Moment-to-moment Engagement Prediction through the Eyes of the Observer: PUBG Streaming on Twitch

David Melhart*
Daniele Gravina*
Georgios N. Yannakakis
david@modl.ai
daniele@modl.ai
georgios@modl.ai
modl.ai

ABSTRACT

Is it possible to predict moment-to-moment gameplay engagement based solely on game telemetry? Can we reveal engaging moments of gameplay by observing the way the viewers of the game behave? To address these questions in this paper, we reframe the way gameplay engagement is defined and we view it, instead, through the eyes of a game's live audience. We build prediction models for viewers' engagement based on data collected from the popular battle royale game *PlayerUnknown's Battlegrounds* as obtained from the *Twitch* streaming service. In particular, we collect viewers' chat logs and in-game telemetry data from several hundred matches of five popular streamers (containing over 100, 000 game events) and machine learn the mapping between gameplay and viewer chat frequency during play, using small neural network architectures. Our key findings showcase that engagement models trained solely on 40 gameplay features can reach accuracies of up to 80% on average and 84% at best. Our models are scalable and generalisable as they perform equally well within- and across-streamers, as well as across streamer play styles.

CCS CONCEPTS

• **Applied computing** → **Computer games**; • **Human-centered computing** → *User models.*

KEYWORDS

Machine learning, artificial neural networks, engagement, viewer modelling, streaming, battle royale games, PUBG

ACM Reference Format:
David Melhart, Daniele Gravina, and Georgios N. Yannakakis. 2020. Moment-to-moment Engagement Prediction through the Eyes of the Observer: PUBG Streaming on Twitch. In *International Conference on the Foundations of Digital Games (FDG '20), September 15–18, 2020, Bugibba, Malta.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3402942.3402958

*Both authors contributed equally to this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
FDG '20, September 15–18, 2020, Bugibba, Malta
© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-8807-8/20/09…$15.00
https://doi.org/10.1145/3402942.3402958

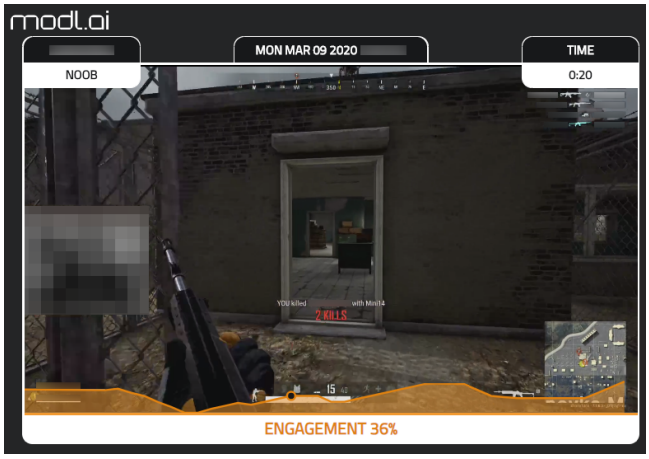

Figure 1: Moment-to-moment engagement prediction of a live PUBG streamer. In the middle of the screen a replay of the game can be seen. The streamer's face is on the left (blurred out to preserve anonymity). Engagement is illustrated as a continuous orange line at the bottom of the video; its current value is displayed underneath. The streamer's name (also blurred in this example) and play style (i.e., *Noob* in this example) are shown at the top left of the dashboard. See Section 4.3 for more details regarding the methods used to identify PUBG play styles. The video on the screenshot is obtained through the Twitch Developer API (fair use), PUBG is a registered trademark of PUBG CORPORATION.

1 INTRODUCTION

The reliable estimation of the moment-to-moment gameplay engagement is arguably of utmost value for game development. Accurate proxies of engagement may not only enhance a game's monetisation strategy, they can also be used for rapidly testing games through artificial game-playing agents that are equipped with such engagement estimates. Artificial intelligence algorithms that are driven by reliable engagement models can, in turn, improve aspects of player experience and lead to the design of entirely new and engaging gameplay via game content generation [48].

With the advent of streaming services and the growing popularity of electronic sport competitions, the engagement of game spectators became increasingly important; if not as important as

the engagement of players themselves. This is evidenced by the exponential growth of game streaming services such as *Twitch*[1] and *Mixer*[2] and the audience of games like *PlayerUnknown's Battlegrounds* (PUBG Corporation, 2017) or *Fortnite* (Epic Games, 2017) in recent years. Unconventionally in this study, instead of looking at the player's behaviour as a predictor of engagement we *reframe* the modelling problem and look at gameplay engagement from the viewers' perspective. We define *gameplay* as the state of a game that is experienced and *engagement* as the active participation of viewers of gameplay. We thus assume there is an unknown mapping between gameplay—that is live streamed to viewers—and the engagement of the audience of that game.

To test this hypothesis we utilise the popular video live streaming service *Twitch* and obtain data from 5 popular streamers of the game *PlayerUnknown's Battlegrounds—PUBG* (PUBG Corporation, 2017) as streamed on *Twitch*. Importantly, three of those streamers are currently ranked within the top 10 PUBG streamer list—rank 1, 2 and 7—in terms of viewership. To construct models of moment-to-moment gameplay engagement in this initial study we investigate the relationship between critical events of the game and the corresponding frequency of messages in the chat feed. In particular, we use artificial neural networks that are able to predict gameplay engagement (as attributed to the viewers' chat frequency) at each critical event in the game (e.g., player death, head-shot, kill etc.). The derived models reach accuracies of up to 80% on average and 84% at best suggesting that gameplay events can form accurate predictors of viewer engagement and that viewer behaviour (through the frequency of chatting) can be attributed to gameplay. Our models are able to predict engagement within and across the five different streamers with similarly high accuracies showcasing the scalability and generalisability of the approach. Moreover, the models can accurately predict engagement—with accuracies up to 75-80% on average—across three different PUBG play styles (*Noob*, *Explorer* and *Pro*) which are identified through data clustering methods. The outcome of this work is a continuous prediction of engagement (engagement line) and play style for any given live PUBG video that is streamed (see Figure 1).

This paper is novel in several ways. First it approaches gameplay engagement from a third-person (viewer) rather than a first-person (player) perspective, as normally done in player modelling studies [47, 48]. Second, it introduces a continuous moment-to-moment predictor of engagement in games with a particular application to a popular live streamed game. Finally, the engagement models obtained are highly accurate and general within and across streamers indicating that the function between viewer engagement and gameplay can be learned accurately. Before delving into the details of our methods, the data we solicited, and the key results we obtained, in the next section we elaborate further on our definition of gameplay engagement and review the literature on predictive models of engagement.

2 ENGAGEMENT

Engagement is a popular yet ambiguous term used in user experience design and research to describe a continuous interest and interaction. The concept of engagement generally encompasses both cognitive and affective processes and is widely associated with attention, arousal [25], information interaction [40], the flow state [9], aesthetics [18], novelty, and challenge [30]. In the remainder of this section we first review the relationship between viewing behaviour and engagement, we then move onto surveying the links between chat messaging and engagement, and finally we cover core aspects of engagement prediction.

2.1 Gameplay Engagement via Viewers

Can viewer behaviour reveal anything about gameplay engagement? According to Yee [49], the main factors that motivate online gameplay are *immersion, social interest,* and *achievement*. While these factors attempt to describe why people play games, they can also inform us why people watch others playing. Contemporary studies of Sjöblom et al. [36, 37], for instance, reveal similar motivations for watching game streams in the form of *affective, social,* and *tension release* needs of viewers.

In this paper we assume there is function between the gameplay state and the engagement of the viewers of that game and we define *engagement* as the active participation of viewers of gameplay. The theoretical grounding of this assumption builds on the *theory of mind* [8] pointing to our cognitive ability to attribute mental states to ourselves and to others and feel how they might feel. The relationship between a player's and a viewer's engagement has also been described in player experience frameworks such as those of Lazzaro [19]—the *people* factor of "fun"—or the player archetype taxonomy by Bartle [4]—the *socialiser* archetype. In practical terms, this attribution of gameplay engagement to viewers can be seen as a form of *third-person* annotation which is the dominant practice for obtaining reliable labels of ground truth in affective computing [6]. Given the above, the underlying hypothesis explored in this study focuses on people engaging with a game as *viewers* instead of players.

2.2 From Chat Messages to Engagement

Although playing is a generally more interactive activity compared to spectating, online viewers are not entirely passive [30]. The participatory communities on Twitch streams, for instance, encourage social engagement while qualitative studies reveal a connection between interaction and pivotal points of streams both in terms of novelty and emotion [17]. While spectators react to the stream content often in an emotionally charged manner—producing rapid, unique patterns of crowd communication [16, 29]—their engagement is entangled with the streamer's focus and the para-social nature of the streamer–spectator relationship [44]. Beyond the suspense of the streamed game content, however, spectators may also engage with the streamer and their online personality.

Intuitively it seems appropriate to associate viewer engagement to high frequencies of chatting behaviour. Recent evidence, however, suggests that higher message frequencies might not always correspond to more engaging content [29]. One explanation of this phenomenon lies within the dynamics between the streamer and the spectators as the continuous interaction of viewers is, in part, mediated by para-social interactions with the streamer [44].

[1] https://www.twitch.tv
[2] https://mixer.com/

Consequently, when the streamer's attention is directed to the immediate gameplay, spectators lose a point of interaction, causing a drop in their message frequency until a cathartic point is reached, prompting an emotional response. Another explanation lies within the ways viewers manage the incongruity between the novelty of the stimuli and their internal mental models [33]. Encountering a boring segment, the viewers' engagement with the video drops; to maintain the level of stimuli, however, the engagement with the chat increases. This can explain the drive behind "spamming" behaviour and emoji cascades [3] which spike during frustrating or boring sections of streams [29].

Based on the aforementioned studies we argue that the design of a measurement (or a proxy) of engagement that attributes engagement to higher frequencies of viewer-player interactions can be misleading. In particular, given the dynamics of the examined game—which include long stretches of low tension gameplay (see Section 3.1 for more details)—in this paper we approximate spectator engagement with the gameplay content as a *function inverse to the viewers' chat message frequency.*

2.3 Continuous Engagement Prediction

As mentioned earlier, watching streamed gameplay content involves active participation in the form of submitting video recommendations and posting comments. Traditional data analytics methods rely on these metrics—in addition to passive viewership numbers—to calculate the engagement of videos [12, 21] in terms of dropout, re-engagement, and engagement levels. Dropout and re-engagement can generally be measured and predicted similarly to churn and rely on the user's interaction with a whole platform rather than the streamed content [11, 43]. Predicting the engagement level of the video per se, however, often relies on data specific to the streamed content. These metrics focus on the number of interactions during the video such as comments and chat messages [29]. Predictive modelling, in such cases, builds on the language and emotional content of the messages via natural language processing and sentiment analysis [3]. These approaches may also integrate qualitative analyses via visualisation methods [31] or statistical aggregations of chat logs.

Player profiling—a dominant practice in the games industry [14, 15]—relies mostly on basic statistical approaches and unsupervised learning methods that derive emergent patterns and distinct groups within the behaviour of players [5, 7, 13]. These methods integrate well into existing industry practices and provide valuable information for both designers and industry stakeholders about how people are interacting with their content in general. Increasingly larger games—with previously unseen amounts of content to stream—come with unique challenges, however, which clustering and profiling methods are unequipped to solve. A response to such growth in available content volume is dynamic player modelling. Player modelling relies on large amounts of data and models the behaviour and experience of players, thereby providing dynamic feedback beyond large-scale observations [47, 48]. In particular, player modelling methods that rely on various supervised learning techniques have already been applied successfully to predict churn [32, 41], player behaviour [2, 22], motivation [26], and experience [45, 46].

In contrast to the aforementioned studies on media and gameplay engagement, in this paper we focus on a time-continuous prediction of engagement. While traditional analytics focus on evaluating a piece of content (such as a gameplay stream) as a unit we, instead introduce a method for a fine-grained, moment-to-moment prediction of engagement. With this method it is possible to model the moment-to-moment change in engagement—not just highlighting more and less engaging sessions—but providing time-continuous feedback on how engagement changes within a game session. We also introduce a novel proxy for engagement in the form of reverse chat frequency that is used to generate a continuous trace of engagement labels for streamed content. While this ad-hoc metric has to be cross-verified against annotated engagement, our results showcase that features of gameplay content can predict such a measure with supreme levels of accuracy within and across streamers.

3 DATASET

This section outlines the dataset used in this study with an emphasis, on the one hand, on the particular game selected and the telemetry features considered (see Section 3.1) and, on the other hand, on the engagement annotations we obtained through Twitch (see Section 3.2).

3.1 PUBG & Extracted Features

In our attempt to predict engagement via streamed gameplay content we selected PUBG (PUBG Corporation, 2017) as the test-bed for all reported experiments. Our selection is based primarily on two core factors: a) the game's popularity and b) the availability of detailed streaming data.

PUBG (PUBG Corporation, 2017) heralded the rise of Battle-Royale style games and reached high levels of popularity with streamers, who broadcast their gameplay for a wide audience. PUBG is a multiplayer online shooter game, in which a group of players (up to 100 at a time) are dropped into a large open map and left to scavenge for weapons and items, eventually engaging each other in combat until only the winner remains; see Figure 2. The gameplay dynamic is characterised by long stretches of traversal and preparation which are inter-cut by fast bursts of action. As the game progresses, the playable area shrinks, forcing the remaining players closer together, increasing the likelihood of combat. If players remain outside the area of the playable radius they take constant damage; this area is refereed to as the *Blue Zone*. The shrinking of the *Safe Zone* encompassed by the *Blue Zone* is played out in phases. In each phase an *Evacuation Zone* is designated, outside of which players get a *warning* to evacuate the area. The *Blue Zone* then shrinks gradually the *Safe Zone* to the size of the *Evacuation Zone*. The pacing of the game is occasionally broken up by the bombardment of a random localised area, which is indicated by a *Red Zone* and forces players to take shelter inside buildings or evacuate the area.

PUBG Corporation provides an API and telemetry service[3], through which developers and researchers can generate dense datasets of gameplay telemetry. Each session is logged in detail in a hierarchical structure, organised by gameplay events and objects (such as players, pickups, vehicles, and weapons). There are 40

[3] https://documentation.pubg.com/

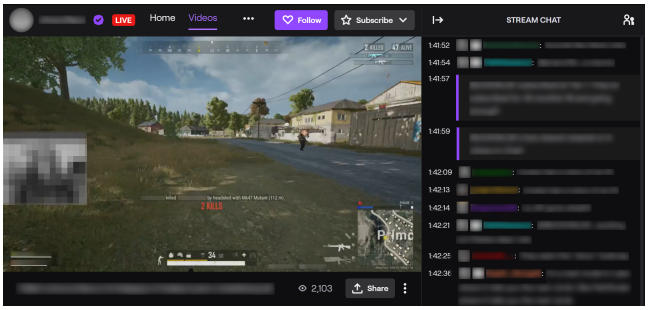

Figure 2: To the left: Gameplay stream; to the right: live chat feed. Screenshot of PUBG obtained from Twitch (fair use). PUBG is a registered trademark of PUBG CORPORATION. Identifying information of the streamer and the content of the chat window is blurred out to preserve anonymity.

gameplay events and 10 objects available through the API, which cover all players on the level and general game states as well. As this study focuses only on the streamer's content, data relating to other players (e.g., their position, actions, and combat periods which do not involve the streamer) is filtered out. This filtering is also necessary due to the unique structure of *battle royale*-style games. Given the generally large map sizes and the initial scattered distribution of players, one can spend long stretches of the game alone, scavenging for weapons and items, while other players are locked in a battle elsewhere. We only focus on the streamer in an effort to limit the noise of the dataset as most enemy action is completely hidden from the streamer (and their audience) and actual combat happens in short bursts.

Apart from the aforementioned filtering, we are making use of the full extent of the PUBG API, extracting all event based features without any hand-selection. In particular, we extract 40 PUBG gameplay features for the experiments reported in this paper. The features can be broken down to 5 main categories: *Health, Traversal, Combat, Item Use,* and *General Game State*. The *Health* category includes the streamer's *Health Level* and a number of boolean events: *Healing, Reviving, Receiving Revive, Armor Being Destroyed, Made Groggy, Taking Damage,* and *Being Killed*. The *Traversal* category includes the distance travelled since the last event (*Delta Location*), and the *In Blue Zone, In Red Zone, Swim Start, Swim End, Vault Start, Vehicle Ride, Vehicle Leave* boolean game events. The *Combat* category includes the *Shot Count, Damage Done* scalar values and the following boolean features: *Is Attacking, Weapon Fired, Caused Damage, Destroyed Object, Destroyed Armour, Destroyed Wheel, Destroyed Vehicle, Made Enemy Groggy*. The *Item Use* category keeps track of the *Item Drop, Item Equip, Item Unequip, Item Pickup, Item Pickup From Carepackage, Item Pickup From Lootbox, Item Use, Item Attach, Item Detach* boolean events. Finally, the *General Game State* category includes the *Elapsed Time* (in seconds), *Number of Alive Teams* and *Number of Alive Players* and the *Phase* of the game (i.e., *Blue or Red Zone*).

3.2 Twitch & Engagement

For the purposes of this study we obtained live PUBG gameplay data from Twitch; currently the largest streaming platform. Although Twitch is a general-purpose live-streaming platform, much of the site's traffic is generated by videogame streaming, both casual and competitive. As eSports and game streaming become more and more popular, the need for selecting more engaging streams, or parts of streams, rises. This is especially true to videogame streaming where fast rising trends can upend previously successful genres and new consumer darlings can skyrocket a company. While Twitch connects streamers with viewers, it also provides a platform for viewers to connect with each other. As it can be seen in Figure 2 chatting while watching streamers is a large part of the shared experience. Indeed, contemporary studies on the motivation behind Twitch viewership show that the strongest motivations are social, followed by affective and tension release needs [37]. While viewers do receive some level of gratification from watching streams and engaging with other viewers, cognitive (i.e., learning) and personal integrative (i.e., recognition by peers) needs are less pronounced [36] in the users of the platform.

As mentioned earlier in this paper we measure moment-to-moment engagement as the inverse frequency of chat messages in between two consecutive events of the game. This value is computed as the number of chat messages between two consecutive events and normalised between 0 and 1. To account for the reactionary nature of spectator chat, we look at the number of messages not congruently but following gameplay events (i.e., the number of messages between the observed event and the next event). It is important to note that our metric focuses on the *game content-related* engagement of *the spectators*, and not the player's engagement with the game. Following the study of Makantasis et al. [23] in this paper we view the prediction of engagement as a binary classification task, in which the objective is to predict "high" or "low" engagement labels. In particular, we consider *low* and *high* engaging those events with a message frequency higher and lower, respectively, than a selected threshold, α. While it might seem surprising to associate lower frequencies as moments of viewer's high engagement, by qualitatively inspecting the videos we observed that the chat room tends to be more quiet when fast-pace action is happening on the screen (i.e., viewers are paying more attention to the screen) and chat more when there are calmer slow-pace moments (e.g., as a manifestation of boredom).

3.3 Streamer Data Collection & Preprocessing

To test to which degree we can predict the PUBG engagement through telemetry events, we solicit in-game events and corresponding chat messages from the PUBG API and Twitch API, respectively, from 23 August 2019 to 12 January 2020. In particular we collected data from five anonymised streamers—referred in this paper as *A*, *B*, *C*, *D* and *E*—based on their popularity and the availability of datasets which are large enough to be explored through machine learning. Table 1 presents the streamers' ranking [4], the number of videos and matches collected, the average number of viewers [5], the average duration, the number of chat messages, and the number of events collected within the selected timeframe, for each of the five streamers. Based on these statistics we can observe directly that

[4]Ranked by the total viewership hours (hours live × average viewers) obtained obtained at the time of writing from https://www.twitchmetrics.net/. Only English language speakers are considered in this ranking.
[5]Live value; last accessed at the time of writing.

Table 1: Rank, number of videos, number of matches, average number of views (per video), average match duration (in seconds, per match), average number of chat messages (per match), and number of events (per match) across the five selected streamers. Standard deviation is shown in parentheses.

Streamer	Rank	# Videos	# Matches	# Viewers	Duration	# Chat	# Events
A	1	8	74	3789.6 (195.2)	478.2 (516.9)	279.7 (363.5)	290.8 (298.6)
B	2	2	48	2150.0 (861.3)	636.4 (572.5)	261.9 (329.8)	456.2 (392.3)
C	7	3	89	460.6 (35.3)	512.7 (558.9)	91.6 (104.2)	382.6 (372.8)
D	14	3	34	893.0 (342.5)	813.6 (563.4)	129.1 (139.6)	387.6 (289.3)
E	N/A	3	79	1175.3 (334.4)	429.9 (402.6)	92.8 (110.9)	369.9 (359.4)
Average	6.0 (5.9)	3.8 (2.4)	64.8 (22.9)	1693.8 (1325.8)	574.2 (154.1)	171.0 (92.5)	377.4 (58.9)

the two top ranked streamers, *A* and *B*, have a substantially higher number of viewers and chat messages per match compared to the other three streamers, who have comparable numbers among them. An interesting exception to this popularity ranking is the average match duration of *D* who seems to be playing roughly two times longer than the other streamers.

After the extraction and preprocessing of the input features (see Section 3.1) and the transformation of the message frequencies into binary labels (see Section 3.2), we obtain a total of 119,345 labelled events. Independently of the class splitting threshold (α) value chosen, the dataset presents a highly unbalanced ratio between the two classes, with a majority of the labels being classified as *high* engagement. To balance the dataset, we oversample—by randomly sampling the available samples with replacement—and undersample—by randomly selecting a given number of samples—the minority and majority classes, respectively, resulting to baseline accuracies of 50%. We follow this process individually for the training and test sets so that we eliminate any data leakage. Ideally the over and undersampling method proposed could be isolated on the training set; doing so, however, would yield highly unbalanced test sets that would not ease the analysis in this study. As long as the data processing method we followed does not allow for data leakage between training and test partitions the accuracy of the models obtained is generalisable to potentially highly-unbalanced unseen data.

4 EXPERIMENTS

For all experiments included in this paper we employ artificial neural networks (ANNs) as our prediction models. We picked ANNs in this initial study because of their evidenced training efficiency in large-scale datasets compared to other machine learning approaches such as support vector machines [1]. More importantly for this work (and potential future studies), ANNs can be easily extended to fuse different input modalities (e.g., pixel and telemetry data), which cannot be easily accomplished with other machine learning techniques [20]. A number of different ANN architectures and set of hyperparameters have been tested. In particular, we performed a sensitivity analysis across three different parameters: learning rate, number of hidden nodes and dropout rate. Based on that extensive parameter tuning process the ANNs we use feature a single fully-connected hidden layer composed of 128 nodes, followed by a dropout layer [38]; the network has an output node that predicts *high* (1) or *low* (0) engagement. All nodes use the ELU activation function [10], the learning rate equals 10^{-5}, and the ANN is trained for 100 epochs.

Table 2: Best configurations of α and ϵ values for each streamer (see Section 4.1) and cluster (see Section 4.3).

	Streamer					Cluster		
	A	B	C	D	E	Noob	Explorer	Pro
α	0.3	0.3	0.3	0.3	0.3	0.2	0.3	0.3
ϵ	0.0	0.05	0.05	0.02	0.08	0.00	0.05	0.08

In the first round of experiments (Section 4.1), we train and test our model individually on each of the five streamers. In the second set of experiments (Section 4.2), we test the scalability of our engagement models across all the streamers. In Section 4.3 we, instead, identify and model the different play styles across all streamers and finally in Section 4.4 we discuss qualitatively about the engagement lines produced from our models across two representative videos.

4.1 Individual Streamer Models

In this first set of experiments, we collect and machine learn data coming from each streamer individually. We validate our models using a 5-fold cross-validation scheme in which the matches are distributed randomly within the folds.

To assess which splitting criteria lead to the best model performances, we explore four different threshold α values (0.0, 0.1, 0.2, 0.3). Earlier work, however, suggests that this naive approach may lead to split criteria biases, as the model may learn to classify high and low engagement based on trivial differences in the frequency of the events [23, 24, 45, 46]. To address this challenge, we employ an *uncertainty bound* (ϵ) when we split the data so that we filter out any unambiguous datapoints close to the selected threshold value; in particular, we omit all the events that fall within the range $\alpha \pm \epsilon$. In addition to the four α values we explore three different values for $\epsilon = \{0.02, 0.05, 0.08\}$, we examine all the possible combinations of α and ϵ exhaustively, and we select the configuration with the highest 5-fold cross-validation accuracy. Table 2 shows the setup selected for each streamer.

Figure 3 shows the performances achieved for each streamer for the selected hyper-parameters. All individual streamer models

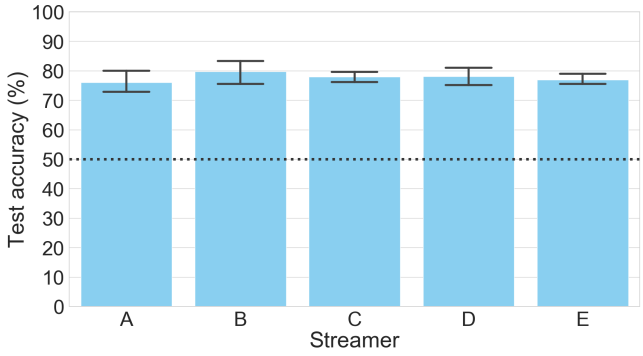

Figure 3: Individual-streamer engagement models: Average 5-fold cross validation accuracies. Error bars denote 95% confidence intervals.

Table 3: Engagement models across all streamers: Average accuracies and their 95% confidence intervals for different splitting criteria (α) and uncertainty bound (ϵ) configurations. The highest accuracy appears in boldface.

	α			
ϵ	0.0	0.1	0.2	0.3
0.0	70.1% ±1.5%	70.2% ±3.2%	72.8% ±1.4%	**74.7%** ±3.6%
0.02	70.2% ±2.4%	71.1% ±3.6%	73.1% ±1.3%	70.1% ±11.8%
0.05	70.3% ±2.0%	71.9% ±3.4%	74.1% ±2.6%	71.2% ±11.8%
0.08	70.4% ±1.8%	73.4% ±3.3%	74.2% ±3.2%	71.2% ±11.9%

of engagement achieve similar performance which reaches 76% to 80% on average. In particular the best accuracies are observed for the streamers B (79.7% on average; 84.3% at best), D (78.0% on average; 82.4% at best), and C (77.8% on average; 80.43% at best) while slightly lower values are obtained with E (76.8% on average; 80.8% at best), and A (76.0% on average; 83.2% at best). These results already indicate that our method can capture the relationship between streamer telemetry and viewer engagement with a very high accuracy across four different streamers.

4.2 Models Across All Streamers

The findings of the previous set of experiments showcase that capturing the engagement of individual streamers is possible with a very high accuracy. In this section we examine to which degree the models can generalise further and capture the engagement values of unseen streamers. To test the models' generality we employ the demanding *leave-one-streamer-out* cross-validation scheme [23], in which we train our model based on the data collected from four streamers, and we test it against the remaining streamer. This process is repeated five times, one for each streamer, and the results are averaged. The results obtained with this validation method are a robust indicator of the generalisation capacity of the proposed method, as by subdividing the data into the five available streamers we cannot overfit to a particular streamer data distribution (given the different number of viewers for each streamer) and data leakage is avoided by design.

For all the reported experiments (Table 3), we select the best parameter setup based on an exhaustive search of all combinations of α and ϵ values as performed in Section 4.1. The best model we could find (74.7% on average; 78.7% at best) yields a lower accuracy compared to the accuracies of the models tested on the data of individual streamers. This is unsurprising as a model's generality within-streamer is far easier to achieve than a model's generality across-streamers.

4.3 Models of Streamer Play Styles

Given the results obtained in the first two rounds of experiments it becomes apparent that a general model of engagement across streamers is a rather challenging task. Our hypothesis is that streamers depict varying (non-consistent) behaviours across the matches they play which, in turn, makes any attempt to model engagement across them very challenging for machine learning. We assume, instead, that there are general patterns of play across streamers that machine learning could capture and associate to engagement in an easier manner.

To investigate whether the five streamers show different play styles, we cluster the data collected. The raw data used in the moment-to-moment engagement prediction, however, is too sparse to extract any meaningful clusters. Therefore we first aggregate the 119, 345 events to 324 matches—i.e., we sum the boolean events (e.g., *Healing*) and we average the scalar values (e.g., *Delta Location*)—and for each match we normalise the data with min-max normalisation.

To determine the number of clusters present in the data we follow the approach proposed in [13], we employ two different clustering algorithms—k-means and hierarchical clustering [42]—and we test the consistency of their outcomes. We first apply k-means to the normalised data for k ranging from 1 to 10, and we compute the quantisation error—i.e., the sum of the distances of every data point to the corresponding cluster centroid. The results show that the percent decrease of the quantisation error when k increases is particularly high with two and three clusters, with a decrease of 53% and 20%, respectively. With higher values of k ($k \geq 4$) the difference is more contained (between 1% and 10%). Similar results are obtained with the *silhouette coefficient* method [34]. The silhouette coefficient (s) is equivalent to the difference of the mean intra-cluster distance and the mean nearest-cluster distance; higher silhouette coefficient values correspond to better defined clusters, bounded between 1 and −1. The results show that for $k = 2$, $k = 3$ and $k = 4$ we obtain the highest coefficients, with $s = 0.45$, $s = 0.28$ and $s = 0.26$ respectively; higher values of k, instead, produce lower silhouette coefficients, between $s = 0.18$ and $s = 0.2$.

An alternative approach to find the appropriate number of clusters is to partition the data in a hierarchical manner starting from every single match and then observe the relationship between the number of clusters and the corresponding squared Euclidean distance that separates those clusters. In our application of hierarchical clustering we use the *Ward* distance metric [42], which minimises the total within-cluster variance. This approach yields comparable results to k-means: the dendrogram of Figure 4 shows that a squared Euclidean distance threshold higher than 6.6 yields three clusters, while a threshold higher than 10.3 yields two clusters. The analysis

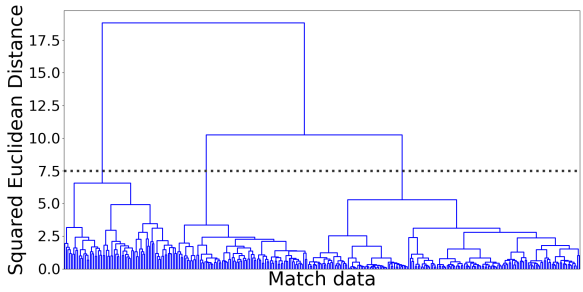

Figure 4: Dendrogram resulting from hierarchical clustering with the *Ward* method. Indicatively, a distance threshold of 7.5 yields 3 clusters.

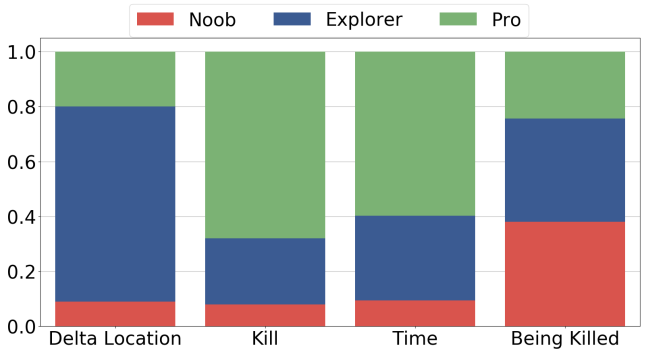

Figure 5: Normalised stacked bar chart: mean values of four representative features across the three play styles. *Noob*, *Explorer* and *Pro* are depicted in red, blue and green, respectively.

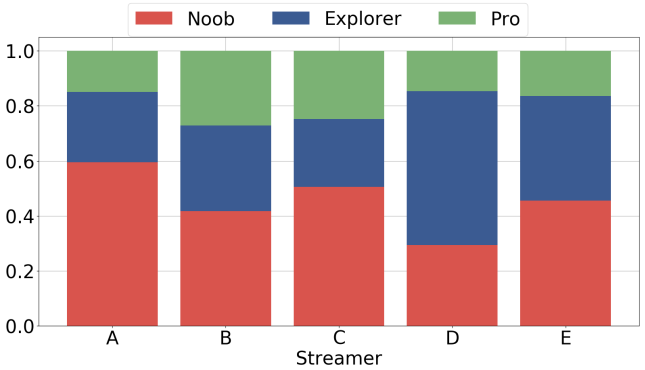

Figure 6: Normalised stacked bar chart of the three play styles across the five streamers.

performed with these two unsupervised learning algorithms collectively indicates that the most appropriate number of data clusters lies between two and three. Two clusters partition the data into highly unbalanced clusters, with 86 matches (74, 947 events) for the first cluster and 238 matches (44, 398 events) for the second cluster. Similarly, four clusters yield an unbalanced distribution, with 152 matches (14, 266 events) for the first cluster, 53 matches for the second cluster (47, 088 events), 95 for the third cluster (35, 367 events) and 24 for the fourth cluster (22, 624 events). Three clusters, however, yield a more uniformly distributed match data partitioning, with 155 (14, 858 events), 105 (42, 878 events) and 64 matches (61, 609 events) for the first, second and third cluster, respectively. If we use the information entropy (H) [35] as a measure of the balance of the distribution of the matches obtained, we notice a higher entropy ($H = 0.95$) with three clusters compared to two ($H = 0.84$) and four clusters ($H = 0.87$). Given the high imbalance of matches partitioned with two clusters, and the similarity of results obtained by the two clustering algorithms it appears that the most reliable way to partition this dataset is through three clusters.

To label the three player styles clustered, we investigate how the features of gameplay are grouped within each cluster. Figure 5 shows the distribution of four representative features across the three clusters. The features displayed are *Delta Location* (distance covered in a match), *Kill* (number of opponents killed in a match), *Taking Damage* (damage taken by the player in a match), and *Time* (match duration in seconds). Using popular game culture terminology we label the first cluster as *Noob* play style as in those matches the streamer does not play particularly well, he reaches a low number of kills and is more likely to be killed. Meanwhile, the matches are much shorter, most likely because the streamer dies within the first minutes of the match. The second cluster of play style is labelled as *Explorer*: in those matches the streamer explores the map far more—as the *Delta Location* feature is higher compared to the other two clusters—but the performance of the player is still average, as shown by the *Kill* and *Being Killed* features. Finally, we label the third play style as *Pro* as it features matches where the streamer has played his best: he tends to kill more players, to die less often compared to the other two clusters, and while it takes a considerable amount of damage he survives longer (i.e., higher *Time* values), most likely winning the match.

Figure 6 shows the distribution of the three play styles across the five streamers and the variation of play styles the different streamers depict, validating our hypothesis. In particular, *A* is labelled as a *Noob* in the majority of his matches, while *D* appears to be more of a *Explorer* player type. *C*, *B*, and *E* show a more uniform distribution of the three play styles in their gameplay.

Given the three different play styles we obtained we test the generalisability of moment-to-moment engagement models that are built on the play styles, instead of the streamers. We thus train a separate engagement model per play style. Following Section 4.2, we perform an exhaustive search of the predetermined values of α and ϵ for each play style model. To compare the results obtained, we validate our models using a *leave-one-streamer-out* cross-validation scheme. Figure 7 illustrates the average test accuracies obtained for the three different play style models of engagement. All models are predicting engagement with high degrees of accuracy (over 75% on average) but the model for the *Noob* play style performs better (78.0% on average, 84.2% at best) than the models for the *Explorer* (77.0% on average, 81.4% at best) and the *Pro* play style (75.4% on average, 80.7% at best). The key findings in this section suggest

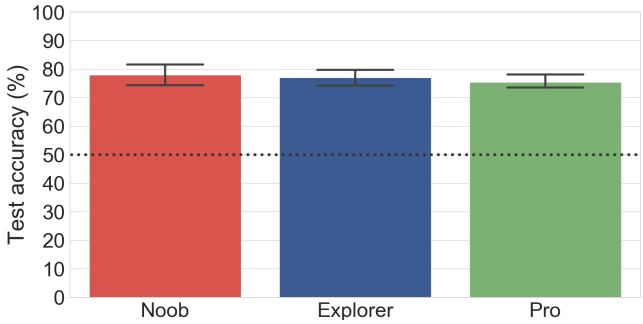

Figure 7: Test accuracies across the three play styles averaged from 5 runs of leave-one-streamer-out validation. Error bars denote 95% confidence intervals.

that constructing models of engagement across streamer play styles instead across streamers offers a higher generalisability potential for the model.

4.4 Engagement Line Analysis

In this section we discuss two indicative examples of PUBG matches with their corresponding engagement line prediction. Figure 8 shows the two example engagement lines as associated with video frame sequences taken from streamer *A*. To extract and display a continuous line of engagement between events, we apply a moving average (sampled every second) to the output of the engagement model.

At the top graph of Figure 8 we observe a steady increase of engagement as the player starts the match searching for enemies and exploring around the map but with no battle events occurring. Towards the end of the match we observe a further increase of the engagement value which is associated with a fast-pace action phase in which the player engages in battle with several opponents. The player at the bottom graph of Figure 8 is shooting and healing himself during the initial phase of the match (first 20 seconds); as a result the model predicts high engagement values for this initial phase. In the middle phase of the match (200 – 500 seconds) the player drives around the map and hence the model yields low levels of engagement. Towards the end of the match, however, the engagement value increases rapidly as the player gets shot and he is engaged in a battle against another player in a house.

5 DISCUSSION

The key findings of this paper suggest that it is not only possible to rely solely on a number of key gameplay events and predict the level of viewer engagement in a continuous fashion but that it can be done with high levels of accuracy.

The approach, however, needs to be tested across a number of varying properties of the dataset considered. Even though the results obtained on 5 PUBG streamers—and several thousand game events—already support the scalability of the method within this game, a larger dataset across more streamers will make our findings even stronger. In addition to the size of the dataset alternative machine learning methods will need to be tested involving deep learning methods as these are more appropriate for larger datasets. It is already highly encouraging, however, that accuracies of over 80% could be reached with relatively simple and shallow ANN architectures. The type of method will also depend on the types of data the model will be trained on. Future studies will consider the use of natural language processing for the chat boxes, facial expression and speech recognition for the streamer, as well as computer vision methods for the pixels of the video stream—as e.g., in [23]—in an attempt to reach more accurate models of engagement. While adding more modalities of input to the predictor of engagement might be beneficial to the accuracy of the models it makes the model more dependable on specific input types and, hence, less versatile. The experiments reported in this paper already suggest that simple telemetry features of the game suffice for the construction of engagement models of high accuracy.

Our notion of gameplay engagement is associated with viewer behaviour and, in particular, with the inverse chatting frequency. While such a proxy of engagement is theoretically grounded, it is supported by recent evidence in the literature, and it can be predicted from gameplay telemetry, other ground truths of engagement are planned to be designed and modelled. Any ad-hoc proxy of engagement—as the one proposed here—will need to be empirically cross-verified against annotation data obtained via video annotation tools such as PAGAN [27, 28]. It is important to note, however, that verifying the ad-hoc engagement proxy we designed in this initial study is beyond the focus of this paper; the core outcome of this study, instead, is that engagement (as defined here) is both theoretically supported and can be predicted accurately from gameplay characteristics in a moment-to-moment fashion. Given the absence of any engagement annotation or emotion labelling in the PUBG game (and most streamed games), reframing the problem of engagement and looking at it from the lens of the viewers' behaviour offers a general-purpose ground truth that can be easily obtained without further human intervention and tedious annotation processes.

The presented results are relevant to researchers and game industry stakeholders alike. The presented methodology can serve as a basis for future studies towards a more holistic understanding of engagement in games and beyond. Since our ad-hoc metric focuses on game content-related engagement of spectators, it highlights the elements of the gameplay experience, which can be controlled by developers. Thus, games that are developed with streaming content in mind can largely benefit from this type of engagement approximation from the early stages of creation. Our system could also aid industry stakeholders and streaming services who face challenges of information-overload and are in need of algorithmic ways to sort and highlight engaging content. Finally, predicting spectator engagement with the streamed game content can also be used for the procedural generation of play with the aim of creating and curating artificial streams [39]. In light of the highly promising results, this study offers some initial evidence that engagement of gameplay videos can be predicted through game telemetry in a particular game of a particular genre. Future studies will focus on testing the proposed methodology across different games of the battle royale genre and also across dissimilar game genres.

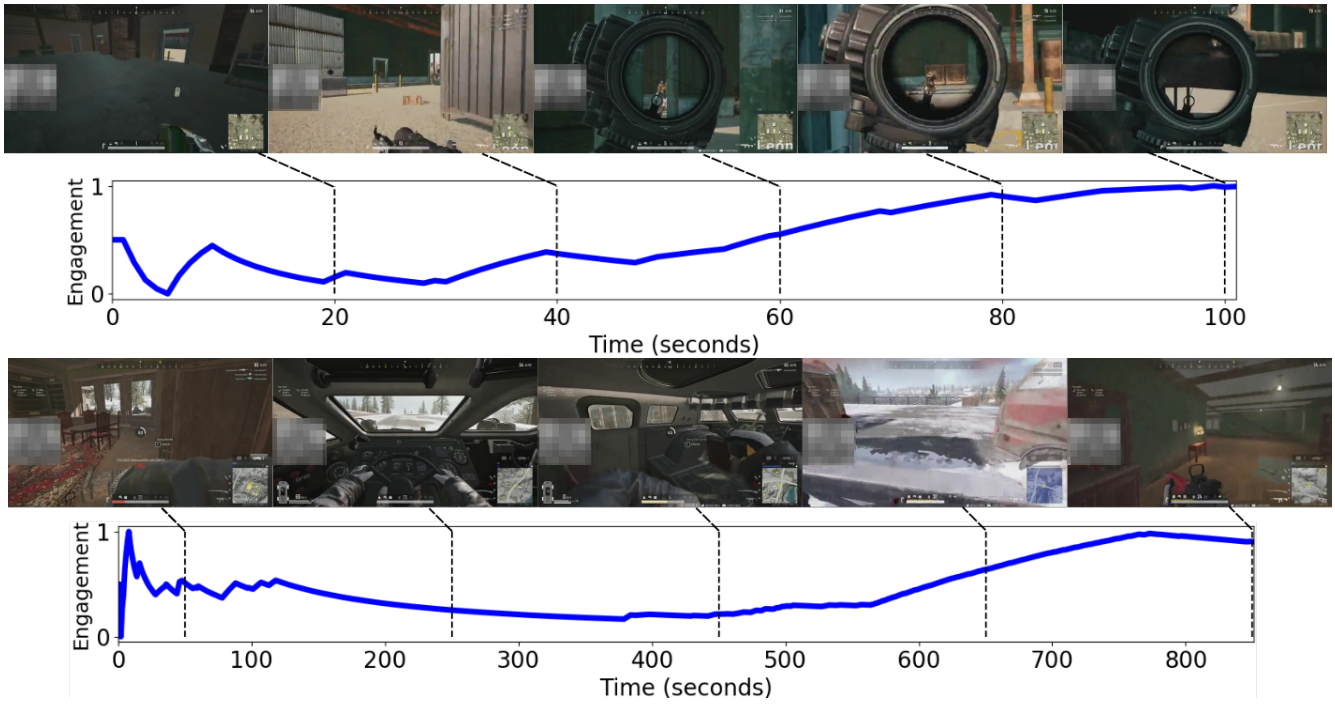

Figure 8: Two indicative examples of video streams taken from the streamer *A* with the corresponding engagement lines as predicted by our ANN model. The figure depicts stream screenshots at particular events during the game. The video on the screenshots is obtained through the Twitch Developer API (fair use), PUBG is a registered trademark of PUBG CORPORATION.

6 CONCLUSIONS

In this paper we reframe the way gameplay engagement is naturally viewed: from a first person to a third person perspective. In particular, we attempt to predict the moment-to-moment engagement of viewers of live streamed games through their chatting activity. We test our hypothesis that gameplay can be a good predictor of viewer chat frequency in PUBG (PUBG Corporation, 2017) live streams that are obtained over 5 Twitch streamers. We employ shallow ANNs and we model engagement as a function of player metrics across critical events of the game. Our results showcase that modelling engagement in a continuous fashion is not only possible but that viewer engagement can be predicted with accuracies, that reach 80% on average (and 84% at best). The models appear to be versatile within and across different streamers as well as across different play styles of streamers. This initial study showcases the potential of the approach for measuring moment-to-moment engagement in game streams (and beyond) through simple yet critical events in the game.

REFERENCES

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*. 265–283.

[2] Sander CJ Bakkes, Pieter HM Spronck, and Giel van Lankveld. 2012. Player behavioural modelling for video games. *Entertainment Computing* 3, 3 (2012), 71–79.

[3] Francesco Barbieri, Luis Espinosa Anke, Miguel Ballesteros, Juan Soler, and Horacio Saggion. 2017. Towards the understanding of gaming audiences by modeling twitch emotes. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*. 11–20.

[4] Richard Bartle. 1996. Hearts, clubs, diamonds, spades: Players who suit MUDs. *Journal of MUD research* 1, 1 (1996), 19.

[5] Christian Bauckhage, Anders Drachen, and Rafet Sifa. 2015. Clustering game behavior data. *IEEE Transactions on Computational Intelligence and AI in Games* 7, 3 (2015), 266–278.

[6] Rafael A Calvo, Sidney D'Mello, Jonathan Matthew Gratch, and Arvid Kappas. 2015. *The Oxford handbook of affective computing*. Oxford University Press, USA.

[7] Alessandro Canossa, Sasha Makarovych, Julian Togelius, and Anders Drachenn. 2018. Like a DNA String: Sequence-Based Player Profiling in Tom Clancy's The Division. In *Proceedings of the 14th Artificial Intelligence and Interactive Digital Entertainment Conference*.

[8] Peter Carruthers and Peter K Smith. 1996. *Theories of theories of mind*. Cambridge University Press.

[9] Peter Chapman, Sanjeebhan Selvarajah, and Jane Webster. 1999. Engagement in multimedia training systems. In *Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences. 1999. HICSS-32. Abstracts and CD-ROM of Full Papers*. IEEE, 9–pp.

[10] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. 2015. Fast and accurate deep network learning by exponential linear units (ELUs). *arXiv preprint arXiv:1511.07289* (2015).

[11] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*. ACM, 191–198.

[12] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, et al. 2010. The YouTube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems*. ACM, 293–296.

[13] Anders Drachen, Alessandro Canossa, and Georgios N Yannakakis. 2009. Player modeling using self-organization in Tomb Raider: Underworld. In *2009 IEEE symposium on computational intelligence and games*. IEEE, 1–8.

[14] Anders Drachen and Shawn Connor. 2018. Games Analytics for Games User Research. In *Games User Research*, Anders Drachen, Pejman Mirza-Babaei, and

Lennart E Nacke (Eds.). Oxford University Press, 333–355.
[15] Magy Seif El-Nasr, Anders Drachen, and Alessandro Canossa. 2016. *Game analytics*. Springer. 792 pages.
[16] Colin Ford, Dan Gardner, Leah Elaine Horgan, Calvin Liu, AM Tsaasan, Bonnie Nardi, and Jordan Rickman. 2017. Chat speed op pogchamp: Practices of coherence in massive twitch chat. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 858–871.
[17] William A Hamilton, Oliver Garretson, and Andruid Kerne. 2014. Streaming on twitch: fostering participatory communities of play within live mixed media. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1315–1324.
[18] Morgan Jennings. 2000. Theory and models for creating engaging and immersive ecommerce websites. In *Proceedings of the 2000 ACM SIGCPR conference on Computer personnel research*. ACM, 77–85.
[19] Nicole Lazarro. 2004. Why we play games: Four keys to more emotion without story. In *Game developer's conference, San Jose*.
[20] Antonios Liapis, Daniele Gravina, Emil Kastbjerg, and Georgios N Yannakakis. 2019. Modelling the Quality of Visual Creations in Iconoscope. In *International Conference on Games and Learning Alliance*. Springer, 129–138.
[21] Lassi A Liikkanen. 2013. Three Metrics for Measuring User Engagement with Online Media and a YouTube Case Study. *arXiv preprint arXiv:1312.5547* (2013).
[22] Tobias Mahlmann, Anders Drachen, Julian Togelius, Alessandro Canossa, and Georgios N Yannakakis. 2010. Predicting player behavior in Tomb Raider: Underworld. In *Proceedings of the Symposium on Computational Intelligence and Games (CIG)*. IEEE, 178–185.
[23] Konstantinos Makantasis, Antonios Liapis, and Georgios N Yannakakis. 2019. From Pixels to Affect: A Study on Games and Player Experience. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 1–7.
[24] Hector P Martinez, Georgios N Yannakakis, and John Hallam. 2014. Don't classify ratings of affect; rank them! *IEEE transactions on affective computing* 5, 3 (2014), 314–326.
[25] Akhil Mathur, Nicholas D Lane, and Fahim Kawsar. 2016. Engagement-aware computing: Modelling user engagement from mobile contexts. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 622–633.
[26] David Melhart, Ahmad Azadvar, Alessandro Canossa, Antonios Liapis, and Georgios N Yannakakis. 2019. Your Gameplay Says it All: Modelling Motivation in Tom Clancy's The Division. *Proceedings of IEEE Conference on Games (CoG)* (2019).
[27] David Melhart, Antonios Liapis, and Georgios N Yannakakis. 2019. PAGAN: Platform for Audiovisual General-purpose ANnotation. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 75–76.
[28] David Melhart, Antonios Liapis, and Georgios N Yannakakis. 2019. PAGAN: Video Affect Annotation Made Easy. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 130–136.
[29] Ilya Musabirov, Denis Bulygin, Paul Okopny, and Ksenia Konstantinova. 2018. Between an arena and a sports bar: Online chats of esports spectators. *arXiv preprint arXiv:1801.02862* (2018).
[30] Heather L O'Brien and Elaine G Toms. 2008. What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American society for Information Science and Technology* 59, 6 (2008), 938–955.
[31] Rui Pan, Lyn Bartram, and Carman Neustaedter. 2016. TwitchViz: a visualization tool for twitch chatrooms. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 1959–1965.
[32] África Periáñez, Alain Saas, Anna Guitart, and Colin Magne. 2016. Churn prediction in mobile social games: towards a complete assessment using survival ensembles. In *Proceedings of the International Conference on Data Science and Advanced Analytics (DSAA)*. 564–573.
[33] Matthias Rauterberg. 1995. About a framework for information and information processing of learning systems. In *Information System Concepts*. Springer, 54–69.
[34] Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* 20 (1987), 53–65.
[35] Claude Elwood Shannon. 1948. A mathematical theory of communication. *Bell system technical journal* 27, 3 (1948), 379–423.
[36] Max Sjöblom and Juho Hamari. 2017. Why do people watch others play video games? An empirical study on the motivations of Twitch users. *Computers in Human Behavior* 75 (2017), 985–996.
[37] Max Sjöblom, Maria Törhönen, Juho Hamari, and Joseph Macey. 2017. Content structure is king: An empirical study on gratifications, game genres and content type on Twitch. *Computers in Human Behavior* 73 (2017), 161–171.
[38] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* 15, 1 (2014), 1929–1958.
[39] Ruck Thawonmas and Tomohiro Harada. 2017. AI for Game Spectators: Rise of PPG. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*.
[40] Elaine G Toms. 2002. Information interaction: Providing a framework for information architecture. *Journal of the American Society for Information Science and Technology* 53, 10 (2002), 855–862.
[41] Markus Viljanen, Antti Airola, Jukka Heikkonen, and Tapio Pahikkala. 2018. Playtime measurement with survival analysis. *IEEE Transactions on Games* 10, 2 (2018), 128–138.
[42] Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association* 58, 301 (1963), 236–244.
[43] Sarah Wassermann, Nikolas Wehner, and Pedro Casas. 2019. Machine Learning Models for YouTube QoE and User Engagement Prediction in Smartphones. *ACM SIGMETRICS Performance Evaluation Review* 46, 3 (2019), 155–158.
[44] Tim Wulf, Frank M Schneider, and Stefan Beckert. 2018. Watching players: An exploration of media enjoyment on Twitch. *Games and Culture* (2018), 1555412018788161.
[45] Georgios N Yannakakis, Roddy Cowie, and Carlos Busso. 2017. The Ordinal Nature of Emotions. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 248–255.
[46] Georgios N Yannakakis, Roddy Cowie, and Carlos Busso. 2018. The Ordinal Nature of Emotions: An Emerging Approach. *IEEE Transactions on Affective Computing* (2018).
[47] Georgios N Yannakakis, Pieter Spronck, Daniele Loiacono, and Elisabeth André. 2013. Player modeling. In *Dagstuhl Follow-Ups*, Vol. 6. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.
[48] Georgios N. Yannakakis and Julian Togelius. 2018. *Artificial Intelligence and Games*. Springer Nature. http://gameaibook.org.
[49] Nick Yee. 2006. Motivations for play in online games. *CyberPsychology & behavior* 9, 6 (2006), 772–775.