Overwatch. In quantitative analysis, we conducted a gender-based analysis of the conducted survey collected from the Overwatch Reddit Forum with different rank levels, asking each of the participants 32 questions about their feeling about gender stereotype of different races, and then we asked another 10 professional Overwatch players to label all the collected data from the anonymous player to check if they have the gender stereotype according to their answer. The collected data are further used to fit the state-of-art machine learning models to find the top features which cause the gender-based stereotype. Qualitative analysis is conducted interviews with participants to their experiences and perceptions of playing Overwatch toward gender stereotype toward female players.

3.1.Pipeline

![Figure 3. Scheme of the machine learning analysis pipeline](image)

The pipeline of the overall machine learning analysis process is shown in Figure 3. First, we use all the collected questionnaires as a dataset, and manually labeled the dataset with gender-based stereotype data. After removing the non-compliant questionnaire, the dataset was cleaned and sorted into the characteristics required for the experiment. Then, we send the feature vectors into 7 state-of-art supervised learning classifiers together with the proposed ensemble late fusion (ELF) method, and use 10-fold cross-validation and grid parameter search to predict the results of gender stereotype samples in Overwatch. In addition, we also used the feature selection analysis library embedded in the classifiers to select the top features separately and find the features that are most gender-discriminatory to female players.

3.2.Training Protocols

In the machine learning analysis procedure, we applied 10-fold cross-validation for the training process. Cross-validation is a re-sampling technique used to evaluate machine learning models with limited data samples. The procedure has a single parameter called k, which refers to the number of groups a given sample of data is divided into. Therefore, this procedure is often referred to as the k-fold cross-validation. If we choose a specific value for k, you can use it in the model reference instead of k. For example, k = 10 is a 10-part cross-test. Cross-validation is mainly used in applied machine learning to evaluate the capabilities of machine learning models for invisible data. That is, a limited sample is used to evaluate how the model should generally perform when making predictions on data that was not used during model training. This is a common method because it is easy to understand and generally has less bias in model capabilities and optimistic estimation than other methods such as simple training/testing splits.
3.3. Feature Selection

Feature selection is one of the fundamental concepts of machine learning that has a significant impact on model performance. The functions used to train machine learning models have a significant impact on achievable performance. Unused or partially used features can affect model performance. When selecting features, the predictors of interest or characteristics that contribute most to the output are selected automatically or manually. Unrelated features in the data reduce model accuracy and can train the model based on extraneous features. The advantages of feature selection are listed as follows:

- Reduction of overfitting: less redundant data means less chance of making decisions based on noise.
- Better accuracy: Less misleading data means more accurate modeling.
- Reduced training times: the fewer data points, the less complex the algorithm and the faster the training process of the algorithm.

3.4. Supervised Learning Classifiers

In this paper, we applied the following 7 state-of-art classifiers and the proposed ensemble late fusion (ELF) to our collected dataset to predict if gender-based stereotypes toward female players are existing in Overwatch.

1) Decision tree (DT): Decision tree is based on the known occurrence probability of various situations, by constructing a decision tree to obtain the probability that the expected value of the net present value is greater than or equal to zero. In machine learning, a decision tree is a predictive model, which represents a mapping relationship between object attributes and object values. Decision tree is a very commonly used classification method. It is a kind of supervised learning. The so-called supervised learning is to give a bunch of samples, and each sample has a set of attributes and a category. These categories are determined in advance. Then through learning, a classifier can be obtained. The objects are given the correct classification.

2) K-Nearest Neighbor Algorithm (KNN): KNN is characterized by estimating the distance between various features. Its main idea is: given a test data, if most of the K training data closest to it belong to a certain category, then the test data is considered to also belong to this category.

3) Support Vector Machine (SVM): Support Vector Machine (SVM) is a method of supervised learning, which can be widely used in statistical classification and regression analysis. Support vector machines belong to generalized linear classifiers, and it can minimize empirical errors and maximize geometric edge regions at the same time, and then it can be transformed into the solution of a convex quadratic programming problem.

4) Naive Bayes (NB): Statistically, Naive Bayes is a simple family of “rational classifications” based on the application of Bayes theory of complex independent concepts between symbols. They are one of the simplest models of the Bayesian network.

5) Random forest (RF): Random Forest consists of many random decision trees and each decision tree in the random forest is not related. When a new input sample enters, each decision tree in the forest makes a judgment separately to see which category the sample belongs to, and then according to the majority vote, the new sample will be predicted to the corresponding category. Random Forest can handle both discrete values and continuous values. Besides, Random Forest can handle both discrete values and continuous values, and can also be used for unsupervised learning clustering and outlier detection.
6) Logistics regression (LR): Logistics regression is a statistical model that uses its basic model for logistics operations, which returns the dependence of voluntary return, although there is a more complex extension. Analytical regression, logistic regression defines the boundaries of the logistics model (a type of binary regression).

7) Multilayer perceptron (MLP): Multilayer perceptron is a type of feedforward neural network, which contains at least three node layers: input layer, hidden layer and output layer. Each node is a neuron with nonlinear activation function. Except for input nodes, MLP uses a supervised learning method called back-propagation in training process. By activating multiple nonlinear layers, MLP is different from linear perceptrons, and it’s able to distinguish non-linearly separated data.

3.5. Ensemble Late Fusion (ELF)

1) ELF method: The proposed ELF method shown in Figure 3 unifies the predicted result of traditional machine learning and deep classifier into a joint prediction. Instead of using the binary prediction generated in the traditional classifiers, ELF utilizes the returned predicted probability from each classifier that a particular participant belongs to the positive class (has gender stereotype toward the female player) vs. the negative class (no gender stereotype toward the female player). Then, ELF assigns weights (adding up to 1) to each classifier to each returned probability. To determine the best assigned weights, ELF performs a grid search to identify the approximate weight without touching the test set.

To elaborate ELF in more detail, we define the Ensemble Late Fusion Probability (ELFP) as the weighted sum of each returned predicted probability of the positive class from each classifier:

\[
ELFP_j = \sum_{i \in \text{classifiers}} w_i \times prob_j^i
\]

where \( w_i \) is the assigned weights for \( i \)-th classifier when outputting a probability \( prob_j^i \) of belonging to class 1 in the prediction for sample \( j \), and \( \sum_i w_i = 1 \). Then, if \( ELFP_j \geq 0.5 \), the sample \( j \) is assigned label 1, otherwise 0.

2) Optimal training weights: Now we describe how to determine the optimal weight used for the ELF. First, we split the data set into two parts: training set and test set. We only use the training set to identify the best weight. We perform 10-fold CV for all potential combinations of assigned weights for each traditional machine learning classifiers, which sums up to 1. We determine the performance of the validation based on the weighted probability in the proposed equation 1. The best weight combination corresponds to the weight of the average of 10 best performances obtained from the validation set. Then the best ELF result is obtained using the best weights found during the average performance in this process.

4. Experimental Setup

In this section, we provide details regarding the collected datasets through questionnaires and the evaluation metrics.

4.1. Datasets

In our quantitative analysis, we first randomly collected a total of 244 Overwatch players (106 females, 138 males) on Google Form and Overwatch Reddit Forum [5], with similar Overwatch player statistical system distribution. The age of the respondents ranges from 10 - 40, the races of the
respondents are Asian, African American, First Nation, Caucasian, Latino Americans, and Middle Easterners. In the questionnaire, we also designed 31 survey questions for the respondents, including the field like game time, cost in the game, preferred roles in the game, willingness to play with female players, etc. The detailed questions in the survey are as follows:

- What is your age?
- What is your gender?
- What is your race?
- How many days in a month do you play Overwatch?
- How many hours do you play Overwatch per day?
- What do you like the most in Overwatch?
- How much money did you spend in Overwatch including the money spent to buy the game?
- Would you recommend this game to female friends?
- Do you prefer to play casually or competitively?
- How many friends do you usually play with?
- What role do you prefer in this game?
- Do you think you’re good at playing this game?
- What is your rank in Tank?
- What is your rank in Damage?
- What is your rank in Support?
- What strategy do you prefer?
- What is the female player percentage of the game in your opinion?
- In your opinion, which gender is better at the game?
- Why do you think this gender is better?
- Do you team up with male or female more?
- Would you refuse or dislike to play with players other than your race?
- Would you refuse or dislike playing with players speaking other languages?
- Do you think the attack role is more important than the tank or support role?
- Do you think there are more male players in damage role?
- Do you think there are more female players in tank or support role?
- If you choose to play the Damage role, it is most likely because?
- If you choose to play the Tank role, it is most likely because?
- If you choose to play the Support, it is most likely because?
- Do you think Overwatch has a high learning curve for female players?
- Do you think the negative environment from teammates may discourage female players to play the game?
- Have you ever played the game together with your boyfriend/girlfriend?

4.2 Metrics

After building the machine learning model for our classification problem, the effect of the model needs to be evaluated. The evaluation indicators that are often used in the industry are Precision, Recall, F-Measure, etc. The following figure shows the evaluation indicators of different machine learning algorithms.

In general, we have the following 4 basic definitions of the metrics: TP — True Positive. Values that are actually positive and predicted positive; FN — False Negative. Values that are actually positive but predicted to negative; TN — True Negative. Values that are actually negative and predicted to negative; FP — False Positive. Values that are actually negative but predicted to positive. On the basis of these defined basic metrics, we define the following indicators to measure the performance of the built machine learning model.
1) **Accuracy**: The percentage of the correct prediction results in the total sample, the formula is as follows: 
\[
\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}
\]

2) **Precision**: The probability of the actual positive sample among all the samples predicted to be positive is as follows: 
\[
\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}
\]

3) **Recall**: The probability that the actual positive sample is predicted to be a positive sample is as follows: 
\[
\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}
\]

4) **F1-score**: In order to integrate the performance of precision and recall, to find a balance between the two, and F1-score appeared. 
\[
\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

5) **AUC**: AUC (area under the curve) In order to calculate the points on the ROC curve, we can use different classification thresholds to evaluate the classification model multiple times, but this is very inefficient. However, AUC can provide us this kind of information. The coverage of both positive and negative samples should be 50%, which means random effect. The steeper the ROC curve, the better, so the ideal value is 1, a square, and the worst random judgment is 0.5, so the general AUC value is between 0.5 and 1.

6) **False Positive Rate**: False Positive Rate (FPR) = FP / Negative.

7) **False Negative Rate**: False Negative Rate (FNR) = FN / Negative.

### 5. Results

#### 5.1. Classification Results

In this section, we describe the experiments we conducted to assess the performance of our gender-based playing style stereotype surveys in Overwatch using different state-of-art classifiers. We utilized the common 10-fold CV protocol, and in each fold, we divide our dataset into training and test sets. The goal is to find the best parameter combination using the training set alone and then make a prediction on the test set. In order to evaluate the experiment in distinguishing between players who have stereotype against female players and those who do not have stereotype against female players as Figure 4 shows, we tested a suite of 7 classifiers, including both state-of-art machine learning classifiers and a deep learning classifier: Decision Tree (DT), k-Nearest Neighbor (KNN), Random Forest (RF), Naive Bayes (NB), Support Vector Machines (SVM), Logistic Regression (LR), Multi Layer Perceptrons (MLP) and proposed Ensemble Late Fusion (ELF). We additionally changed the input features used to the classifiers to understand which are the top features are affect the classification result most.

![Gender-based Stereotype Percentage](image)

Figure 4. Percentage of players who respond to the questionnaire (244 in total)

Table 1 shows the metrics (e.g. AUC and F1) of using different state-of-art classifiers. We see that in the traditional machine learning classifiers, Random Forest generates the best results, achieving up to 0.8288 AUC and 0.7878 F1, while other classifiers can also generate AUCs from 0.6899 (MLP) to 0.8198 (DT). In the meantime, our proposed Ensemble Late Fusion (with best weights: DT: 0.24, KNN: 0.02, RF: 0.58, NB: 0.1, LR: 0.06) generate AUC achieves 0.9044 and F1-score 0.8837, which outperforms the best result RF or other traditional classifiers generate.
Table 1. Multiple Metrics On Gender-Based Playing Style Stereotype In Overwatch.

| Classifier | Accuracy | Precision | Recall | F1    | AUC     | FPR    | FNR    |
|------------|----------|-----------|--------|-------|---------|--------|--------|
| DT         | 0.8163   | 0.7176    | 0.7142 | 0.7692| 0.8198  | 0.1666 | 0.1935 |
| KNN        | 0.7755   | 0.5726    | 0.7058 | 0.6857| 0.7526  | 0.3333 | 0.1612 |
| RF         | 0.8571   | 0.6667    | 0.8666 | 0.7878| **0.8288** | 0.2777 | 0.0645 |
| NB         | 0.7959   | 0.5816    | 0.75   | 0.7058| 0.7688  | 0.3333 | 0.1290 |
| SVM        | 0.8163   | 0.5579    | 0.8461 | 0.7096| 0.7732  | 0.3888 | 0.0645 |
| LR         | 0.8571   | 0.6357    | 0.9230 | 0.7741| 0.8172  | 0.3333 | 0.0322 |
| MLP        | 0.7551   | 0.3963    | 0.8    | 0.5714| 0.6899  | 0.5555 | 0.0645 |
| **ELF**    | **0.9382** | **0.7971** | **0.95** | **0.8837** | **0.9044** | **0.1739** | **0.0172** |

5.2. The Performance of ELF

We observe that the Ensemble Late Fusion classifier has the best classification result at distinguishing between players who have stereotypes against female players and those who do not have stereotypes against female players as Table 1 shows, not only about high results on AUC and F1, also false-positive rate and false-negative rate as low as 0.1739 and 0.0172, respectively. Through the best weights combination, we find that Random Forest contributes the most. RF is trained based on several decision trees, and the reason for the good performance is that RF algorithm selects a subset of the training data, and each of the different sub-tree can output a binary prediction (the player has stereotype toward female players in Overwatch or not) and then by a majority of votes in all the prediction results, a final decision was made. Because RF considers many different subsets of the training data, it avoids potentially over-fitting due to the reduced dimension of the training data. Also unsurprisingly, it outperforms the decision tree classifier, which is trained only once on all training data.

5.3. Recognizing Top Features

We also examined and highlighted top features that identify players who have stereotypes against female players or not. Figure 5 shows the heat-map of the top 9 features which identify players who have stereotypes or not. Each feature number corresponds to a specific question in the survey we conducted (e.g., Feature 26 relates to the attitude of if the player likes to play with female players). For example, Figure 5 shows that the top 1 feature is the 31st feature, that is, Would you refuse or dislike to play with female players? The answer to this question has the highest correlation with whether there is a gender-based stereotype against female players in Overwatch. Obviously, if the respondent’s answer to this question is yes, it means that the player refuses or does not like teaming up with female players in the game, and it is likely that the player only likes to team up with male players in the game. This feature can reveal the fact that the players who have stereotype toward female players believe female players may drag down the performance of the entire team, or do not like the way the female players behave in the game. The second top feature is feature 25, Do you team up with males or females more? Similarly, if the answer is yes to male players, this type of respondent will have the subjective impression that female players cannot play the game better than male players, ignoring the fact that there are still top female players who have high ranking as male players in Overwatch.
Figure 5. The heat-map of top features toward gender-based stereotype in Overwatch Dataset

The details of some top features and the percentage of answers are listed as follows:

- Feature 31: What is the female player percentage of the game in your opinion? (0 - 20% female: 31.9% participates; 20 - 40% female: 29.5% participates; 40 - 60% female: 23.3% participates; 60 - 80% female: 15.3% participates; 80 - 100% female: 0% participate)
- Feature 3: In your opinion, which gender is better at the game? (82% male; 18% female)
- Feature 20: Do you think Overwatch has a high learning curve for female players? (77% yes; 23% no)
- Feature 24: Do you think the negative environment (from teammates) may discourage female players to play the game? (69% yes; 31% no)
- Feature 29: Do you think there are more male players in damage role? (92% yes; 8% no)
- Feature 12: Do you think there are more female players in the tank or support role? (76% yes; 24% no)
- Feature 28: If you choose to play the Damage role, it is most likely because? (Damage role is more important: 34.6%; Enjoy dealing damage: 30.2%; To fill the empty role in a team: 26.3%; To get In-game Reward: 8.9%)
- Feature 25: Do you team up with male or female more? (81% yes; 19% no)
- Feature 26: Would you refuse or dislike to play with female players? (62% yes; 38% no)

6. Conclusion

Based on a blended method and applied statistical and machine learning methods, this article examines Overwatch’s gender stereotypes and how these stereotypes affect female player’s choices and playstyle preferences. Our findings show that gamers prefer to select characters in the game whose gender is compatible with their real gender, and gamers tend to choose “supporting” roles during MOBA play.

Our study also noticed that male players are more belligerent, violent, and challenged with their enemies, regardless of the perceived preference between female and male players, whereas female players are more cautious, evasive, and conservative when they are playing the same positional positions. Our findings indicate that male players perform better on deaths, injuries, murders, etc. In practice, while female players perform better in all positions in terms of treatment and healing. This is
in line with recent research that has shown male to be more competitive than female. The findings of our interview also reveal that some players tend to play “Support” to avoid conflict with their rivals. This result is consistent with previous research in which females view their abilities as worse than males due to gender stereotypes, even though they have the same abilities. Players often play “Duo-Support” to avoid being mocked or insulted. Owing to the hostile play climate in the League, female players are more likely to be targeted for insults and assaults due to gender stereotypes.

Finally, we can conclude with the following statements after analyzing the collected questionnaire:

- Male is more competitive than female, and female players often underestimate their skills.
- The players are also more conservative and less confrontational.
- Female players believe “Support” to be the easiest position to play, and they prefer to play “Support” because they are scared of being blamed for not doing well.
- Females are kind and excellent at healing others, which has turned out to be a typical gender stereotype in games.
- The learning curve of Overwatch is steep.
- A negative feedback loop that further discourages females from studying and exploring in Overwatch is created by the steep learning curve and hostile environment.
- Male players are more experienced when playing and willing to devote more time to the game.
- Female players may receive more pressure and criticism from male players who have more experience in Overwatch, forcing females to take on simple roles to avoid guilt.
- Female players prefer to play alongside their romantic partners and play more loving roles while they are with their romantic partners.
- The female players felt more secure and better at supporting their teammates.
- Long-standing involvement in the game may further discourage women from playing in Overwatch.
- Fewer female characters are available than male characters, and this drastically restricts the opportunity to learn and explore different roles for players who choose to play female characters.

While the percentage of female players in Overwatch grew, in competitive games, male players remained the majority population. Further study activities will concentrate on how we can promote female player’s engagement in professional games and create a safe gaming climate for them. Only by getting rid of traditional game discourse can we arouse the attention of the entire game society to female players. To eliminate gender stereotypes in games, male game designers and male players must first change the perspective of objectifying female players and show sufficient respect for them. More importantly, female players also need to jump out of the gender myths woven by current games, improve their gender awareness, actively participate in game design, and develop critical reflections on current game works. Only in this way, female players and designers can confidently write their own stories in the game world.

Acknowledgments

We thank the anonymous respondents for their contributions and answers to our datasets toward gender stereotypes in Overwatch. We additionally thank ViaX for excellent suggestions and comments on our manuscripts.

References

[1] How the Overwatch Cities & Countries Album came to be: https://www.redbull.com/za-en/overwatch-cities-and-countries-album
[2] MOBA: https://en.wikipedia.org/wiki/Multiplayer_online_battle_arena
[3] Overwatch: https://en.wikipedia.org/wiki/Overwatch_(video_game)
[4] Overwatch League: https://overwatchleague.com/zh-cn/
[5] Overwatch Reddit: https://www.reddit.com/r/Overwatch/
[6] Overwatch Roles: https://overwatch.gamepedia.com/Roles
[7] The status of discrimination against women: http://games.ifeng.com/special/female/
[8] Myat Aung, Valerio Bonometti, Anders Drachen, P Cowling, Athanasios Vasileios Kokkinakis, Christian Yoder, and A Wade. Predicting skill learning in a large, longitudinal moba dataset. In 2018 IEEE Conference on Computational Intelligence and Games (CIG), pages 1–7. IEEE, 2018.
[9] Kelly Bergstrom, Jennifer Jenson, and Suzanne de Castell. What’s’ choice’ got to do with it? avatar selection differences between novice and expert players of world of warcraft and rift. In proceedings of the International Conference on the Foundations of Digital Games, pages 97–104, 2012.
[10] James L Hilton and William Von Hippel. Stereotypes. Annual review of psychology, 47(1):237–271, 1996.
[11] Rosa Mikeal Martey, Jennifer Stromer-Galley, Jaime Banks, Jing Si Wu, and Mia Consalvo. The strategic female: gender-switching and player behavior in online games. Information, Communication & Society, 17(3):286–300, 2014.
[12] Yi Mou and Wei Peng. Gender and racial stereotypes in popular videogames. In Handbook of research on effective electronic gaming in education, pages 922–937. IGI Global, 2009.
[13] Supun Nakandala, Giovanni Luca Ciampaglia, Norman Makoto Su, and Yong-Yeol Ahn. Gendered conversation in a social game-streaming platform. arXiv preprint arXiv:1611.06459, 2016.
[14] Ji Hoon Park, Nadine G Gabbadon, and Ariel R Chernin. Naturalizing racial differences through comedy: Asian, black, and white views on racial stereotypes in rush hour 2. Journal of Communication, 56(1):157–177, 2006.
[15] Aashish Prasad. Estimating video game success using machine learning. 2019.
[16] Rabindra A Ratan, Nicholas Taylor, Jameson Hogan, Tracy Kennedy, and Dmitri Williams. Stand by your man: An examination of gender disparity in league of legends. Games and culture, 10(5):438–462, 2015.
[17] F. Safara, A. S. Mohammed, M. Yousif Potrus, S. Ali, Q. T. Tho, A. Souri, F. Janenia, and M. Hosseinzadeh. An author gender detection method using whale optimization algorithm and artificial neural network. IEEE Access, 8:48428–48437, 2020.
[18] Aleksandr Semenov, Peter Romov, Sergey Korolev, Daniil Yashkov, and Kirill Neklyudov. Performance of machine learning algorithms in predicting game outcome from drafts in dota 2. In International Conference on Analysis of Images, Social Networks and Texts, pages 26–37. Springer, 2016.
[19] Robin Smit. A machine learning approach for recommending items in league of legends. 2019.
[20] Teresa L Thompson and Eugenia Zerbinos. Gender roles in animated cartoons: Has the picture changed in 20 years? Sex roles, 32(9-10):651–673, 1995.
[21] Weiqi Wang. Predicting multiplayer online battle arena (moba) game outcome based on hero draft data. PhD thesis, Dublin, National College of Ireland, 2016.