

A large-scale study of Wikipedia users' quality of experience

Flavia Salutari

Telecom ParisTech

flavia.salutari@telecom-paristech.com

Gilles Dubuc

Wikimedia Foundation

gilles@wikimedia.org

Diego Da Hora

Telecom ParisTech

diego.hora@gmail.com

Dario Rossi

Huawei Technologies, Co. Ltd

dario.rossi@huawei.com

ABSTRACT

The Web is one of the most successful Internet application. Yet, the quality of Web users' experience is still largely impenetrable. Whereas Web performances are typically gathered with controlled experiments, in this work we perform a large-scale study of one of the most popular Web sites, namely Wikipedia, explicitly asking (a small fraction of its) users for feedback on the browsing experience. We leverage user survey responses to build a data-driven model of user satisfaction which, despite including state-of-the art quality of experience metrics, is still far from achieving accurate results, and discuss directions to move forward. Finally, we aim at making our dataset publicly available, which hopefully contributes in enriching and refining the scientific community knowledge on Web users' quality of experience (QoE).

CCS CONCEPTS

- Information systems → World Wide Web;
- Networks → Network measurement; Network performance analysis;

ACM Reference Format:

Flavia Salutari, Diego Da Hora, Gilles Dubuc, and Dario Rossi. 2019. A large-scale study of Wikipedia users' quality of experience. In *Proceedings of The Web Conference (WWW'19)*. ACM, New York, NY, USA, Article 4, 7 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

Since its inception, the World Wide Web has sometimes been dubbed as World Wide "Wait" [9]. Slow rendering of webpages happened due to dial-up connections in the 80s, slow 2G connections in the 90s and so on, but it also persists nowadays for several reasons including unexpected sources of latencies [16], interactions between network protocols [22], the growingly more complex structure of websites [45], an increased usage of mobile devices [19, 34] and the emergence of new protocols [40]. Yet, whereas the study of Web performance is commonly [19, 22, 29, 34, 40, 45–47] tackled via simple objective metrics [20], and rather typically the Page Load Time (PLT), the quality of Web users' experience is still largely impenetrable [15, 28]. As such, a number of alternative metrics that attempt at better fitting the human cognitive process (such as SpeedIndex, user-PLT etc., see §2) have been proposed as a proxy of users' quality of experience (QoE).

At the same time, studies involving more advanced metrics are typically validated with rather small-scale experiments, either with

a small number of volunteers, or by using crowdsourcing platforms to recruit (cheap) labor and produce a dataset labeled with user opinion. Often, *videos* of Webpages rendering process is used (as opposite to actual browsing), with possibly very specific instruction (e.g., such as in A/B testing, by clicking on the fastest of two rendering processes) that are however rather different from the cognitive process in action during the typical user browsing activities. Additionally, such tests are carried on a limited number of fixed conditions, with a small heterogeneity of devices, OSs and browsers, and are not exempt from cheating so that ingenuity is needed to filter out invalid answers from the labeled dataset [24, 44]. Finally, because these tests are carried on a limited number of pages, it is possible to evaluate computationally costly metrics, such as those that require processing the visual rendering of the webpage, which would hardly be doable in the World "Wild" Web.

Our aim is instead to take a completely different approach and perform a large-scale study of a popular Web site in operation, by explicitly asking a fraction of users for feedback on their browsing experience. Clearly, the approach is challenging but it opens the possibility to gather more relevant user-labels, as they are issued from *real users of a real service*, as opposite to crowdworkers payed to play a game (e.g., find which video completes first as in A/B testing). We do so by launching a measurement campaign over Wikipedia, that at time of writing has gathered over 62k survey responses in nearly 5 months. We complement the collection of user labels with objective metrics concerning the user browsing experience (ranging from simple PLT [20] to sophisticated SpeedIndex [1]), and harvest several data sources to further enrich the dataset so that each user survey answer is associated with over 100 features. Summarizing our main contributions:

- first, we use survey data to characterize user satisfaction, finding that on average 85% of users are satisfied;
- second, we build a supervised data-driven model of user experience: despite our features include state-of-the art quality of experience metrics, we find that they fall short to model user QoE in operational settings;
- third, in spirit with research reproducibility, we plan to release both the collected *dataset* and our developed *code* as open-source, as we hope this can help the scientific community in refining its understanding of Web users' experience.

After overviewing the related work (§2), this paper describes our feedback collection process and dataset (§3), that we leverage to build a data-driven model of Wikipedia user experience (§4), finally discussing its current limits and directions to circumvent them (§5).

2 BACKGROUND

Assessment of Web users' quality of experience can be traced back to [36], that was among the first to adapt classic results of psycho-behavioral studies gathered in the *computer* domain [30] (in turn inspired by work by Weber and Fechner in the late 1800s), to the *computer-network* domain. This knowledge was later embedded into standards ITU-T G1030 [26, 39] (and models [23]) that encode the Weber-Fechner logarithmic [26, 39] (or exponential [23]) relationship between a stimulus (e.g., a delay) and its perceived impact (e.g., nuisance for Web users). However, while logarithmic models are valid for simple waiting tasks (e.g., file downloads), the case of interactive Web browsing is knowingly much more complex, as ITU-T G1031 [27] and [21] first pointed out.

Still, with few exceptions [15, 18, 44, 48] most studies still rely on simple metrics such as the Page Load Time (PLT) to assess the expected impact of new Web protocols [22, 40, 45, 47], Web accelerators [29, 46] and devices [34, 37]. While reducing delay is clearly a desirable objective, it is however unclear if (and by how much) a latency reduction translate into a better perceived experience, which is the ultimate goal of the above studies. In other words, while the importance of *delay* in human perception is agreed upon, the exact relationship between the Web response time and user satisfaction appear much less clear than it appeared to be [33], and motivated a proliferation of new metrics proposals and validation studies attempting at going beyond PLT.

Web QoE metrics: As we are interested in measuring browsing experience on individual pages, *engagement* metrics such as those used in [11, 31] are clearly out of scope. As such, objective metrics of interest for Web user QoE can be divided in two classes. On the one hand, there are metrics that either *pinpoint precise time instants*: notable examples include the time at which the DOM is loaded or becomes interactive (TTI), the time at which the first element is painted (TTFP) or the time when the Above The Fold (ATF) portion of the page is rendered [14] etc. Most of these metrics are available from the browser navigation timing [20], and are easy to include (though not necessarily relevant) as proxy of user experience.

On the other hand, there are metrics that *integrate all events of the waterfall* representing the visual progress of the page, such as SpeedIndex [1] and variants [2, 12, 24], that have received significant attention lately. Denoting with $x(t) \in [0, 1]$ the visual completeness ratio of a page, metrics in the SpeedIndex family are defined as the integral of the residual completion $\int (1 - x(t))dt$ and differ in the way they express $x(t)$. Initial definitions in this family required capturing movies of the rendering process [1], or to further use similarity metrics SSIM [24], making them difficult to use outside a lab environment. To counter this issue, simple approximations such as the ObjectIndex/ByteIndex [12] that merely count the fraction of objects/bytes received (over the total amount), or as the RUM SpeedIndex (RSI) [2] that use areas of rectangles for objects as they are painted on screen (over the total screen size) have been proposed. In this paper, we use RSI, which is among the most advanced Web QoE metrics considered to be the current industry standard. Finally, while we are aware that more complex approaches involving the spatial dimension (i.e., eye gaze) also exist [15, 28], we prefer to leave them for future work (cfr §5).

Metrics Validation: At the same time, the above metrics suffer from a limited validation with user feedback. Typical approaches are to crowdsource the validation with A/B testing [24, 44], or by performing experiments on real pages in controlled conditions [18, 33, 35]. Both approaches have their downsides. Controlled experiments with real HTTP server/clients and emulated network conditions for a more faithful and interactive browsing experience, but are harder to scale, topping to few hundreds users and few thousands data points [18]. A/B tests try to circumvent this limit, but introduce other limitations. First and foremost, A/B testing is hardly representative of Web browsing activity, since crowdworkers are instructed to select which among two videos, that they are passively screening side-by-side and that correspond to two different Web rendering processes, appears to finish first – whereas it is known that even for a simple Web browsing task such as information seeking, already different types of searches are rather different from the user standpoint in terms of cognition, emotion and interaction [32]. In other words, these experiments inform us that humans can perceive differences in these rendering processes, however they fail to signify if these perceptible rendering changes would impact the user satisfaction through the course of a normal browsing session.

The time at which users consider the process finished is denoted as user-PLT (uPLT)[44] or Time To Click (TTC) [24] and is often used as a ground truth of user perception. Yet, when users select a uPLT in [44], they are proposed with similar frames at earlier times, which has the beneficial effect of clustering answers and make uPLT more consistent at the price of possibly inducing a bias. Similarly, [24] employs SpeedIndex and TTC to forecast which among the left or right video was selected by the user at time TTC: the classifier in [24] is accurate in predicting which of the two videos is perceived as fastest by users. Yet, findings in [24] are not informative about whether the user would have been dissatisfied from the slower rendering had s/he been truly browsing.

Our contribution: To get beyond these limitations, in this work we are the first to query, at scale, Web users for their feedback on the quality of their browsing experience, which, to the best of our knowledge, has not been attempted before on the wide and wild Web. Instead of collecting user feedback on a 5-grade ACR scale, we ask for a (slightly more than) binary feedback (see §3), which let us formulate a simple (yet hard, see §4) binary classification problem.

Compared to recent literature , we are the first to involve a large number of real users (62k from 59k distinct IP addresses) accessing a diverse set of pages (46k Wikipedia pages, which are more likely similar among them than the set of different Websites used in other studies), gathering over 62k user responses overall (more than twice the survey responses collected in similar large-scale Wikipedia studies [41]). Particularly, whereas most of the studies involving lab volunteers & crowdworkers employ a single browser and hardware (since crowdworkers are shown videos rendered with a single browser and hardware combination) on a relatively small set of synthetic controlled network conditions, in our dataset we observe 45 distinct browsers software used on over 2,716 hardware devices on 3,827 ISPs – a significant change with respect to artificial and controlled lab conditions, which make the dataset of particular interest.

3 USER FEEDBACK COLLECTION

Wikipedia is, according to Alexa [3], the 5th most popular website, with over 1 billion monthly visitors, that spend over 4 minutes over 3 pages on average per day on the site. We engineer a survey that is triggered after the page ends loading and collects user feedback (§3.1), that we augment with additional information (§3.2).

We note that, while this paper is not the first in leveraging Wikipedia surveys in general (see e.g., [41]) this is the first to gather user feedback on quality of Web browsing experience from operational websites, for which we believe releasing the dataset can be valuable for the community. To make sharing of the dataset possible¹, we take special care into making user and content deanonymization as hard as possible, without hurting the dataset informative value as much as possible (§3.3). In this section, we also perform a preliminary assessment of the collection methodology, to confirm the absence of bias in the response process (§3.4).

3.1 Technical aspects of the survey collection

Due to limitations in Wikimedia's caching infrastructure, the survey is injected into the page via client-side code. Wikimedia continuously collects navigation timing performance of a randomly selected sample \mathcal{T} of page views (less than 1 every 1,000 pageviews); the survey is displayed to a randomly selected sub-sample \mathcal{S} of this population (less than 1 every 1,000 of the pageviews with navigation timing information) and only part of the surveys do receive an answer \mathcal{A} . Since $\mathcal{A} \subset \mathcal{T}$, several features (that we detail in §3.3) related to page loading performances are also available for pages sampled in the survey responses.

The survey appears on Russian, French and Catalan Wikipedias, as well as English Wikivoyage, and it is displayed in the appropriate language to the viewer. We collect the survey on mobile & desktop version of the site (but not on the mobile app). Instead of asking users a 5-grade Absolute Category Ranking (ACR) score, we opt for a simpler yet still very relevant feedback [25], as users can respond with a *positive*, *neutral* or *negative* experience. Neutral feedback is meant for, e.g., users that have no honest opinion, as well as users who were not paying attention during the rendering, or users that do not understand the question, etc. to avoid biasing the results (§3.4). For the sake of completeness, a snapshot of the survey as it is rendered for English readers is reported in Fig. 1.

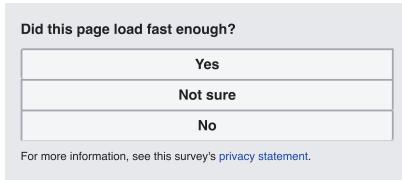


Figure 1: Appearance of the Survey in the English Wikipedia (answer order is randomized).

The survey is injected in the DOM after the page finished loading (i.e., when `loadEventEnd[20]` fires). In order to give the survey visibility, it is consistently inserted in the top-right area of the

¹The Wikimedia feature vetting process is still ongoing at time of writing.

Table 1: Collected corpus of Wikipedia users' QoE feedback.

Period	May 24th – Oct 15th		
No. of survey requests	$ \mathcal{S} =$	1746799	
No. of survey answers	$ \mathcal{A} =$	62740	$ \mathcal{S} / \mathcal{A} =3.6\%$
No. of positive answers	$ \mathcal{A}^+ =$	53208	$ \mathcal{A}^+ / \mathcal{A} = 84.8\%$
No. of neutral answers	$ \mathcal{A}^0 =$	4838	$ \mathcal{A}^0 / \mathcal{A} = 7.7\%$
No. of negative answers	$ \mathcal{A}^- =$	4694	$ \mathcal{A}^- / \mathcal{A} = 7.5\%$

wiki article, ensuring that it typically appears above the fold, with randomization of answer order. However, as the users can freely browse the page before the survey appears, it might be out of sight when it's injected in the DOM, which is why we also record the time elapsed between the `loadEventEnd` and the moment the user sees the survey. Also users that are shown the survey are free *not* to respond to the survey, or might as well respond very late (e.g., possibly browsing to other tabs in the meanwhile).

Overall, users responded as reported in Tab.1 to about 3.6% of the over 1.7M surveys that have been displayed in the period, for a total of over 62k answers: 84.8% of the users respond positively to the survey with an almost equal split of the remaining answers to a neutral (7.7%) or negative (7.5%) grades.

3.2 Collected features

We enrich the collected corpus with external sources that are instrumental to the purpose of feedback prediction (§4). A terse summary of the metrics collected is reported in Tab.2, while rationales of the selection for those publicly available is given in §3.3.

Page: For each page, we record 15 features that concerns it (e.g., its URL, revision ID, size, etc.) and that thus are critical from a privacy point of view. We additionally record the time lapse at which the survey is shown to users, which is instead innocuous.

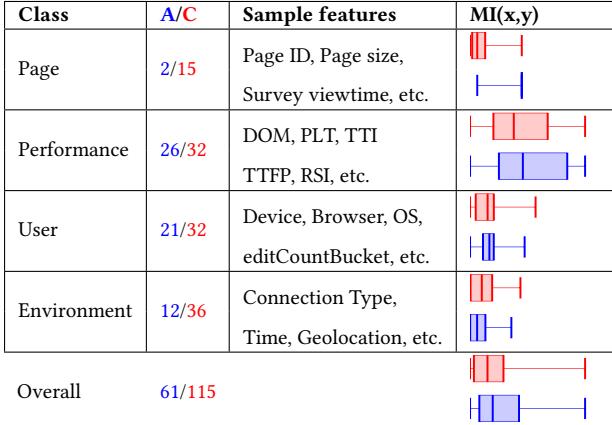
Performance: Since $\mathcal{S} \subset \mathcal{T}$, then all the 32 navigation-timing performance-related metrics (such as DOM, PLT, TTI, TTFP, connection duration, number of HTTP redirect, DNS wait time, SSL handshake time, etc.) are also collected. Finally, we compute the page download speed which is a simple, yet non linear, transformation of page size and connection duration. These informations are specific to page views, and are less critical to be shared.

User: The 32 collected user-related metrics include the browser, device and OS families. Additionally, we know whether users are logged in Wikipedia, if they are accessing Wikipedia through a tablet device and the number of edits that users have made (coarse bins). These informations are of course highly critical.

Environment: The 36 environmental collected features pertain time, network, geolocation and techno-economic aspects. With the exception of time information, which are directly available from the survey query, we extensively use external data sources to extract environmental features.

As for the network, we leverage MaxMind [4] for IP to ASN and ISP mappings and for geolocation at country (and city) granularity. ISP and ASN mappings are potentially interesting as it can be

Table 2: Summary of the available/collected (A/C) features that are associated to each users' survey response. The mutual information between the survey answer and A/C features in the class is reported as a boxplot.



expected that performances (for the same access technology) vary across ISPs (access technology is also available for about 2/3 of the samples). Concerning geolocation, whereas databases are known not to be reliable for city-level geolocation of server addresses [38], they are generally sufficiently accurate for resolving customer IP addresses, and especially when only ISO-3166-2 country-level precision is required. Country-level precision also allows us to relatively compare performances across users in the same environment, i.e., we normalize the page download speed with respect to the median per-country speed observed in our dataset (in terms of ratio, absolute and relative error).

Additionally, ties between country wealth and network traffic volumes have been established in the literature (particularly, deviation from expected volume [42]): it is thus worth investigating whether there also exist ties between wealth and users' impatience. We use the Gross Domestic Product (GDP) information made available by the World Bank Open Data project [5]. The per-country economic features we consider (namely, per-country GDP, country GDP rank, per-country per-capita GDP, etc.) are expressed in terms of Geary-Khamis dollars, which relate to the purchasing power parity, i.e., how much money would be needed to purchase the same goods and services in two countries. The rationale in so doing is that, albeit Web users perception is tied to psychophysics laws [39], there may be environmental conditions that tune this law differently in each country. For instance, a fixed amount of delay (the stimulus) may have a smaller perceptual value to users of countries with poor Internet access which GDP-related features might capture: e.g., in other words, one can expect users in a high-GDP country to have better average performance and thus be more impatient than users from a low-GDP one.

Finally, we expect user-home gateways [43] and particularly end-user devices [19, 34] to have a direct impact on the overall performance. As such, we complement the ISP-level view with a device-level information. Particularly, we harvest the Web [6] to find techno-economic information about user devices and in

particular, collect device CPU, memory and pricing² information. Intuitively, this information complements the per-country GDP information as, e.g., there may be further perceptual differences between users with a costly smartphone in low-GDP vs high-GDP countries. We recognize that device CPU and memory specs are only an *upper-bound* of the achievable performance, as it is the mixture of applications installed and running on a device that determine the amount of *available* CPU and RAM resources, from which user perception will be ultimately affected [19, 34]. Missing this information on a per-sample basis, we attempt to at least construct the per-device statistics, by considering navigation timing information of a large representative sample of Wikipedia users. Particularly, we consider the month of August 2018 during which we observe over 30 million navigation time samples from 29,336 different devices, including all 2,716 devices in our survey. We then construct *deciles* of per-device performance (e.g., of page load time and similar timing information): indeed, it can be expected that users of knowingly slow devices be less impatient, which this additional data source could provide.

3.3 Ethics

The dataset we collect contains obviously sensitive information allowing to deanonymize Wikipedia visitors (such as IP addresses, version of their browser and handsets), as well as linking them to the content they visited (e.g., page, revision ID, time of their visit, etc.). Despite the dataset release policy explicitly forbids user deanonymization, in the interest of respecting personal privacy we have to obscure information so to render user deanonymization as hard as possible, while still allowing meaningful information to be extracted from the data.

Method: Specifically, we opt for an approach where we transform data in a non bijective way (e.g., IP to ASN and ISP mappings that provide network-related properties, while preventing user deanonymization at the same time), or aggregate at a sufficiently coarse grain (e.g., country-level geolocation; obfuscation of browser major/minor version; aggregation of unpopular devices, etc.). For the same reason, we decide to aggregate time-related information at a coarse-grain (hour-level) and drop most content-related features (e.g., page ID). We quantize the page size with a resolution of 10KB, to also make it hard to reverse-engineer which page was visited. We maintain most of the navigation timing related performance, that have the highest mutual information, which we obfuscate wherever necessary (e.g., given that with precise PLT and download speed one could easily reverse engineer the page size, and thus the content, we quantize the download speed in steps of 100Kbps). We point out that, since the Wikimedia feature vetting process is still ongoing at time of writing, shall the bag of available features ultimately differ from the one described here, it will be properly documented.

Results: As a consequence, this loss of information potentially has an impact on the global prediction accuracy, which we assess in §4: at the same time, from results presented in Tab. 2, we can

²Note that we collect pricing information at the time of our query, and not at the time when the device was actually bought; we also ignore price differences among countries, and per-ISP offer bundles.

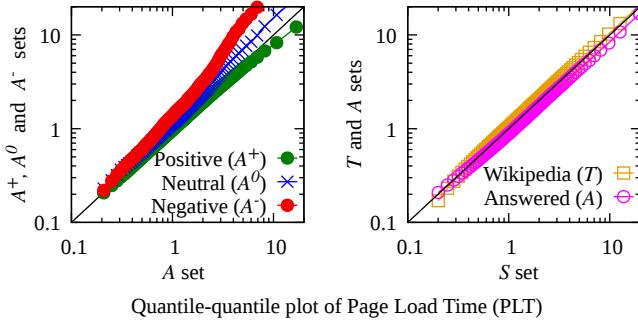


Figure 2: Quantile-quantile plot of PLT statistics for different sets ($\mathcal{T} \supset \mathcal{S} \supset \mathcal{A} = \mathcal{A}^+ \cup \mathcal{A}^0 \cup \mathcal{A}^-$).

expect this effect to be rather limited. Indeed, Tab. 2 reports the number of the publicly available out of the collected (A/C) features in each class (second column), and additionally reports boxplots of the mutual information $MI(x, y)$ between features in the class and the survey answer (last column). MI expresses the amount of information (in bits) that can be obtained about the survey answers through the observed variable. Tab. 2 shows that, while we only report about half of the collected features, the available features overall have a *higher mutual information* (particularly, note that the 25th percentile, median and 75th percentile are higher in the available feature set, and the feature having the maximum MI is also exported). Particularly, under this angle it is fortunate that features belonging to the performance class, which are those exhibiting the highest mutual information with the user grade, are also the least critical to share.

3.4 Validity of the collection methodology

Despite our care in engineering the survey questioning process, we cannot exclude a-priori the existence of bias in the user survey answer process. For instance, users might refrain to answer when the page loading experience was positive, and be more willing to express their opinion in case of bad experience, which would lead to under-estimate the user satisfaction.

To assess whether our survey collection methodology yields to such (or other) biases, we compare three sets of page view experiences, namely (i) the set \mathcal{T} where we record navigation timing information from the browser (ii) the set \mathcal{S} where users have been *shown* the survey (iii) the set \mathcal{A} where users have actually *answered* to the survey. Finally, we further slice the set of answered surveys \mathcal{A} according to the answer in three additional datasets with (iv) positive \mathcal{A}^+ , (v) neutral \mathcal{A}^0 and (vi) negative \mathcal{A}^- grades.

Among the numerous features we collect, without loss of generality we now limitedly consider the Page Load Time (PLT) distribution. Since $\mathcal{S} \subset \mathcal{T}$ is selected with uniform random sampling, by construction we have that \mathcal{S} and \mathcal{T} are statistically equivalent as far as individual features, such as PLT, are concerned. However, in case where users *decision* to answer to the survey would be biased by the performance of the page, then the PLT statistics should differ among the set of displayed \mathcal{S} vs answered \mathcal{A} surveys. The right-side of Fig. 2 reports a quantile-quantile (QQ)-plot of the empirical

PLT distribution, using quantiles of \mathcal{S} on the x-axis and \mathcal{T} , \mathcal{A} on the y-axis, from which one can clearly remark the absence of such bias.

Conversely, one would expect that, shall the PLT affect the actual grading of the browsing experience, then PLT statistics should differ among the $\mathcal{A}^+ \cup \mathcal{A}^0 \cup \mathcal{A}^- = \mathcal{A}$ sets. This is shown in the left-side of Fig. 2, comparing the quantiles of the answer set \mathcal{A} on the x-axis to its per-grade slices on the y-axis. Several remarks are in order. First, it can clearly be seen that browsing experience with negative scores fall above the equality line, confirming as expected that pages with longer download time yield to poor experience. Second, similar considerations hold for neutral (slightly above) and positive (slightly below) answers, although they are less visible – in part, this is due since positive grades represent the bulk of the answers $|\mathcal{A}^+|/|\mathcal{A}| = 84.8\%$, for which the PLT statistics of \mathcal{A}^+ and \mathcal{A} are mechanically more similar (we will take care of class imbalance when appropriate later on in §4). Third, notice that a range of PLT values are present in the set of positive, neutral and negative answers, indicating as expected that the PLT alone cannot fully capture user perception.

4 USER FEEDBACK PREDICTION

Problem formulation: Disregarding the neutral scores, we now build data-driven models to forecast user answers. This allows to turn the problem into a binary classification one: this simple formulation enables immediate and intuitive statements of performance objective, that we express in terms of the classic information retrieval metrics.

Clearly, from an operational standpoint a *conservative* estimation of user satisfaction is preferable. Indeed, the service operator wants to avoid that a malfunctioning service that is truly affecting user experience goes undetected, as when the ratio of dissatisfied users increases above a given level this can prompt alert to repair or ameliorate the service. In our settings, conservative prediction results translate into *maximizing the recall of negative scores*.

Reference classification results: Given the class imbalance, we have to preliminarily downsample the dataset³: indeed, given that after discarding neutral scores 92% of the users are satisfied, a naïve 0-R classifier that just learns the relative frequency of the scores and systematically answers with the majority class, would achieve 0.92 accuracy – but would entirely miss negative scores, having thus a null \mathcal{A}^- recall. Hence, a more appropriate baseline for recall of unsatisfied users is that of a uniform random selection.

Fig. 3 reports a confusion matrix, additionally highlighting the average accuracy, precision and recall of the unsatisfied users. Results are gathered on all the 115 collected features with a 20-trees random forest [13] on a 10-fold validation using 80% of the samples for training. We obtain very similar results with XGBoost [17], with a slightly higher accuracy but lower \mathcal{A}^- recall, which we do not report to avoid cluttering the pictures. Prediction outcome is clearly deceiving and only slightly better than the naïve baseline, despite the relatively large number of features collected: only 62% of the unsatisfied users are correctly captured, with a precision of

³We prefer to avoid the diametrically opposite approach of synthetically generating users score, which is in stark contrast with the very same nature of our work.

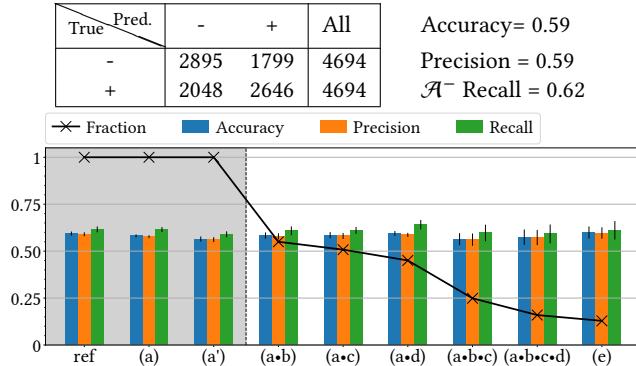


Figure 3: Classification results: Confusion matrix for all collected features as a reference (top) and performance obtained by limiting the (a) 61 available features, (a') 26 features of the performance class, and restricting the attention to (b) Chrome-only browser, (c) Russian population (d) Android OS and (e) top-1000 pages (and combinations thereof).

0.59. Interestingly, not shown in the picture, performance on the *remaining dataset*, i.e., the set of positive scores filtered out due to class imbalance, remains consistent with an accuracy of 0.65.

Feature subsampling: We next consider how the classification results change with respect to the above reference when considering only the 61/115 publicly available features, which is shown in Fig. 3-(a): as expected, since features in the available set are fewer but with better mutual information, classification results are practically unaffected. We reduce this set even further by only considering the 26 features of the performance class in (a'), which shows a slightly higher, but still very limited, reduction of classification performance: on the one hand, performance-related features consistently rank high in terms of Gini importance, though on the other hand they lack discriminative power for telling user answers apart.

Dataset subsampling: We finally spatially condition the dataset, investigating whether classification performance mechanically improves by reducing the heterogeneity in the dataset, in an attempt to recreate more homogeneous conditions as usually done in the lab studies. Particularly, (a-b) uses publicly available features and restrict the attention to the most popular browser, namely Chrome, considering both mobile and desktop flavors. In (a-c) we instead restrict to users of the prevalent country, and in (a-d) to Android users. We also combine these filters altogether (a-b-c) and (a-b-c-d), and finally consider (e) the top-1000 pages in our dataset. Clearly, conditioning the dataset implies that a smaller fraction of original dataset is available, which we also have to re-balance (solid black line in Fig. 3): in turn, confidence intervals for the metrics of interest increase for decreasing dataset fractions, which is expected. Yet, it is easy to gather that classification performances are only minimally affected in all the above cases, so that the state-of-the performance metrics we collect are apparently not enough to discriminate satisfied vs unsatisfied users.

5 DISCUSSION

In this paper we engineer, collect and (plan to) share at [7, 8] user survey scores pertaining to the quality of their Web browsing experience. Out of over 1.7 million queries, we gather over 62k answers corresponding to either positive (84.8%), neutral (7.7%) or negative (7.5%) experiences. We then develop data driven models to predict user scores: under this angle, the most important (and equally disturbing) takeaway is that it is surprisingly hard to predict even a very coarse-grained indication of user satisfaction. This can be tied in part to the lack of more informative indicators in our dataset, and also raises a number of interesting community-wide challenges, which we discuss next.

Collection and validation methodologies: We remark that this work is the first to collect user feedback from real users in real browsing activity, from an operational deployment. This is in stark contrast with most lab research, where volunteers or crowdworkers are exposed to a very limited heterogeneity (e.g., single device/browser), are not carrying on a browsing activity (e.g., A/B testing uses videos) and are not asked about their satisfaction but about other metrics as a proxy (e.g., which video finished first?). We argue that lab/crowdsourcing experiments and collection in the wild should *coexist*. Particularly, we argue that surveys such as those we are carrying on should be kept *running continuously*, as it is commonplace for VoIP applications that regularly poll their users for a QoE opinion. Operating continuously would lower barriers for further experiments [10], empower Website operators with a very relevant performance indicator for their service, informing them in near-real time about impact of new features deployment, and ultimately helping to ameliorate data-driven models.

RSI: not needed, or not enough ?: Concerning Web user QoE metrics, this study seems to suggest a poor discriminative power of the RumSpeedIndex (RSI) so as to predict users scores, at least for Wikipedia users. In part, this may be due to the structure of Wikipedia pages (where, e.g., text may be more prevalent than in other pages in the Alexa top 100 typically considered in similar studies, see §2), which nevertheless raises the question so as to whether it is possible to design more specific metrics that are better fit to the spatial structure of any given page.

Per-device statistics: Given that “computation activities are the main bottleneck when loading a page on mobile browsers” [34], collecting per-device statistics [19] seems a mandatory step. Unfortunately, average per-device performance we considered in this work are not telling enough, as they merely report the resource *upper-bound* (i.e., CPU and RAM capacity) as opposite to the *actual state* of the device (i.e., free RAM and available CPU cycles) corresponding to the page view that the user answered about – which could hopefully ameliorate prediction performance.

ACKNOWLEDGEMENTS

We wish to thank Leila Zia, Nuria Ruiz and Tilman Bayer for the fruitful discussion.

REFERENCES

- [1] [n. d.]. <https://sites.google.com/a/webpagetest.org/docs/using-webpagetest/metrics/speed-index>.
- [2] [n. d.]. <https://github.com/WPO-Foundation/RUM-SpeedIndex>.
- [3] [n. d.]. <https://www.alexa.com/topsites>.
- [4] [n. d.]. <https://www.maxmind.com/>.
- [5] [n. d.]. <https://data.worldbank.org/>.
- [6] [n. d.]. <https://www.gsmarena.com>.
- [7] [n. d.]. <https://webqo.e.telecom-paristech.fr>.
- [8] [n. d.]. https://meta.wikimedia.org/wiki/Research:Study_of_performance_perception_on_Wikimedia_projects.
- [9] 2010. World Wide Wait. <https://www.economist.com/science-and-technology/2010/02/12/world-wide-wait>.
- [10] Eytan Bakshy, Dean Eckles, and Michael S. Bernstein. 2014. Designing and Deploying Online Field Experiments. In *Proc. of the 23rd International Conference on the World Wide Web (WWW)*. <https://doi.org/10.1145/2566486.2567967>
- [11] Athula Balachandran, Vaneet Aggarwal, Shobha, He Yan, et al. 2014. Modeling Web Quality-of-experience on Cellular Networks. In *Proc. ACM MOBICOM*. ACM. <https://doi.org/10.1145/2639108.2639137>
- [12] Enrico Bocchi, Luca De Cicco, and Dario Rossi. 2016. Measuring the Quality of Experience of Web Users. In *Proc. ACM SIGCOMM, Internet-QoE Workshop* (2016). <https://doi.org/10.1145/3027947.3027949>
- [13] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (oct 2001), 5–32. <https://doi.org/10.1023/A:1010933404324>
- [14] Jake Brutlag, Zoe Abrams, and Pat Meenan. [n. d.]. Above the fold time: Measuring Web page performance visually. <http://conferences.oreilly.com/velocity/velocity-mar2011/public/schedule/detail/18692>.
- [15] Michael Butkiewicz, Daimeng Wang, Zhe Wu, Harsha V. Madhyastha, and Vyta Sekar. 2015. Klotski: Reprioritizing Web Content to Improve User Experience on Mobile Devices. In *Proc. of the USENIX Conference on Networked Systems Design and Implementation (NSDI)*.
- [16] Vint Cerf, Van Jacobson, Nick Weaver, and Jim Gettys. 2012. BufferBloat: what's wrong with the internet? *Commun. ACM* 55, 2 (2012), 40–47. <https://doi.org/10.1145/2076450.2076464>
- [17] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 785–794. <https://doi.org/10.1145/2939672.2939785>
- [18] Diego Da Hora, Alemany Arese, Vassilis Christophides, Renata Teixeira, and Dario Rossi. 2018. Narrowing the gap between QoS metrics and Web QoE using Above-the-fold metrics. In *Proc. Passive and Active Measurement Conference (PAM)*. <https://doi.org/10.1007/978-3-319-76481-8>
- [19] Malleesham Dasari, Santiago Vargas, Arani Bhattacharya, Aruna Balasubramanian, Samir R. Das, and Michael Ferdman. 2018. Impact of Device Performance on Mobile Internet QoE. In *Proc. ACM Internet Measurement Conference*. <https://doi.org/10.1145/3278532.3278533>
- [20] Zhiheng Wang (Ed.). 2012. Navigation Timing. <http://www.w3.org/TR/2012/REC-navigation-timing-20121217/>. In "W3C Recommendation".
- [21] Sebastian Egger, Peter Reichl, Tobias Hößfeld, and Raimund Schatz. 2012. "Time is bandwidth"? Narrowing the gap between subjective time perception and Quality of Experience. In *Proc. IEEE International Conference on Communications (ICC)*. <https://doi.org/10.1109/ICC.2012.6363769>
- [22] Jeffrey Erman, Vijay Gopalakrishnan, Rittwik Jana, and K. K. Ramakrishnan. 2013. Towards a SPDY'ler Mobile Web?. In *ACM CoNEXT*. 303–314. <https://doi.org/10.1145/2535372.2535399>
- [23] Markus Fiedler, Tobias Hößfeld, and Phuoc Tran-Gia. 2010. A generic quantitative relationship between quality of experience and quality of service. *IEEE Network* 24, 2 (2010), 36–41. <https://doi.org/10.1109/MNET.2010.5430142>
- [24] Qingzhu Gao, Prasenjit Dey, and Parvez Ahmad. 2017. Perceived Performance of Top Retail Webpages In the Wild: Insights from Large-scale Crowdsourcing of Above-the-Fold QoE. In *Proc. ACM SIGCOMM, Internet-QoE Workshop*. <https://doi.org/10.1145/3098603.3098606>
- [25] Tobias Hößfeld, Poul E Heegaard, Martín Varela, and Sebastian Möller. 2016. QoE beyond the MOS: an in-depth look at QoE via better metrics and their relation to MOS. *Quality and User Experience* 1, 1 (2016), 2.
- [26] ITU-T. 2014. Estimating end-to-end performance in IP networks for data application.
- [27] ITU-T. 2014. QoE factors in web-browsing.
- [28] Conor Kelton, Jihoon Ryoo, Aruna Balasubramanian, and Samir R Das. 2017. Improving User Perceived Page Load Time Using Gaze. In *Proc. of the USENIX Conference on Networked Systems Design and Implementation (NSDI)*.
- [29] Yun Ma, Xuanzhe Liu, Shuhui Zhang, Ruirui Xiang, Yunxin Liu, and Tao Xie. 2015. Measurement and Analysis of Mobile Web Cache Performance. In *Proc. of the 24th International Conference on the World Wide Web (WWW)*. <https://doi.org/10.1145/2736277.2741114>
- [30] Robert B Miller. 1968. Response time in man-computer conversational transactions. In *Proc. AFIPS Fall Joint Computer Conference*. ACM.
- [31] Ben Miroglio, David Zeber, Jofish Kaye, and Rebecca Weiss. 2018. The Effect of Ad Blocking on User Engagement with the Web. In *Proc. of the 27th International Conference on the World Wide Web (WWW)*. <https://doi.org/10.1145/3178876.3186162>
- [32] Yashar Moshfeghi and Joemon M. Jose. 2013. On Cognition, Emotion, and Interaction Aspects of Search Tasks with Different Search Intentions. In *Proc. of the 22nd International Conference on the World Wide Web (WWW)*. <https://doi.org/10.1145/2488388.2488469>
- [33] Fiona Fui-Hoon Nah. 2004. A study on tolerable waiting time: how long are web users willing to wait? *Behaviour & Information Technology* 23, 3 (2004), 153–163.
- [34] Javad Nejati and Aruna Balasubramanian. 2016. An In-depth Study of Mobile Browser Performance. In *Proc. of the 25th International Conference on the World Wide Web (WWW)*. <https://doi.org/10.1145/2872427.2883014>
- [35] Ravi Netravali, Anirudh Sivaraman, Keith Winstein, Somak Das, Ameesh Goyal, and Hari Balakrishnan. 2014. Mahimahi: a lightweight toolkit for reproducible web measurement. In *ACM SIGCOMM Computer Communication Review*, Vol. 44. ACM, 129–130. <https://doi.org/10.1145/2740070.2631455>
- [36] Jakob Nielsen. [n. d.]. Response Times: The 3 Important Limits. <https://www.nngroup.com/articles/response-times-3-important-limits/>
- [37] Ashkan Nikravesh, Hongyi Yao, Shichang Xu, David Choffnes, and Z. Morley Mao. 2015. MobiLyzer: An Open Platform for Controllable Mobile Network Measurements. In *Proc. ACM MobiSys*. <https://doi.org/10.1145/2742647.2742670>
- [38] Ingmar Poese, Steve Uhlig, Mohamed Ali Kaafar, Benoit Donnet, and Bamba Gueye. 2011. IP Geolocation Databases: Unreliable? *ACM SIGCOMM CCR* 41, 2 (2011). <https://doi.org/10.1145/1971162.1971171>
- [39] Peter Reichl, Sebastian Egger, Raimund Schatz, and Alessandro D'Alconzo. 2010. The logarithmic nature of QoE and the role of the Weber-Fechner law in QoE assessment. In *Proc. IEEE International Conference on Communications (ICC)*. <https://doi.org/10.1109/ICC.2010.5501894>
- [40] Sanae Rosen, Bo Han, Shuai Hao, Z. Morley Mao, and Feng Qian. 2017. Push or Request: An Investigation of HTTP/2 Server Push for Improving Mobile Performance. In *Proc. of the 26th International Conference on the World Wide Web (WWW)*. <https://doi.org/10.1145/3038912.3052574>
- [41] Philipp Singer, Florian Lemmerich, Robert West, Leila Zia, Ellery Wulczyn, Markus Strohmaier, and Jure Leskovec. 2017. Why We Read Wikipedia. In *Proc. of the 26th International Conference on the World Wide Web (WWW)*. <https://doi.org/10.1145/3038912.3052716>
- [42] Chris Smith-Clarke and Licia Capra. 2016. Beyond the Baseline: Establishing the Value in Mobile Phone Based Poverty Estimates. In *Proc. of the 25th International Conference on the World Wide Web (WWW)*. <https://doi.org/10.1145/2872427.2883076>
- [43] Srikanth Sundaresan, Walter de Donato, Nick Feamster, Renata Teixeira, Sam Crawford, and Antonio Pescapé. 2011. Broadband Internet Performance: A View from the Gateway. In *ACM SIGCOMM*. 134–145. <https://doi.org/10.1145/2018436.2018452>
- [44] Matteo Varvello, Jeremy Blackburn, David Naylor, and Konstantina Papagiannaki. 2016. EYEORG: A Platform For Crowdsourcing Web Quality Of Experience Measurements. In *Proc. ACM CoNEXT*. <https://doi.org/10.1145/2999572.2999590>
- [45] Xiao Sophia Wang, Aruna Balasubramanian, Arvind Krishnamurthy, and David Wetherall. 2014. How Speedy is SPDY?. In *Proc. of the USENIX Conference on Networked Systems Design and Implementation (NSDI)*.
- [46] Xiao Sophia Wang, Arvind Krishnamurthy, and David Wetherall. 2016. Speeding up Web Page Loads with Shandian. In *Proc. of the USENIX Conference on Networked Systems Design and Implementation (NSDI)*.
- [47] Torsten Zimmermann, Jan Ruth, Benedikt Wolters, and Oliver Hohlfeld. 2017. How HTTP/2 Pushes the Web: An Empirical Study of HTTP/2 Server Push. In *Proc. IFIP Networking*. <https://doi.org/10.23919/IFIPNetworking.2017.8264830>
- [48] Torsten Zimmermann, Benedikt Wolters, and Oliver Hohlfeld. 2017. A QoE Perspective on HTTP/2 Server Push. In *Proc. ACM SIGCOMM, Internet-QoE Workshop*. <https://doi.org/10.1145/3098603.3098604>