Learning a QoE Metric from Social Media and Gaming Footage

Catalina Alvarez
EPFL
Lausanne, Switzerland
catalina.alvarezinostroza@epfl.ch

Katerina Argyraki
EPFL
Lausanne, Switzerland
katerina.argyraki@epfl.ch

Abstract

Defining a universal metric for Quality of Experience (QoE) is notoriously hard due to the complex relationship between low-level performance metrics and user satisfaction. The most common metric, the Mean Opinion Score (MOS), has well-known biases and inconsistency issues. We propose an alternative that leverages (a) social-media comments on network performance and (b) streaming footage that includes performance numbers. We argue that our proposal is feasible for online gaming, and it may apply to other applications in the near future. We discuss its potential to enable a direct mapping from low-level performance metrics to accurate QoE scores—the golden standard for assessing user satisfaction.

Keywords

Internet Measurement; Quality of Experience; Social Media

1 Introduction

Quality of Experience (QoE) is meant to capture user satisfaction or annoyance with a service [13]. It is a vital quantity for service providers, as there is a direct relationship between user satisfaction and business success. Ideally, a service provider would want to know how to achieve a satisfactory QoE while minimizing resource requirements and, more generally, saving costs [26].

The ideal QoE metric would accurately capture user satisfaction while enabling a direct mapping from low-level performance (e.g., packet loss, latency, or throughput) to a QoE score. Designing such a metric, however, has proven challenging due to the complex, sometimes counterintuitive, relationship between low-level performance metrics and user satisfaction. E.g., prior work has shown that, in video services, there is a non-monotonic relationship between throughput and user satisfaction [4]. User context (e.g., emotional state and personality), as well as the particular type of application and content, introduce additional confounding factors [17].

Over the years, there have been many attempts to design such a QoE metric, with different degrees of success. The industry standard is the Mean Opinion Score (MOS), which relies on asking users to rate their experience on a scale; it is simple to compute but has well-known biases and inconsistency issues [29]. The typical way to address MOS's limitations is by taking the human out of the loop and focusing on specific applications. E.g., VMAF [18] leverages machine learning to learn a mapping from low-level video-performance metrics to MOS scores; this does remove the biases and inconsistencies due to human subjectivity, but is susceptible to manipulation [28].

We propose a different approach to QoE assessment, which leverages (a) social-media comments on application performance and (b) streaming footage that includes low-level performance numbers. In particular, previous work has shown that users go to social media to express their dissatisfaction with application performance [24]. We propose to collect such comments and apply sentiment analysis to them; the QoE score will simply be the outcome of sentiment analysis. A key challenge is linking social-media comments to the underlying performance that caused them. We can solve this challenge for a particular application—online gaming—by processing publicly available gaming footage and extracting the latency numbers that it includes.

We argue that this approach has the potential to combine the best of the state of the art: consider users' opinions without the biases that result from explicitly asking; and enable a direct mapping from low-level performance numbers to QoE scores.

Ethical considerations. Our work does raise ethical issues because it uses data posted on social media. We take extreme care to not leverage any information that a streamer did not clearly intend to share publicly (e.g., the fact that the streamer owns a particular social-media account). We also take extreme care to not violate the Terms of Service of any social-media or streaming platform (e.g., we do not store any video streams—only thumbnails from video streams, explicitly made available by the streaming platform). We have formal approval from our institution's Ethical Review Board for all the data collected. However, our approach raises a more complex issue: is it, in principle, too invasive to extract particular elements from streaming footage? We discuss this issue and outline a direction that has the potential to address it at the end of the paper.

2 Problem

The de-facto metric for measuring QoE is the Mean Opinion Score (MOS) [12]. An "opinion score" is "the value on a predefined scale that a subject assigns to their opinion of the performance of a system" [13]. The International Telecommunications Union (ITU) recommends a 5-value scale: {Excellent, Good, Fair, Poor, and Bad}. A system's MOS is simply the average opinion score of multiple users.

MOS is valued for two main reasons: (1) It is simple: The typical way to collect opinion scores is to conduct a short user survey, e.g., as was done to assess the impact of latency on gaming [3] and the QoE of mobile applications over WiFi [8]. There exist even simpler approaches, such as OneClick [6] and HostView [14], which allow users to give their opinion score instantaneously by clicking a button. (2) It is a mature and well-defined standard, refined over more than 20 years, with many readily available tests and detailed guidelines to ensure their validity.

At the same time, MOS has the limitations that result from explicitly asking and relying on users' opinions. A human's answer is often biased by the very fact that they are participating in a survey; their emotional state; their personality, or likes and dislikes [6]. All these factors may be unrelated to the application performance that MOS is supposed to assess. The personality bias can be corrected by normalizing a user's answers based on their history, however, this requires polling the same users often, which is considered a nuisance. In general, these biases can be corrected by performing the user surveys in controlled environments, however, this limits the scale of what can be measured [29].

Moreover, MOS has the limitations that result from asking users to rate their experience on an absolute scale—a task that is subject to several cognitive biases [20]. For example, humans tend to evaluate relatively rather than absolutely ("centering bias"), and to adjust their scores in order to span the whole range ("range-equalizing bias") [30]. Moreover, the same scale may be perceived as linear or non-linear by different users, e.g., depending on the language in which the survey is carried out [29].

The alternative to MOS is to define application-specific, objective QoE metrics, which can be extracted automatically (without asking for the user's opinion). One line of work measures QoE based on how well users perform application-specific tasks that depend on it. For example, Durin et al. [9] measure the quality of a telecommunication system based on how it affects digit or letter recognition (in particular, reaction times and error rates). Of course, this approach is limited to applications where there is a measurable relationship between QoE and user performance. Another line of work defines objective QoE metrics and learns how to automatically map them to MOS scores. E.g., VMAF [18] uses learning to create a map from objective video-specific metrics, e.g., Visual Information Fidelity (VIF) [27], to MOS scores. It has been shown that VMAF's scores can approximate well the scores that humans would give [25]. However, it is vulnerable to pre-processing methods that distort the video to artificially increase its VMAF score (without improving its user-perceived quality) [28].

3 Opportunity

Considering the strengths and weaknesses of MOS and its alternatives, we believe that there is a sweet spot somewhere between the two, i.e., between subjective opinion scores and objective application-specific metrics. On the one hand, we want to avoid the biases that result from asking humans to rate on an absolute scale. On the other hand, we expect that any approach that completely excludes the human user will be sensitive to pre-processing, and hence can be gamed. Hence, we need a metric that somehow factors in users' opinions without directly asking for them, and without involving an absolute scale.

Instead of asking for users' opinions, one can passively listen to the opinions they freely express on social media. The latter constitute easily accessible platforms for users to share information and freely express their thoughts, which occasionally include their opinions on application performance. In particular, previous work has shown that users go to social media, such as Twitter, to complain when services are not available [19]; other work has compared tweets where users complain about a service with customer-care tickets, and it has found a direct correlation between the two, and that problems often appear on Twitter before they show up in tickets [24]; finally, social media has served as a source of data showing user satisfaction with a service [23].

The challenge with collecting users' opinions from social media is linking each opinion to the underlying low-level performance metrics that led to it. For example, when an application explicitly asks the user to rate a call, it can directly link that score to latency, packet loss, rebuffering events, etc. Social-media platforms, on the other hand, are typically decoupled from the user's environment.

Gaming footage has the potential to solve this challenge in the context of a specific—yet arguably very important—application: Gaming is exceptionally sensitive to latency and packet loss, so network problems are likely to bother players [21]. As a result, many games display on-screen the latency and/or packet loss between the game server and the client. Live-streaming of online gaming is on the rise [1], giving us plenty of access to gaming footage with latency and packet-loss numbers. At the same time, gaming and streaming are social activities, and players who stream tend to have an active online presence [10].

4 Proposal

Our idea is to learn an application's QoE metric from publicly available data: collect comments on application performance from social media; find any matching application-streaming footage and extract any visible low-level performance numbers from it; and perform sentiment polarity analysis [22] on the comments to map the low-level performance numbers to a QoE score. The QoE score will be simply the outcome of sentiment analysis, normalized based on the user's history of comments. Currently, this is applicable only to gaming; however, it is plausible that some of the interactive applications of the future are (a) social activities and (b) network-performance sensitive, hence it is plausible that they will generate streaming footage with visible performance numbers.

Sentiment polarity analysis (from now on simply "sentiment analysis") takes as input a piece of text, and it outputs a score, from -1 to 1, which represents the level of satisfaction, happiness, or, in general, "positive feelings" reflected by the text. "-1" and "+1" represent, respectively, "strongly negative" and "strongly positive" feelings, while "0" represents the absence of positive or negative feelings (e.g., a legal text should yield "0").

We think that this approach has the potential to remove some of MOS's limitations: First, collecting opinions that users freely express on social media removes the bias resulting from explicitly asking users' opinions. Second, applying sentiment analysis removes the biases resulting from asking users to rate on an absolute scale. Intuitively, for most users, it should be easier to describe their experience in their own words than to assign an absolute score to it.

If naïvely implemented, our approach could be as vulnerable as MOS to users' personality and emotional state; however, we hypothesize that a user's social media and streaming accounts provide context that can be used to correct the resulting biases. For example, by applying sentiment analysis to a user's history of social-media posts, it should be possible to quantify the user's predisposition toward (dis)satisfaction. Or, by considering a user's emotional state during many different occurrences of the same/similar low-level performance, it should be possible to get a reasonable estimate of whether/how that low-level performance affects the user's emotional state. We are by no means saying that personality or emotional-state biases can be completely eliminated; only that social media and streaming accounts provide significantly more context than MOS has available to correct these biases.

Using the output of sentiment analysis as a QoE metric has other significant advantages: It is application-agnostic, and its output will evolve with users' expectations over time. So, if our approach becomes applicable to different applications, one can imagine using the same sentiment-extraction algorithm on any user comments, regardless of the application they concern, or the social-media platform they come from, trivializing comparison across applications and over time. Moreover, sentiment analysis is an important and growing research topic in its own right [5], which will only benefit from the advances in Natural Language Processing (NLP) that are expected over the coming decades. So, redefining QoE measurement as an application of sentiment analysis seems like a good bet.

Is it feasible to extract low-level performance numbers from streaming footage in the first place? We have built a system, called Tero [2], that does it: it extracts latency numbers from gaming footage provided by the Twitch[1] streaming platform. In summary, Tero operates as follows: (a) It periodically downloads thumbnails—images extracted from live video streams—of gaming footage posted on Twitch's Content Distribution Network (CDN). (b) It extracts from each thumbnail a latency number, by combining Optical Character Recognition (OCR) and heuristics that leverage game user interfaces; 96.3 ± 0.40% of the latency numbers extracted this way are correct, i.e., not the result of OCR error or the latency number being obstructed on screen. Figure 1 shows examples of latency numbers visible on thumbnails. (c) It tries to associate each sequence of latency numbers extracted from a stream with a social-media account, by searching for voluntary connections that the streamer may have created between their streaming and social-media accounts. Since March 2021, Tero has extracted 96 million latency numbers from 184 thousand streamers with intentional connections between their streaming and social-media accounts.

5 Preliminary Evidence

We answer the following preliminary questions: Can we find streamers who (a) complain on social media about network performance and (b) explicitly link from their social-media to their streaming accounts? What does sentiment analysis say about their complaints, e.g., are they more negative than their other comments? (§5.1) Can we find network-performance problems that are visible on streaming footage? (§5.2)

5.1 Sentiment Analysis of Complaints

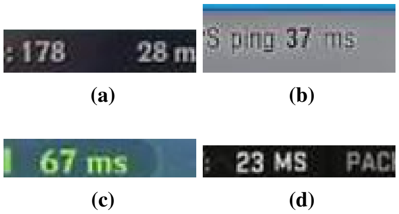
Our source of complaints is tweets that we collected during the first week of June 2023 (but may have been posted at any point until we collected them). We collected three sets of tweets:

(a) **Complaints.** We searched for the tags and terms shown in Table 1, which were partly inspired by prior work [15, 16, 19]. Each search was a combination of "#twitch" and one other tag/term, e.g., "#twitch #internetproblems" or "#twitch lag spikes." Table 2 shows examples of searches and the complaints that they yielded. For each search, we downloaded all the resulting tweets, removed all retweets and non-network-related comments, and kept only the complaints and original authors. For each Twitter user who authored a complaint, we looked for an explicit, publicly disclosed link to a Twitch streaming account, as well as a geographical location at the granularity of a city, state, or country. In the end, we obtained 2,442 tweets with complaints about network performance, from 1,587 unique users with an explicit link to a Twitch account; we call the latter "complainers." We note that complaints are not common: on average, each user made 1.5 complaints.

(b) **Complainer tweets without complaints.** These are all the tweets posted by all complainers, excluding the tweets with complaints. We used this set to assess the difference in sentiment between complaints and no-complaint tweets.

(c) **Non-complainer tweets.** These are all the tweets posted by a set of randomly selected non-complaining users. We created this set as follows: for each active Twitch streamer A, we looked for a Twitter profile P with the same username as A; if we found such a profile, we checked whether P included an explicit link to A; if yes, we associated P and A, and we added this streamer to our set. We used this set to assess the

[1] https://www.twitch.tv/



(a) (b)

(c) (d)

Figure 1: Examples of latency numbers visible on Twitch thumbnails.

difference in the sentiment expressed by non-complainers and complainers in non-complaint tweets (to confirm that complainers do not complain because of their personality).

Sentiment-analysis tool. We computed the sentiment of each downloaded tweet with VADER [11], a lexicon-based sentiment model that specifically targets social media. To evaluate the sentiment polarity of a tweet, VADER classifies each word as "positive," "negative," or "neutral" using a dictionary of words generally labeled according to their semantic orientation. Then, VADER compiles the proportion of each category of words in the whole sentence; from these three values, it computes a compound score by adding the 3 values and applying heuristic rules that approximate human intuition [11]. The compound score varies from -1 (a completely negative, or unpleasant, sentiment) to +1 (a completely positive, or pleasant, sentiment).

Complaint sentiment vs baseline. We compared the median sentiment of (a) a complainer's complaints vs (b) the same complainer's non-complaint tweets (we call the latter the complainer's "baseline"). Figure 2a shows the Cumulative Distribution Function (CDF) of the percentage difference between these two quantities; each point corresponds to one complainer; a negative X-axis value indicates sentiment that is more negative than the baseline. For 80% of the complainers, the median complaint is (as expected) more negative than the baseline; for 50% of the complainers, the relative difference between their median complaint sentiment and their baseline is at least -25%. However, for 20% of the complainers, the median complaint sentiment is (surprisingly) more positive than their baseline. Manual inspection revealed three explanations for this: (a) *Mixed signals*, where the complaint mixes a negative and a more positive situation, e.g., "My internet died during my birthday. Hurray!" (b) *Sarcasm*, e.g., "Thank you provider", "Looking forward to my high ping." (c) *Progress reports* about a network problem being worked on, which is interpreted as positive, e.g., "Working with tech support to fix the major lag spikes."

Complainer vs non-complainer sentiment. We computed the sentiment median and span (the difference between the 75th and 25th percentiles) of (a) each complainer's tweets without complaints and (b) each non-complainer's tweets. Figure 2b shows the results; each blue circle corresponds to one non-complainer, and each orange cross to one complainer. We see that the two groups have comparable sentiment median and span values. As a side note—and perhaps contrary to what one would expect—all users have a positive baseline. We argue that this makes sense because the Twitter users that we selected are also Twitch streamers. Streamers tend to use Twitter as an advertisement platform, i.e., they tweet positive messages in order to attract viewers. For example, many streamer tweets enthusiastically invite people to check out their streams.

In summary: Used out-of-the-box, a standard sentiment-analysis tool classified most tweets that complain about network performance as more negative than the same user's baseline. At the same time, the tool classified a non-negligible number of complaints as positive due to mixed signals, sarcasm, or the complaint including a progress report. We conclude that sentiment polarity is a promising QoE metric; however, we may need to develop better techniques for detecting sarcasm or negative comments that are embedded in a more positive context.

5.2 Latency Problems on Gaming Footage

Ideally, we would process the gaming footage of our complainers and non-complainers (§5.1), extract their latency numbers, and check whether there is a statistically significant correlation between poor latency and complaints about network performance. However, we do not (yet) have access to a sufficient amount of footage to perform such analysis. So, we set the more modest goal of finding concrete examples of complainers with poor latency that is visible on their gaming footage.

Collected latency numbers. Our source of latency numbers is gaming footage from Twitch and, in particular, thumbnails generated between June 2 and June 23, 2023. Twitch's Terms of Service[2] may be interpreted as forbidding the processing of live video streams; however, they do generate a thumbnail from each stream every 5 minutes and make that publicly available on their CDN. Of the 1,587 complainers (§5.1), we identified those who streamed during this period and contributed at least 10 latency numbers; for each of them, we identified all the non-complainers who streamed during the same period and from the same geographical area; and we retained only {complainer, non-complainers} groups with at least 5 non-complainers. We manually filtered the 54 remaining complainers and retained only those whose complaints were clearly related to network performance. In the end, we retained 11 complainers and 2,262 non-complainers, and we extracted, respectively, 1,518 and 360,852 latency numbers from them.

Complainer vs non-complainer latency. We compared the latency distribution of each complainer against the latency distribution of the non-complainers who streamed from the same geographical area. Figure 2c shows the results as a boxplot; each box specifies median latency (horizontal line across), 25th and 75th percentile (box top and bottom), and 1st and 99th percentile (whiskers). Each blue box corresponds to a set of non-complainers, while each orange box corresponds

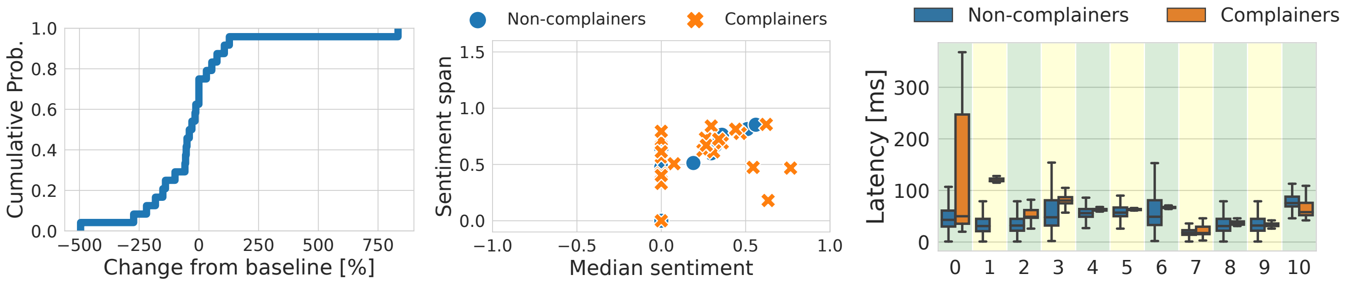
[2] https://www.twitch.tv/p/en/legal/terms-of-service/#7-license

Problem tags	#badinternet, #framedrops, #internetproblems, #internet#down, #internetissues, #lagkills, #outage, #packetloss, #lag, #shitinternet, #shittyinternet, #streamlag, #throttling
ISP tags	#isp1, #isp2, #isp3, #isp4, #isp5
Search terms	"lag spikes", "packet loss", "high ping", "internet issues", "internet problems", "internet outage"

Table 1: Search tags and terms used to find complaints on Twitter.

#twitch #badinternet	Tried Streaming today but my net has decided to not cooperate this time >.< #gaming #twitchstreamer #twitch #badinternet
#twitch #internetproblems	Will the internet be fixed? Find out on the next episode on twitch! #smallstramer #twitch #twitchstreamer #internetproblems
#twitch #isp1	WHY THY DO I HAVE TO LIVE THROUGH THESE DARK AGES?! My internet is being naughty. #Twitch #isp1 #Naughty
#twitch #isp2	#ISP2 IS THE LITERAL DEVIL #twitch #live #streaming
#twitch internet issues	***[Case 1] Internet issues are great! Finally #live on #twitch!!*** ***[Case 10] Unfortunatly due to internet issues I have had to postpone my 24 hour gamethon until the 9th may.#internetissues #twitch*** I have sacrificed multiple bamboo sticks on the altars of the Electric and Internet gods. Hopefully, this appeases them and we have no issues today. #twitchstreamer #twitch.
#twitch packet loss	***[Case 0] hitting far away running headshots WITH packet loss :) #twitch***
#twitch #lag	***[Case 2] Almost quit because of the LAG. Ended up winning with 11k instead.***

Table 2: Complaints found on Twitter. The highlighted complaints with case numbers correspond to Fig. 2c.



(a) Complaint sentiment vs user baseline.

(b) Complainer vs non-complainer sentiment.

(c) Complainer vs non-complainer latency.

Figure 2: Preliminary data analysis.

to a complainer. In cases 0 to 6, the complainer's latency median or variance is clearly worse than the non-complainers'. In cases 7 and 8, the complainer's median latency is still worse, but the difference is less significant. In cases 9 and 10, the latency distributions do not indicate any reason for complaint. Table 1 shows some of the corresponding complaints: Case 0 is a clear complaint with negative sentiment, consistent with the author's poor latency variance. Cases 1 and 2 are examples of sarcasm and mixed signals, respectively; we expect that with the proper sentiment-analysis technique, these would also yield negative sentiment, consistent with their authors' relatively poor latency. Case 10 is a clear complaint, but it refers to an event in May, so, it is plausible that any network-performance problems that caused the complaint were resolved by June (which is when we collected the gaming footage).

In summary: We encountered concrete examples where a streamer who complained on Twitter about network performance also experienced poor latency that is visible on their gaming footage. We also encountered counter-examples, where a streamer complained on Twitter, but the gaming footage that we collected does *not* indicate poor latency. The former suggests that it is plausible to connect social-media complaints with poor latency that is visible on gaming footage; however, we need a broader measurement study to draw any conclusions.

6 Discussion

Is all this ethical? We start from the easier aspects of ethics and conclude with a more challenging issue:

We take the obvious steps to respect streamers' privacy: We do not use any data that a streamer did not clearly intend to share publicly. In particular, we use data from a streamer's social-media (Twitter) account only if the streamer explicitly left a link from that account to their streaming (Twitch) account. Also, we store only the data that is necessary for our approach: latency numbers and approximate geographic locations (city, state, and/or country). We discard any intermediate data, e.g., streamer usernames or footage thumbnails, as soon as we have processed them. To link subsequent thumbnails to the same streamer, we consistently map each streamer username to a random identifier.

We interpret the Terms of Service of each platform conservatively: First, we draw a connection from a Twitter to a Twitch account only if the Twitter user clearly intended the connection to be drawn; this is consistent with Twitter allowing the drawing of "reasonable" connections[3]. Second, as already mentioned (§5.2), we do not download any live streams or streamer profiles from Twitch.

However, there exists a more challenging issue: Is it too invasive to extract particular elements from streaming footage? Streamers explicitly make their footage public (that is the whole point of streaming). Still, we ask ourselves whether what we are proposing could be directed toward bad purposes, e.g., to extract from footage elements that a streamer thought would go unnoticed.

This question is taking us toward a slightly different direction, which keeps all the benefits of extracting performance metrics from streaming footage while avoiding the ethical ambiguity: Streaming platforms are starting to offer "extensions"—small overlays that streamers explicitly add to their screen to enhance their own and/or their viewers' experience. For example, in a competitive game, an extension may show the streamer's win/lose record or statistics over time. We are considering the idea of writing extensions that display the performance numbers that the corresponding application collects. Installing such an extension provides a clear way for a streamer to give explicit consent for the content displayed by the extension to be extracted from their footage.

Is it easy to manipulate? In a world where QoE is computed from social-media posts and streaming footage, one can imagine mischievous users launching campaigns to complain on social media about network performance without real cause, just to wreak havoc on QoE computation. We hypothesize that it is possible to detect and control the impact of such behavior through data analysis. For example, if a user experiences no change in their normal network performance (according to their streaming footage), yet they start complaining about it after other such complaints appear on social media, that may constitute evidence of manipulation; complaints from such "suspicious" users could be conservatively ignored.

Are gamers representative Internet users? No; we expect a typical gamer to have a better Internet connection, and to require lower and stabler latency to be satisfied relative to a typical Internet user. In general, we expect the users of different applications to have different expectations, so, it makes sense to learn a different QoE metric per application.

Could our idea work for other apps? There are reasons to believe so. First, virtual-reality applications share a lot of properties with gaming (some of them actually *are* games). For instance, they also have a strong social aspect, so people are likely to comment on them on social media and stream them. Also, they are at least as sensitive to network performance as gaming, so it is plausible that they start including low-level performance numbers in their UIs. But even less exotic applications, e.g., video-streaming services like YouTube and Netflix, or peer-to-peer (P2P) networks, display network statistics like throughput on-screen—and P2P networks have even been used to study network performance [7]. Of course, one wonders whether a user would (or should) ever stream themselves watching a movie or downloading something from a P2P network. For better or worse, it is not as crazy as it sounds, given the growing trend to stream arguably trivial activities. After all, Twitch has a dedicated—and quite successful—category for streaming one's sleep [4].

To summarize: Users comment on the performance of their network on social media; it makes sense to apply sentiment analysis to these comments and use the outcome as the measure of their QoE. The challenge is correlating these comments with low-level performance numbers; it is plausible to extract the latter from streaming platforms—for now directly through the streaming footage, but eventually through extensions that would delineate clearer privacy boundaries. If this works, it could yield a direct mapping from different low-level performance metrics to QoE—the golden standard for assessing user satisfaction.

Acknowledgments

We would like to thank the HotNets reviewers for their insightful feedback, especially regarding our preliminary experimental evaluation.

References

[1] [n. d.]. Streamlabs and Stream Hatchet Q3 2022 Live Streaming Report. https://streamlabs.com/content-hub/post/streamlabs-and-stream-hatchet-q3-2022-live-streaming-report. ([n. d.]). Accessed: 2023-10-23.

[2] Catalina Alvarez and Katerina Argyraki. 2023. Using Gaming Footage as a source of Internet latency information. In *Proceedings of the 23rd ACM Internet Measurement Conference*. 606–626.

[3] Rahul Amin, France Jackson, Juan E Gilbert, Jim Martin, and Terry Shaw. 2013. Assessing the impact of latency and jitter on the perceived quality of call of duty modern warfare 2. In *Human-Computer Interaction. Users and Contexts of Use: 15th International Conference, HCI International 2013, Las Vegas, NV, USA, July 21-26, 2013, Proceedings, Part III 15*. Springer, 97–106.

[4] Athula Balachandran, Vyas Sekar, Aditya Akella, Srinivasan Seshan, Ion Stoica, and Hui Zhang. 2012. A quest for an internet video quality-of-experience metric. In *Proceedings of the 11th ACM Workshop on Hot Topics in Networks*. 97–102.

[5] Marouane Birjali, Mohammed Kasri, and Abderrahim Beni-Hssane. 2021. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems* 226 (2021), 107134.

[6] K-T Chen, C-C Tu, and W-C Xiao. 2009. Oneclick: A framework for measuring network quality of experience. In *IEEE INFOCOM 2009*. IEEE, 702–710.

[3] Off-Twitter matching https://developer.twitter.com/en/developer-terms/agreement-and-policy

[4] https://streamhatchet.com/blog/blog-im-only-sleeping-category-on-twitch-grows-to-62m-in-2022/

[7] David R Choffnes, Fabián E Bustamante, and Zihui Ge. 2010. Crowdsourcing service-level network event monitoring. In *Proceedings of the ACM SIGCOMM 2010 Conference*. 387–398.

[8] Alexandre De Masi and Katarzyna Wac. 2019. Predicting quality of experience of popular mobile applications from a living lab study. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 1–6.

[9] Virginie Durin and Laetitia Gros. 2008. Measuring speech quality impact on tasks performance. In *Ninth Annual Conference of the International Speech Communication Association*.

[10] 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.

[11] Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, Vol. 8. 216–225.

[12] ITU-T. 2016. Recommendation p. 800.1: Mean opinion score (MOS) terminology. (2016).

[13] ITU-T. 2017. Vocabulary for Performance, Quality of Service and Quality of Experience. (2017).

[14] Diana Joumblatt, Jaideep Chandrashekar, Branislav Kveton, Nina Taft, and Renata Teixeira. 2013. Predicting user dissatisfaction with internet application performance at end-hosts. In *IEEE INFOCOM 2013*. IEEE, 235–239.

[15] Dennis Kergl, Robert Roedler, and Gabi Dreo Rodosek. 2017. Towards internet scale quality-of-experience measurement with Twitter. In *Security of Networks and Services in an All-Connected World: 11th IFIP WG 6.6 International Conference on Autonomous Infrastructure, Management, and Security, AIMS 2017, Zurich, Switzerland, July 10-13, 2017, Proceedings 11*. Springer, 108–122.

[16] Ege Cem Kirci, Martin Vahlensieck, and Laurent Vanbever. 2022. " Is my internet down?" sifting through user-affecting outages with Google Trends. In *Proceedings of the 22nd ACM Internet Measurement Conference*. 290–297.

[17] Hendrik Knoche, Hermann G De Meer, and David Kirsh. 1999. Utility curves: Mean opinion scores considered biased. In *1999 Seventh International Workshop on Quality of Service. IWQoS'99.(Cat. No. 98EX354)*. IEEE, 12–14.

[18] Zhi Li, Anne Aaron, Ioannis Katsavounidis, Anush Moorthy, and Megha Manohara. 2016. Toward a practical perceptual video quality metric. *The Netflix Tech Blog* 6, 2 (2016), 2.

[19] Marti Motoyama, Brendan Meeder, Kirill Levchenko, Geoffrey M Voelker, and Stefan Savage. 2010. Measuring Online Service Availability Using Twitter. *WOSN* 10 (2010), 13–13.

[20] Jim Mullin, Lucy Smallwood, Anna Watson, and Gillian Wilson. 2001. New techniques for assessing audio and video quality in real-time interactive communications. *IHM-HCI Tutorial* (2001), 1–63.

[21] Manuel Oliveira and Tristan Henderson. 2003. What online gamers really think of the Internet?. In *Proceedings of the 2nd workshop on Network and system support for games*. 185–193.

[22] Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval* 2, 1–2 (2008), 1–135.

[23] Weijie Qi, Rob Procter, Jie Zhang, and Weisi Guo. 2019. Mapping consumer sentiment toward wireless services using geospatial Twitter data. *IEEE Access* 7 (2019), 113726–113739.

[24] Tongqing Qiu, Junlan Feng, Zihui Ge, Jia Wang, Jun Xu, and Jennifer Yates. 2010. Listen to me if you can: tracking user experience of mobile network on social media. In *Proceedings of the 10th ACM Internet Measurement Conference*. 288–293.

[25] Reza Rassool. 2017. VMAF reproducibility: Validating a perceptual practical video quality metric. In *2017 IEEE international symposium on broadband multimedia systems and broadcasting (BMSB)*. IEEE, 1–2.

[26] Raimund Schatz, Tobias Hoßfeld, and Pedro Casas. 2012. Passive Youtube QoE monitoring for ISPs. In *2012 Sixth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*. IEEE, 358–364.

[27] Hamid R Sheikh and Alan C Bovik. 2006. Image information and visual quality. *IEEE Transactions on Image Processing* 15, 2 (2006), 430–444.

[28] Maksim Siniukov, Anastasia Antsiferova, Dmitriy Kulikov, and Dmitriy Vatolin. 2021. Hacking VMAF and VMAF NEG: vulnerability to different preprocessing methods. In *2021 4th Artificial Intelligence and Cloud Computing Conference*. 89–96.

[29] Robert C Streijl, Stefan Winkler, and David S Hands. 2016. Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives. *Multimedia Systems* 22, 2 (2016), 213–227.

[30] Slawomir Zielinski, Francis Rumsey, and Søren Bech. 2008. On some biases encountered in modern audio quality listening tests-a review. *Journal of the Audio Engineering Society* 56, 6 (2008), 427–451.