Can Deep Learning Predict Problematic Gaming?

Qirui Wu
Dept. Computing and Software
McMaster University
Hamilton, Canada
wuq43@mcmaster.ca

Jacques Carette
Dept. Computing and Software
McMaster University
Hamilton, Canada
carette@mcmaster.ca
ORCID 0000-0001-8993-9804

***Abstract*—How does one build a healthy gaming ecosystem? Recent evidence clearly demonstrates the existence of problematic gaming [1]. Predicting problematic gaming is still in its infancy. Here we focus on excessive gaming and model in-game behaviour as a means to continuously predict future play time. This can be used to help players maintain a healthy balance between the virtual and real worlds. To do this, we convert game log data into time-series and label such data with criteria of problematic gaming. Deep learning is then used to solve the resulting multi-class classification problem.**

***Index Terms*—Modelling in-game behavior, Problematic gaming, Deep Learning, Deep Auto Encoder (DAE), Long short term memory (LSTM), Game log**

I. Introduction

Video games occupy an increasing proportion of some players' time, sometimes to the detriment of their physical and psychological health. Finding means to recognize abnormal in-game behavior can help mitigate the harm of problematic gaming. Current studies of problematic gaming are mostly conducted using small-scale surveys, which does not scale. The World Health Organization (WHO) [2] classifies gaming disorder as an international disease and recommends all that gamers be aware of their play time. Thus large-scale prediction and early warning are important.

Using established means of assessing Internet Gaming Disorder (the Weekly Gameplay - WG), we create an effective prediction model, using currently available data. We do this by analyzing game log data of ***League of Legends (LOL)***, which is available publicly. We first describe how problematic gaming could manifest itself in such log data. Then we use deep learning algorithms to analyze in-game behaviour. The dataset, code and configurations for our work can be found online [1]. Finally, we evaluate our model quantitatively against the established assessment of problematic gaming (the Game Addiction Scale - GAS) through a player survey.

II. Related Work

A. History of Problematic Gaming

Atari Home Pong and ***Magnavox Odyssey*** in the 1970s heralded the first generation of video games targeted at a wide audience. As console games became more prevalent in the 1990s, numerous reports of excessive gaming raised awareness of the hazards of problematic gaming. Gaming addiction then began to be assessed using pathological gambling criteria based on the Diagnostic and Statistical Manual of Mental Disorders (DSM-III or DSM-IV), from the American Psychiatric Association (APA) [4]. In early 2000, the rise of online and mobile games led to an explosion in cases of gaming disorder; a 2009 investigation showed that at least 3 million teenagers were highly dependent on video games [5]. In 2013, game addiction was added to DSM-V as a separate category of mental disorders [6]. By 2018 showed that there were 2.3 billion mobile gamers worldwide [7], and rising. Also in 2018, the WHO added "gaming disorder" into its *International Classification of Diseases* [2].

Criterion	Description
Salience	Gaming becomes the highest priority
Tolerance	Play time starts to increase gradually
Mood Modification	Emotion changes, such as releasing anger or stress, when gaming
Withdrawal	Unpleasant emotion associated to reducing play time
Relapse	Strong tendency to return to excessive play after abstinence
Conflict	Arising interpersonal conflicts owing to gaming
Problems	Mental, physical, or behavioral problems caused by excessive gaming

TABLE I
Brief Descriptions of 7-item Game Addiction Scale [3]

B. Techniques for Recognizing Problematic Gaming

Identifying problematic gaming is currently done by conducting a survey such as GAS [3] for distinguishing Internet gaming disorder (IGD) [8]. The reliability of these methods has been verified [9], [10]. Gaming disorder has also been subject to less subjective means, such as a classification model based on electroencephalographs (EEG) [11]. A wearable mobile EEG device logs the frequency attributes of the players' brain waves. After the experimenters labelled abnormal gaming behavior, a logistic regression is used to solve the resulting binary classification problem. Genetic testing was also used to try to estimate the probability of developing gaming addiction [12].

C. Techniques Chosen

We use the principle of *Weekly Gameplay* (WG) [13], as developed by the APA in determining IGD. It defines six levels, defined by time bounds, of 7, 14, 20, 30 and 40 hours

1. https://github.com/LelouchWu/Qirui

978-1-7281-4533-4/20/$31.00 ©2020 IEEE

of play per week (seven consecutive days). If the total play time is less than 7 hours over seven days, this is judged as level 0; 8-14 is level 1, and so on to level 5 for more than 40 hours of play per week. To further validate our model, we conducted a quantitative analysis among college students (gamers). The assessment result of GAS is applied to compare with our predictions. All criteria of GAS, described in Table 1, were assessed on a 5-point scale from 1 ("never") to 5 ("very often"). If 4 criteria are validated ($\geq$3 ("sometimes")), problematic gaming can be identified [3].

III. MODELLING

A. Dataset

We use data from the popular game ***LOL*** released by ***Riot Games*** in October 2009 for both macOS and Windows. It is a multiplayer online battle arena (MOBA) and free-to-play (FTP) game. We used a public API to access the database of log data available at ***www.op.gg***. To obtain our data set of players ranked according to the nine levels of the LOL ranking system, from "Iron" to "Challenger", we used stratified random sampling to randomly extract 1,614 players in each rank. Based on this user list, we collected corresponding log data of each player (14,526 of them) with timestamps from Apr 1, 2019 to Aug 19, 2019.

B. Data Pre-processing and Labelling

Since gamers' behavioral patterns are directly reflected in their daily routine and habits, we need time-series data. As we wish to use a recurrent neural network (RNN) as our prediction model, we need to have adequate data. Although we could have used a regression model, we prefer to learn a prediction model that establishes a connection between game log data and the psychological model. The advantage of this is that the relationships between classes could be analyzed by the confusion matrix method. Such analysis is particularly important for models that use DAE. Thus we need to pre-process and label the log data.

Most games log player's in-game behavior in similar ways. A User ID and timestamp, followed by "lines" that describes information about that player at that particular moment. WG only uses play time to evaluate and predict excessive gaming. We pre-process the data as follows: make the User ID as index, and separate time into discrete periods, that then become "columns". Ultimately, we want a time chain (TMC) labelled by hours that covers all players' in-game behaviour. Thus we aggregate the data, as it is represented too finely in the logs. For each cell of the resulting TMC, the portion of time that exceeds 24 hours will be passed to the subsequent cells in chronological order. We want to continually predict every seven-day play time (SDPT) to push notifications everyday for warning players of potential problems. Thus we use a rolling time window with size = 7 days and step = 1 day in each player's TMC to sum every SDPT. The tags from 0 to 5 are applied to label the corresponding SDPT based on WG.

Algorithm 1: Computing Observation Period (OP)

Input:
1.Set **Seven-day Play Time(SDPT)** $= [T_1, T_2, \ldots, T_n]$
2.$\forall T_i \in$ **SDPT**, $T_i = [t_1, t_2, \ldots, t_m]$
3.Set the range of OP $= [LB, UB]$
4.$i \in [1, n]$, $j \in [1, m]$
Output: List of OP (OPs)

Set **SDPT′**,**OPs** to empty list
for *each* T_i *in the set of* ***SDPT*** **do**
  Set T_i' to empty list
  for *each* t_j *in the set of* T_i **do**
    if $t_j > 7$ *hours (based on the first level of WG)* **then**
      T_i'.Append(Abnormal Play)
    else if $t_j > 0$ **then**
      T_i'.Append(Normal Play)
    else
      T_i'.Append(Null)
  SDPT′.Append(T_i')
for *each* T_i' *in the set of* ***SDPT′*** **do**
  for *each* t_j *in the set of* T_i' **do**
    while $t_j \neq$*Null* **do**
      $Starting = j$
    Break
  for *each* t_j *in the set of* T_i' **do**
    while t_j = *Abnormal Play* **do**
      if $j - Starting > UB$ **then**
        OPs.Append(UB)
        Break
      else if $j - Starting > UL$ **then**
        OPs.Append($j - Starting$)
        Break
      else
        $j = j + 1$
return OPs

C. Observation Period

The observation period (OP) is important for prediction: If it is too brief, neural networks will lose accuracy on the testing set; if too long, we may miss the best opportunity to warn gamers. Thus we calculate a duration based on characteristics of the training set, see Algorithm 1. T represents the SDPT of each player, t is weekly play time in every SDPT, n is the player number, and m is the week number in the TMC. We first set a reasonable value range for the observation period: lower bound of 7 days, upper bound of 28 days. This range is then used to extract a duration required for the development of the first excessive gaming behavior within each player's SDPT time distribution (if it exists). A frequency histogram

over OP is used to establish that the average duration is approximately 14 days, giving us the size of our rolling window. We convert every player's TMC into time-series data, using 14-day windows of 1-day intervals. The output will be a prediction of their future weekly play time labelled by WG.

D. Constructing Deep Neural Networks

Four deep neural networks, RNN (baseline), LSTM, Bidirectional LSTM (Bi-LSTM), and DAE-Bi-LSTM, were tested.

LSTM and Bi-LSTM were used to guard against the potential problem of vanishing gradient and exploding gradient of RNN, as they perform better for long sequences. LSTM needs to be provided with high correlation features and a sufficiently large sample. For each player, LSTM keeps only a window of log data. With the passage of time, the past data will be overwritten, whether or not there is in-game activity. Thus we had to compensate for the lack of sample capacity by crawling data from large numbers of players. To avoid the possibility of data irrelevance within the same batch, we put each player's data in a specific mini-batch. To find the best network architecture, we tested the models by adding additional layers and batch normalization. Dropout was applied to avoid overfitting. Finally, SoftMax was selected as the activation function to calculate the predictive probability of classes.

DAE was introduced to mitigate the impact of uncertainties on log data, such as server maintenance. DAE works before Bi-LSTM to reconstruct the input through encoder and decoder. The encoder takes the original input and compresses it into a feature vector which is then reconstructed by the decoder to full dimensionality. As for the selection of activation function, we decided to use the standard Sigmoid function both on input layers and hidden layers by comparing the optimization effect of the average loss.

IV. Experimental Result

A. Modelling Experiment

The models were trained on over 1.7 million log entries from 12,526 players and tested on log entries from another 2000 players. From class 0 to class 5, the distribution of labels of the training set is 10%, 30%, 22%, 18%, 15%, and 5%. Labels in the testing set distribute as 11%, 33%, 20%, 15%, 14%, and 7%. Average cross-entropy (ACE) was used as the optimization parameter. ACE is a loss function commonly used in multi-class classification to assess optimization, which describes the distance between the output vector and the ideal vector. To represent the forecasting performance visually, overall accuracy (ACC) of six classes was applied. They can be calculated as

$$ACC = \frac{1}{N}\sum_{i=1}^{6}\sum_{Y(x)=i} \mathsf{Equal}(Y(x), \widehat{Y}(x)) \tag{1}$$

$$ACE = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{6} y_i \log \hat{y}_i \tag{2}$$

Days Ahead	RNN(Baseline)		LSTM		Bi-LSTM		DAE-Bi-LSTM	
	ACC	ACE	ACC	ACE	ACC	ACE	ACC	ACE
1	0.802	0.307	0.861	0.202	0.880	0.112	0.684	0.696
2	0.783	0.588	0.811	0.418	0.821	0.205	0.675	0.716
3	0.714	0.763	0.760	0.515	0.771	0.349	0.669	0.729
4	0.663	0.865	0.726	0.570	0.723	0.537	0.661	0.741
5	0.629	0.905	0.692	0.698	0.683	0.662	0.652	0.745
6	0.616	0.987	0.658	0.797	0.656	0.705	0.640	0.752
7	0.590	1.169	0.625	0.823	0.631	0.784	0.635	0.763

TABLE II
AVERAGE CROSS-ENTROPY (ACE) AND OVERALL ACCURACY (ACC) OF TESTING SET OF FOUR NETWORKS WITH THE CHANGE OF DAYS AHEAD

where Equal returns 1 if the classes match and 0 otherwise, and $\hat{y}_i$ represents the prediction probability for each class.

To verify the feasibility of predictions, we conducted seven experiments, as shown in Table II, fixing the number of days in advance as the independent variable, from 1 to 7, to seek which network has optimal performance for each test. RNN is considered as the baseline. Maximum ACC and minimum ACE on the testing set are the primary evaluation criteria and also the dependent variables. In theory, the correlation of data decreases with the increase of days in advance. More narrowly, when predictions are made one day ahead, networks only need to predict play time on the 7th day, as there is a six-day overlap between the target week and observation period. However, such overlap does not exist when predictions are made seven days or more ahead, and there is thus no partial correlation among features and labels. In the first three experiments, Bi-LSTM got the best optimization and accuracy, since it can more accurately find patterns in strongly correlated data. In experiment 4, 5, and 6, as correlation decreased, the advantages of Bi-LSTM gradually decreased and almost were overshadowed by LSTM. For DAE+Bi-LSTM, the ACC failed to get more than 0.69 in all tests, because DAE broke the correlation when rebuilt data. However, in the last experiment, the negative effect of DAE was non-existent, and the reconstructed data also obtained new patterns from compressed feature vectors, which further improved the ACC of Bi-LSTM. More details and hyperparameters for all models can be found at https://github.com/LelouchWu/Qirui.

The Confusion Matrix was used to analyze the accuracy on every class as shown in Figure 1. Due to space constraints, we only give the two most extreme experiments, 1 day and 7 days. From Figure 1(a), we can see that Bi-LSTM is struggling to discriminate between class 4 and 5. We believe this is caused by either very similar behavior patterns or the limited number of samples with weekly play time over 40 hours. In Figure 1(b), the same problem also arose for the DAE-Bi-LSTM. Worse, because the proportion of class 1 was 0.302, after the processing of DAE, all inputs obtained some characteristics from class 1, driving to a decrease of predicted accuracy in other classes, especially for class 0 whose sample may contain a large number of zero data points. However, it is worthy to sacrifice the characteristics of minority classes, since DAE-Bi-LSTM did improve the overall accuracy in the 7th test.

Fig (a) : Prediction of 1 day ahead based on Bi-LSTM

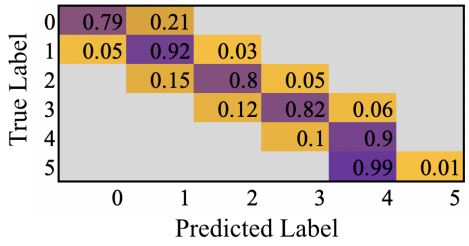

Fig (b) : Prediction of 7 days ahead based on DAE-Bi-LSTM

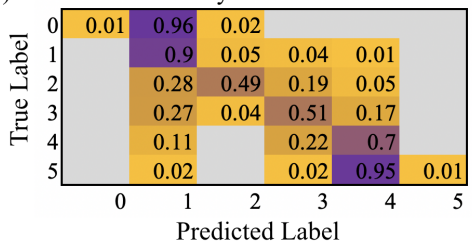

Fig. 1. Confusion Matrix of using Bi-LSTM in experiment 1(a) and DAE-Bi-LSTM in experiment 7(b).

B. Quantitative Research

On May 12, 2020, we conducted an experiment with 115 college students (%56 male), age between 17 and 22 (Mean=19.1, SD=1.69), and filtered out those who did not play LOL from Apr 28 to May 4. We obtained 26 LOL players (%62 male) with age between 18 and 21 (Mean=19.6, SD=1,15). All subjects indicated willingness to participate in the survey and read the related content before taking the survey.

First, we asked each participant to fill the 7-item GAS survey. Second, we used WG to label each player's actual weekly gameplay (AWG) from May 5 to May 11. Third, DAE-Bi-LSTM and players' log data, from Apr 28 to May 4, were then used to predict weekly gameplay (PWG) from May 5 to May 11. Based on [13], problematic gaming of WG can be distinguished by whether the weekly gameplay is greater than 30 hours (class 4 and 5 in our model). While play time is not the only factor that defines problematic gaming, it is a strong correlate and most easily observable.

Fisher's exact test (FET) was applied to analyze the contingency tables shown in Figure 2. Results showed positive problematic gaming in 35% of players from GAS, 27% from AWG, and 27% from PWG, differences that were both statistically significant ($P_A = 0.002, \quad P_B = 0.028, \quad \alpha = 0.05$, FET). Therefore, Study A showed that WG could be used to assess problematic gaming of college students, and Study B proved that our model could also be applied to make a seven-day ahead prediction in student-player community.

V. Conclusion And Future Work

We built a deep learning model to predict excessive gaming. Log data from ***LOL*** gamers was pre-processed and labelled using the WG criteria to model future play time. A quantitative analysis was conducted among student gamers to assess the positive correlation between the validated GAS and predicted results of DAE-Bi-LSTM, which were verified by FET.

Study A: GAS and AWG			Study B: GAS and PWG		
	Positive-AWG	Negative-AWG		Positive-PWG	Negative-PWG
Positive-GAS	6	3	Positive-GAS	5	4
Negative-GAS	1	16	Negative-GAS	2	15

Fig. 2. Contingency tables of assessing problematic gaming among 26 student gamers. Study A: Game addiction scale (GAS) and Actual Weekly Gameplay (AWG). Study B: GAS and Predicted Weekly Gameplay (PWG).

Since the student-player community's behavior patterns are less complicated, the generality of this approach still needs to be investigated by large-scale quantitative experiments. Privacy policies make some valuable features inaccessible to us (from the public log data), such as age, gender, and occupation. With such features, conducting a clustering before prediction will make the results more generalizable and hopefully more precise. In the future, we would also like to work on pay-to-play games, where features such as the history of in-game payment may further improve accuracy.

References

[1] M. D. Griffiths, D. J. Kuss, O. Lopez-Fernandez, and H. M. Pontes, "Problematic gaming exists and is an example of disordered gaming: Commentary on: Scholars' open debate paper on the world health organization icd-11 gaming disorder proposal (aarseth et al.)," *Journal of Behavioral Addictions*, vol. 6, no. 3, pp. 296–301, 2017.

[2] W. H. Organization *et al.*, "Icd-11 for mortality and morbidity statistics," *Retrieved June*, vol. 22, p. 2018, 2018.

[3] J. S. Lemmens, P. M. Valkenburg, and J. Peter, "Development and validation of a game addiction scale for adolescents," *Media psychology*, vol. 12, no. 1, pp. 77–95, 2009.

[4] A. Diagnostic, "statistical manual of mental disorders. american psychiatric association," *Washington, DC*, vol. 886, 1994.

[5] R. Dodd and T. Faust. (2017) A new addiction on the rise: Mobile game addiction.

[6] A. P. Association *et al.*, *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.

[7] D. Lynkova. (2019) How many people play mobile games in 2020.

[8] J. P. Charlton and I. D. Danforth, "Distinguishing addiction and high engagement in the context of online game playing," *Computers in human behavior*, vol. 23, no. 3, pp. 1531–1548, 2007.

[9] B. S. Fabito, R. L. Rodriguez, M. A. Diloy, A. O. Trillanes, L. G. T. Macato, and M. V. Octaviano, "Exploring mobile game addiction, cyberbullying, and its effects on academic performance among tertiary students in one university in the philippines," in *TENCON 2018-2018 IEEE Region 10 Conference*. IEEE, 2018, pp. 1859–1864.

[10] Y. Khazaal, A. Chatton, S. Rothen, S. Achab, G. Thorens, D. Zullino, and G. Gmel, "Psychometric properties of the 7-item game addiction scale among french and german speaking adults," *BMC psychiatry*, vol. 16, no. 1, p. 132, 2016.

[11] M. Hafeez, M. D. Idrees, and J.-Y. Kim, "Development of a diagnostic algorithm to identify psycho-physiological game addiction attributes using statistical parameters," *IEEE Access*, vol. 5, pp. 22 443–22 452, 2017.

[12] D. H. Han, Y. S. Lee, K. C. Yang, E. Y. Kim, I. K. Lyoo, and P. F. Renshaw, "Dopamine genes and reward dependence in adolescents with excessive internet video game play," *Journal of addiction medicine*, vol. 1, no. 3, pp. 133–138, 2007.

[13] H. M. Pontes, O. Kiraly, Z. Demetrovics, and M. D. Griffiths, "The conceptualisation and measurement of dsm-5 internet gaming disorder: The development of the igd-20 test," *PloS one*, vol. 9, no. 10, 2014.