Promoting Positive Post-Click Experience for In-Stream Yahoo Gemini Users

Mounia Lalmas
Yahoo Labs
London, UK
mounia@acm.org

Janette Lehmann*
Freie Universität Berlin
Berlin, Germany
janette.lehmann@fu-berlin.de

Guy Shaked
Yahoo
Tel Aviv, Israel
gshaked@yahoo-inc.com

Fabrizio Silvestri
Yahoo Labs
London, UK
silvestr@yahoo-inc.com

Gabriele Tolomei
Yahoo Labs
London, UK
gtolomei@yahoo-inc.com

ABSTRACT

Click-through rate (CTR) is the most common metric used to assess the performance of an online advert; another performance of an online advert is the user *post-click experience*. In this paper, we describe the method we have implemented in Yahoo *Gemini*[1] to measure the post-click experience on Yahoo mobile news streams via an automatic analysis of advert landing pages. We measure the post-click experience by means of two well-known metrics, *dwell time* and *bounce rate*. We show that these metrics can be used as *proxy* of an advert post-click experience, and that a negative post-click experience has a negative effect on user engagement and future ad clicks. We then put forward an approach that analyses advert landing pages, and show how these can affect dwell time and bounce rate. Finally, we develop a prediction model for advert quality based on dwell time, which was deployed on Yahoo mobile news stream app running on iOS. The results show that, using dwell time as a proxy of post-click experience, we can prioritise higher quality ads. We demonstrate the impact of this on users via A/B testing.

Categories and Subject Descriptors

H.4.0 [**Information Systems Applications**]: General

Keywords

Mobile advertising; Native advertising; Advertising quality; Post-click experience; Dwell time prediction

1. INTRODUCTION

One of the main requirements driving the development of Yahoo Gemini, the Yahoo advertising platform, has been that of caring not only about the sheer number of clicks on ads but also about the user experience. In our case, user experience means to show advertisements[2] that will not discourage users to continue using the platform or, even more challenging, to engage them to click more on ads.

Following from this, we put forward an approach that analyses ad landing pages, and show how these can affect *dwell time* and *bounce rate*. Then we develop a model based on dwell time for predicting the quality of *native* Yahoo Gemini ads, which is currently deployed on Yahoo Gemini. We chose to deploy the version predicting dwell time as we wanted to serve ads on which users spend time, meaning not only they promote a positive post-click user experience but also have the potential of leading to a "conversion" (e.g., making a purchase). Analyzing one month data through A/B testing, we see that returning high quality ads, as measured in terms of the ad post-click experience, not only increases click-through rates by 18%, it has a positive effect on users, as we observe an increase in dwell time (+30%) and a decrease in bounce rate (-6.7%).

As we stated above, we focus in this paper on showing how our approach works on *native* ads. Let us now describe what a native ad is.

Feed-based layouts, or *streams*, are becoming an increasingly common layout in many applications, and a predominant interface in mobile applications. *In-stream* advertising has been increasingly emerging as a popular online advertising because it offers a user experience that fits nicely with that of the stream [9]. It is often referred to as *native advertising.*[3] In-stream (or native) ads have an appearance similar to that of the items in the stream, but are clearly marked with a "Sponsored" label or a currency symbol e.g. "$" to indicate that they are in fact adverts. Major news sites such as the New York Times, Yahoo News, and the Guardian integrate ads into their streams. Yahoo launched *Gemini* in 2013, an unified ad marketplace for mobile search and native advertising. Many of the ads served by Yahoo Gemini are in-streams ads. In this work, we focus on such ads served by Gemini on one of Yahoo mobile news streams.

[*] This work has been carried out while the author was an intern at Yahoo Labs.

[1] `https://advertising.yahoo.com/Blog/gemini-announcement.html`

[2] For simplicity, "advertisement" and "advert" are referred to as "ad".

[3] Sponsored search is also an instance of native advertising, as sponsored results have a similar look and feel to the organic results, while being clearly indicated as being ads.

KDD'15, August 10-13, 2015, Sydney, NSW, Australia.
© 2015 ACM. ISBN 978-1-4503-3664-2/15/08 ...$15.00.
DOI: http://dx.doi.org/10.1145/2783258.2788581.

We refer to the ad impression shown within the stream as the *creative* of the ad. A user decides if he or she is interested in the ad content by looking at its creative, depending on how attractive it is to users. After a user clicks on the creative he or she is redirected to the ad *landing page*, which is either a web page specifically created for that ad, or the advertiser homepage [2]. It is well known that the way user experiences a landing page, the *ad post-click experience*, is an important factor that, if properly measured, can help differentiating between a high quality and a low quality ad. Indeed, when the post-click experience is negative users may tend to click less on ads in the future, if not stop accessing the service at all with disruptive consequences in the overall number of visitors and therefore in total revenue.

Looking at the post-click experience is particularly important in the context of Gemini in-stream ads because (1) the creatives have mostly the same look and feel, as opposed to display ads, and what differs mostly is their landing pages, and (2) accounting for the post-click experience has led to increased performance in many areas, including sponsored search. In this paper, we develop and deploy a method to identify high quality ads, by analysing their landing pages and relate these to the ad post-click experience. Our focus is on mobile in-stream advertising as offered by Yahoo Gemini.

We measure the post-click experience through well-known engagement metrics, *dwell time* and *bounce rate.* Dwell time is the time between users clicking on an ad creative until returning to the stream; bounce rate is the percentage of "short clicks" (clicks with dwell time less than a given threshold). We demonstrate that they can be used as proxy of ad post-click experience. We also show that a negative post-click experience has a negative effect on user engagement and, in particular, future ad clicks.

2. RELATED WORK

Online advertising. Online advertising has been extensively studied in the context of display advertising [1, 22] and sponsored search [3, 23] for desktop users. Studies have been mostly focussed on predicting how an ad will perform according to various effectiveness measures [1, 12, 24].

Many efforts have been devoted to improve the matching between the textual content of web queries (in sponsored search) or web pages (in display advertising) and the ad textual content (creative and/or bid phrases and title). Most works aim at categorizing the ad textual content within some taxonomy and then matched it against the query category [2, 3] or a web page [5, 14, 18] where the ad could be shown. In particular, [2, 3] studied the effect of landing page categorization on user experience by defining a small taxonomy of landing pages. A year later, [6] studied the effect of landing page features in improving ad relevance in textual advertising. They show that augmenting the ad textual features with features from the content of the page increases relevance metrics. Finally, [11] focused on automatically categorizing display ad images into a taxonomy of relevant interest categories. They demonstrate the effectiveness of using image and textual features extracted from the ad landing page in predicting the category of the display ad. We also exploit landing page features to predict the quality of ads with respect to the post-click experience.

Users spend an increasing amount of their time online through their mobile devices. This presents unique opportunities for advertisers interested in promoting their products beyond the desktop. Previous studies have investigated the degree in which mobile advertising is accepted by users [4] and how users perceive display advertisements on mobile [8]. Findings suggest that personal relevance of the ad is not as important as matching the ad topics to the content of the page, and that utility, context and trust are crucial for acceptance of mobile advertising. Efforts have been put in building models to predict when to show an ad [19, 20]. Finally, as highlighted in [9, 21], the dominance of the feed-based structure on mobile makes pop-ups and banners impractical, whereas in-stream ads provide an optimal format as they are seamlessly incorporated to the main feed, thereby promoting relatively similar experience across all ads. This is not the same for the ad landing pages, as advertisers have total freedom in how they design these, which can vary greatly in terms of quality. In this paper, we study how an ad post-click experience, as measured with dwell time and bounce rate, can inform about the quality of its landing page.

Ad quality measures. Several measures have been proposed to evaluate the "performance" of an ad in terms of the user experience. The most common measure is the ad click-through rate (CTR), which is the number of times the ad was clicked out of the number of times it has been shown (number of ad impressions) [1]. The higher the CTR the better the ad is considered to perform; it attracts the users to click on it. However, CTR does not account for how users experience the ad when they land on the ad site, namely their *post-click experience*, for which other measures are better suited. A positive experience with an ad landing page increases the probability of users "converting" (e.g., purchasing an item, registering to a mailing list, or simply spending time on the site building an affinity with the brand) [3, 22]. Even if no conversion occurs, publishers are keen to serve high quality ads, as doing otherwise may affect user long-term engagement: loosing users would mean less clicks on ads, and lower engagement with the site, which in turn negatively impact revenue. In this paper, we indeed show that when users experienced high quality ads post-click, they were more likely to clicks on an ad in the future.

A positive post-click experience does not necessarily mean a conversion. Therefore, although the former can be measured by the latter (e.g., a high conversion rate), this way of assessing the quality of the post-click experience is restrictive. There may be many reasons why conversion does not happen, independent of the quality of the ad served to users. In addition, conversion rate information is not always available. A good proxy of the post-click experience is the time a user spends on the ad site before returning back to the publisher site: "the longer the time, the more likely the experience was positive". The two most common measures used to quantify time spent on a site are *dwell time* [27] and *bounce rate* [23]. These measures have been used as proxies of post-click experience in online advertising and organic search, e.g., to improve the performance of ranking algorithms [12], as well as in recommender systems, e.g., to estimate the relevance of an item to a user [26]. In this work, we show that these are also good proxies of post-click experience in both the mobile and desktop contexts.

Finally, [3] showed that conversion rates differ significantly depending on the type of landing page. Also, [6] showed that landing pages could be leveraged to better select which ads to return to users in sponsored search. In the context of mobile advertising, whether the landing page of an ad is mobile-optimised or not was shown to affect post-click experience [16]. Our research adds to this body of work by analysing other features of landing pages for mobile advertising, and how these help predict user post-click experience.

3. MEASURING POST-CLICK EXPERIENCE

We show that dwell time and bounce rate are good proxy measures of ad post-click experience.

Datasets and metric definitions. We randomly sampled 4,000 ads from a large set of ads served on Yahoo homepage stream in March 2014 both on desktop and mobile. Although our focus is on mobile devices, we look at both mobile and desktop to demonstrate that dwell time and bounce rate are good measures to act as proxy of ad post-click experience. We defined dwell time and bounce rate as follows:

- The average *dwell time* is the average time between users clicking on the ad and returning to the stream.
- The *bounce rate* is the percentage of ad clicks whose dwell time, unless otherwise specified, is lower than or equal to a fixed threshold.

The threshold used to determine the bounce rate on mobile was set to 5 seconds; this was empirically selected based on the dwell time distribution, which showed a "plateau" just around that value. The threshold on desktop instead was set to 12 seconds; this was chosen so to align with the threshold on mobile, by picking the value which corresponds to the same cumulative frequency of "bouncy" clicks on both dwell time distributions. Both these thresholds fall into the range between 5 and 60 seconds proposed in [15].[4]

In addition, we removed all clicks with a dwell time higher than 10 minutes, as these clicks may correspond to cases when users left the mobile or desktop device and came back later. By doing so we removed 1.74% of the total ad clicks. Finally, we only considered ads with at least 10 clicks.

Dwell Time as a Proxy of Ad Post-Click Experience. The fact that a user takes time to return to the news stream after clicking on an ad seems a good indicator that the experience is positive: the user browsed the site, maybe converted (e.g., purchased a product, registered to the site, shared the page, etc.), and finally went back to the stream. We study the extent to which higher dwell time indeed reflects a positive experience.

We used a random sample of 200K ad clicks on the mobile stream for which we have records of a click on the ad site. For desktop, we matched the page views associated with the ad clicks to those contained in the Yahoo toolbar's browsing data,[5] which results in page views from 30K ad clicks.

For both datasets, the Spearman's rank correlation coefficient between the number of clicks on the ad site and the ad dwell time is, respectively, 0.65 for mobile and 0.54 for desktop; the higher the dwell time, the higher the number of clicks on the ad site. Assuming that clicking on the ad site suggests a positive "ad experience", high dwell time is indicative of a positive post-click experience. The probability of a second click as the percentage of users who clicked on a link on the ad landing page, for given dwell time values, is plotted in Figure 1. This probability increases with the time spent *until* the users return to the news stream for both mobile and desktop, further indicating that dwell time is a good proxy of post-click experience. It is also worth noticing that to get the same probability of a second click on mobile and on desktop, the dwell time has to be far larger in the latter than in the former. This suggests that dwell time is generally greater on desktop than on mobile.

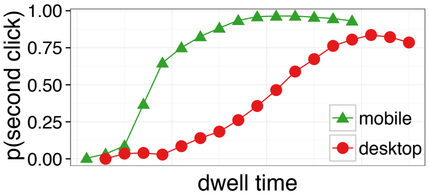


Figure 1: Post-click experience. The probability of a second click given dwell time. The x-axis values (log-scale) are removed for confidential reasons.

Dwell time is *not* the time spent on the ad site, but the time between the click on the ad until the user returns to the stream. It can happen that users visit other sites during their session before returning. With the Yahoo toolbar dataset, we are able to look into this. We saw that the higher the dwell time, the higher the probability that users visited other websites. However, for all ad clicks with a dwell time up to 3 minutes, this happens for only 7.4% of the clicks. For dwell time higher than 3 minutes, this percentage increases to 23.3%. Therefore, in general dwell time is a good proxy of users spending time on the ad site, with longer dwell time reflecting a positive experience with the ad site. Next, we relate this to the long-term effect on user engagement.

Long-term User Engagement. We now focus on mobile. We investigate the effect of the ad post-click experience, as measured by dwell time and bounce rate, on long-term user engagement. In other words, we examine how users are affected when they experience ads on which they spend little time (reflecting a negative ad experience) compared to the opposite (reflecting a positive ad experience)

We divided our dataset into three time-periods, covering a four-week period of user interaction with the mobile stream:[6]

- user *pre-engagement* in a given two-week period;
- user *post-engagement* in the following two-week period;
- user *ad-click-activity* in the last three days of the pre-engagement period and the first three days of the post-engagement period.

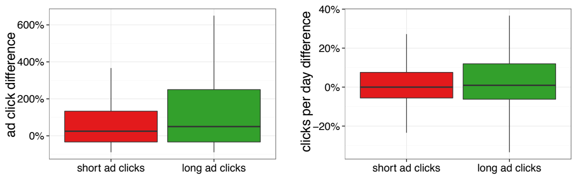
Our objective is to compare the pre- and post-engagement periods depending on the ad-click-activity between the two.

We used the *ad-click-activity* dataset to distinguish between a positive ad post-click experience and a negative one. For each ad a, we calculate its mean dwell time $dt_m(a)$ and its standard deviation $dt_{sd}(a)$. Any click c on ad a with $dt_c(a) \leq dt_m(a) - 0.25 \cdot dt_{sd}(a)$ is referred to as a *short click*, and any click c on ad a with $dt_c(a) \geq dt_m(a) + 0.25 \cdot dt_{sd}(a)$ is referred to as a *long click*. Here, $dt_c(a)$ is the dwell time on ad a for click c. These definitions account for the fact

[4] We also experimented with other thresholds in this range. Similar results were obtained.

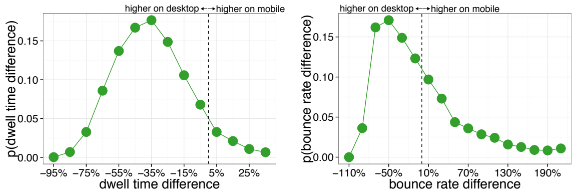
[5] A sample of users gave their consent to provide browsing data via the Yahoo toolbar, which is a browser extension providing additional functionalities and direct access to selected websites.

[6] The study was carried out with other datasets and led to the same conclusions.



(a) Changes in adClicks. (b) Changes in clicksPerDay.

Figure 2: Changes in engagement depending on whether the users experienced short or long clicks.



(a) Difference in dwell time. (b) Difference in bounce rate.

Figure 3: Distributions of the differences in post-click experience between mobile and desktop.

Table 1: Changes in engagement for short and long ad clickers. We report the average and median ($avg|median$) of the ad clicks and clicks per day difference, and the p-values.

	Short ad cl.	Long ad cl.	p-value
AdClicks diff.	90.3%|25.0%	122.6%|50.0%	0.002
ClicksPerDay diff.	2.4%|0.0%	5.5%|0.9%	0.000

that ads differ in terms of their average dwell time. For example, we saw that ads related to beauty products have on average a higher dwell time than those related to finance, simply because the ad experience is different (reading about a product versus registering an interest).

For all users that clicked on at least 3 ads, users with only short clicks were said to have had a negative experience ($shortAdClicker$), whereas those with only long clicks were said to have a positive experience ($longAdClicker$). Having a minimum of 3 clicks allow us to select users that have experienced enough ads to be affected by them. These resulted in two sets of similar size, around 800 users each.

We use two metrics to measure pre- and post-engagement:[7]

- $adClicks$ is the number of ad clicks of a user over the period considered. This metric shows the effect of the ad post-click experience on that user future ad clicks.
- $clicksPerDay$ is the average number of pages a user is viewing within each day over the period considered. This metric shows the effect of the ad post-click experience on the user future interaction with the stream.

We define the change in the engagement between the pre- and post-engagement time periods as follows:

$$\text{eng}_d = (\text{eng}_{\text{post}} - \text{eng}_{\text{pre}})/\text{eng}_{\text{pre}}$$

where eng is either $adClicks$ or $clicksPerDay$. A value above (below) 0 indicates that post-engagement increased (decreased) compared to pre-engagement, and the extent of the increase (decrease). Figure 2 shows the distribution of the two metrics for the two user groups ($shortAdClicker$ and $longAdClicker$). We also report the average and the mean of the two metrics in Table 1. We use a two-sample Kolmogorov-Smirnov test to check whether the distributions are different. The p-values are reported in Table 1.

For both user groups, the ad click activity is increasing. This is likely to reflect that users in both groups are becoming more engaged with the stream and as a result are more likely to click on ads. However, the median $adClicks_d$ for the short ad clicker group is 25.0%; this value is 50.0% for the long ad clicker group. That is, the increase in ad clicks (both in terms of median and average, and distribution) for the long ad clicker group is higher, indicating that a positive ad post-click experience is leading to more ad clicks. The difference is statistically significant (p-value < 0.01).

The metric $clicksPerDay_d$ has a median close to 0.0% for both user groups. This suggests a similar trend in engagement with the stream, some users becoming more engaged, while others becoming less engaged. Looking at our dataset in more depth, we could identify users getting more engaged with the stream as time passed, and users that were very engaged with the stream. As such their future engagement could not increase further (reaching a certain plateau). However, when looking at the average values and the distributions of $clicksPerDay_d$, we see a larger increase for the long click group, compared to the short click group (5.6% and 2.4%, respectively), suggesting a larger increase in engagement with the stream for users in the positive post-click experience group. The difference, although small compared to $adClicks_d$, is significant (p-value < 0.01).

To conclude, using dwell time to measure the ad post-click experience, we showed that a positive experience had a strong effect on users clicking on ads again, and a small effect on user engagement with the stream. The effect is statistically significant for both. Thus, not only can dwell time be used to measure an ad post-click experience, ensuring that high quality ads are served to users is important for long-term engagement. We investigate further the two post-click measures, in particular what they can tell about the mobile post-click experience compared to the desktop one.

Mobile vs. Desktop. We examine the difference in the ad post-click experience, measured by dwell time and bounce rate, between mobile and desktop. First, we compare whether dwell time and bounce rate of an ad differs between the two.

Users experience ads differently depending on the device they are using. In fact, using Spearman's rank correlation coefficient ρ, for dwell time we obtain $\rho = 0.50$; this value is even smaller for bounce rate with $\rho = 0.23$.[8]

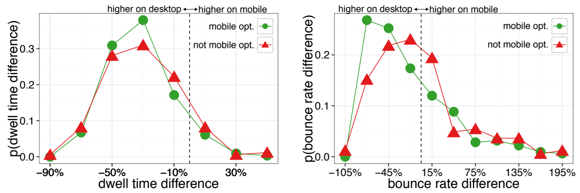
Next, we calculate for each ad the difference in percentage of their dwell time and bounce rate, when shown on desktop compared to mobile (val refers to dwell time or bounce rate):

$$d_{val} = (val_{mobile} - val_{desktop})/val_{desktop}$$

The distributions of the percentage differences are plotted in Figure 3, (a) for dwell time and (b) for bounce rate. For 92.9% of the ads the dwell time is higher on desktop than

[7] Similar results were observed with other metrics.

[8] Similar correlations were observed when restricting to ads with at least 50 clicks. The correlations are 0.63 and 0.29, respectively.



(a) Difference in dwell time. (b) Difference in bounce rate.

Figure 4: Differences in ad post-click experience between mobile and desktop depending on whether the landing page is mobile-optimised or not.

on mobile (Figure 3(a)). This is not surprising, as browsing time on mobile has been shown to be shorter generally outside advertising [25]. The highest decrease in dwell time from desktop to mobile is by 35.0%. From Figure 3(b), we observe that 64.1% of the ads have a higher bounce rate on mobile than on desktop, which is a lower percentage than for dwell time. The highest decrease of bounce rate from desktop to mobile is by 50.0%, which is slightly higher when compared to dwell time. The distribution is skewed to the left however, indicating that many ads have a large increase in bounce rate when shown on mobile. For instance, 18.9% of the ads have a bounce rate increase higher than 50.0%.

Whereas the dwell time differences between desktop and mobile may reflect the browsing behaviour on these two devices (shorter sessions on mobile), the bounce rate differences clearly suggest that the ad post-click experience between mobile and desktop differs. The device has an impact on how users experience the ad.

Mobile-Optimised Ad Landing Pages.

We showed that the ad experience on mobile is different to that on desktop. This can partly be explained by the different ways users interact with their desktop and their mobile. However, a preliminary analysis done on the landing pages of the ads in our dataset showed that, when served on a mobile device, some of the landing pages were *not* mobile-optimised, which is likely to have a negative effect on users [16]. In optimised version, landing pages have typically larger buttons, no long text paragraphs, and a single large image of the product advertised in the middle of the page. To see the effect of this on the ad post-click experience, and the extent to which this reflects the quality of the ad, we designed a mechanism to automatically detect when a landing page is mobile-optimised or not.

We first downloaded the landing pages, rendered them and extracted seven features. We used features such as the size in bytes of the HTML landing page, and whether the page contains an apple touch icon. Next, we manually labelled a sample of 259 ads as mobile-optimised (Opt) or non-optimised (Npt). Our training set consisted of 108 Opt and 151 Npt ads. We fed the feature representation of the landing pages in our annotated dataset to a Gradient Boosted Decision Tree classifier, and we estimated its quality using leave-one-out cross validation (LOOCV). The classifier reached a F_1 score of 0.995, demonstrating its high accuracy.

We then tested if dwell time and bounce rate of a landing page correlates with it being Opt. Using a sample dataset of 500 ads, with approximately 65.0% of them being mobile-optimised and 35.0% non-mobile-optimised, we conducted a similar experiment to the previous section, but with Opt ads and Npt as our classes. The results are shown in Figure 4. We employ a Two-sample Kolmogorov-Smirnov test to look at whether the differences are significant.

The distribution of the dwell time difference is very similar for both groups, Opt and Npt (Figure 4(a)). The average dwell time decreases by 31.8% ($median = 31.8\%$) for Opt landing pages, and by 28.9% ($median = 33.6\%$) for Npt landing pages. The difference is not significant (p-value $= 0.20$). *Whether a landing page is optimised does not influence how long users spend on the mobile ad site.*

When considering bounce rate, however, we observe differences (Figure 4(b)). The average bounce rate decreases by 6.9% (median decreases by 30.4%) for Opt landing pages but increases by 13.4% (median decreases by 11.5%) for Npt landing pages. These differences are statically significant (p-value $= 0.003$). Therefore, *mobile-optimised landing pages have a positive influence on users, as they are less likely to lead to bounce, compared to when shown on the desktop.* Studies looking at ad post-click experience in the mobile context should account for this property of the landing page, which by itself is not surprising. However, since we observe no difference in dwell time other features of the landing pages influence user post-click experience. This is the motivation for the landing page analysis discussed next.

4. PREDICTING HIGH QUALITY ADS

When a landing page has a high dwell time it is likely to be interesting, and *potentially* more likely to lead to a conversion. On the other hand, if a landing page has a high bounce rate it means that most of the users are annoyed by it. The previous section demonstrated dwell time to be a good proxy of user post-click experience and bounce rate provided additional insights on user post-click experience. In this section, we use these metrics to identify *high quality* ads on the basis of their landing pages, as experienced by mobile users on Yahoo mobile news stream.

4.1 High Quality Ads Ranking

Given a request for an ad to be served on Yahoo mobile news stream, we describe how ads are ranked to fulfill that request. Let $P^i_{\text{click}} \in [0, 1]$ be the predicted probability of an ad i being clicked, and let $\text{bid}^i \in \mathbb{R}$ be the amount of money the advertiser is willing to pay for its ad to be shown. Ranking the ads is done by computing the *expected cost per click* $\text{eCPC}^i = P^i_{\text{click}} \cdot \text{bid}^i$ for each ad, and later sorting them in descending order of this value. Our aim is to predict P^i_{SAT}, that is the *conditional probability* of a user being satisfied *given* that he or she clicked on an ad i. The goal is thus to estimate the overall *joint probability* of clicking on an ad *and* being satisfied by its landing page. More formally, we want to compute $P^i_{\text{HQ}} = P^i_{\text{click}} \cdot P^i_{SAT}$.[9] Finally, the ranking of ads will be computed as $\text{eCPC}^i_{\text{HQ}} = P^i_{\text{HQ}} \cdot \text{bid}^i$. The ad with the highest $\text{eCPC}^i_{\text{HQ}}$ will then be served.

We thus want to predict a class label $Y_i \in \{-1, 1\}$ for a given landing page X_i represented by a feature vector $\phi(X_i)$. The class label Y_i is 1 if X_i is a high quality page, -1 otherwise. Concretely, we aim at estimating the following probability density function:

$$P(Y_i = -1 | \phi(X_i)) = 1 - P(Y_i = 1 | \phi(X_i))$$

which, in practice, corresponds to the joint probability P^i_{HQ}.

[9] Note that P^i_{SAT} is conditioned on P^i_{click}.

4.2 Two Definitions of High Quality Ads

In this work, we focus on the quality of an ad as experienced by users on the ad landing page. Within this context, there are many definitions of a *high quality ad* we could use. We consider two definitions, one based on dwell time and the other on bounce rate, which we have shown to be good proxies of the ad post-click experience.

Formally, with a web page X_i accessed by a set of n_{X_i} users we associate two real numbers $\delta_{X_i} > 0$ its mean dwell time computed over the set of n_{X_i} users, and $\beta_{X_i} \in [0, 1]$ its bounce rate computed as the fraction of the n_{X_i} leaving X_i before a given time threshold.

1. *High Dwell Time.* $Y_i = 1$ when $\delta_{X_i} > t_\delta$ where t_δ is a threshold defined on dwell time. Under this assumption $P^i_{\text{HDT}} = P\left(Y_i = 1 | \phi\left(X_i\right)\right) = P\left(\delta_{X_i} > t_\delta\right)$.
2. *Low Bounce Rate.* $Y_i = 1$ when $\beta_{X_i} < \tau_\beta$ where $0 < \tau_\beta < 1$ is a bounce rate threshold value. Under this assumption $P^i_{\text{LBR}} = P\left(Y_i = 1 | \phi\left(X_i\right)\right) = P\left(\beta_{X_i} < \tau_\beta\right)$.

We experiment with these two definitions, to identify whether they can be predicted based on the ad landing page features. Next we define these features.

4.3 Ad Landing Page Features

Inspired by previous work [2, 3, 6, 11] exploiting features extracted from the landing pages to categorise ads, our understanding of the problem at hand and our own expertise and common sense, we define three sets of features.

CONT (C) Features. This group captures the content including the functionality of the landing page:

- *media*: boolean value stating if the site is responsive.[10]
- *clickToCall*: number of clickables linking to a phone call.
- *imageHeight*: height of the landing page.
- *imageWidth*: width of the landing page.
- *numClickable*: number of clickables.
- *numDropdown*: number of dropdown lists.
- *numImages*: number of images.
- *numInputCheckbox*: number of checkboxes.
- *numInputRadio*: number of radio buttons.
- *numInputString*: number of input strings (usually to elicit users details).
- *tokenCount*: number of tokens (words).
- *viewPort*: a boolean feature represents if the site can be tuned to different screen sizes.
- *windowSize*: total width of all `div` tags on the page, which allows detecting carousels.
- *nounsSumOfScores*: number of nouns.
- *numConceptAnnotation*: number of all Wikipedia *entities* (a concept with a Wikipedia entry).
- *summarizabilityScore*: predicts if the page is a good candidate for extracting a summary, where higher value means the landing page is more “newsy”.
- *isMobileOptimised*: the result of the classifier as discussed in Section 3.

SIM (S) Features. This group of features captures the similarity of the landing page with the creative text displayed within the stream. Usually, a user sees the creative and decides to click on the basis of the text written there. If the semantics of the creative text is very different from the semantics of the landing page then the user who clicked may be annoyed and leave immediately the page.

- *similarityNoun*: cosine similarity between creative text and landing page based on nouns.
- *similarityWikiIds*: cosine similarity between creative text and landing page based on Wikipedia entities.

HIST (H) Features. These features captures historical information about the ad past performance.

- *impressions*: number of times the ad was shown.
- *clicks*: number of times the ad was clicked.
- *bouncerate*: bounce rate of the ad.
- *avgdwelltime*: average dwell time of the ad.
- *avgdwelltimenonshort*: average dwell time when short clicks were removed.
- *ctr*: click-through rate of the ad.
- *cpx*: cost per click of the ad.

4.4 Prediction Quality

Using the features listed above we train several models to predict P^i_{HQ} using three well-known learning methods: Logistic Regression (LogReg) [28], Support Vector Machines (SVM) [7], and Gradient Boosted Decision Trees (GBDT) [10]. We use the implementations of these methods available in the Python scikit-learn package.[11] The probability values are extracted using the implementation available with this framework. We adopt standard parameters for each method. For the LogReg classifier we set C, the inverse of regularization strength, to 100, and $L1$ as penalty norm, and 0.01 as the tolerance value for stopping the optimization. For the SVM classifier we adopt a *RBF* kernel with a penalty parameter C of the error term equal to 1.0, 0.0 as the gamma kernel coefficient, and 10^{-3} as the tolerance used in the stopping criterium. Finally, for the GBDT classifier we generate a forest of 100 trees with a max depth of each tree of 4, and a learning rate of 0.01.

4.5 Offline Model Evaluation

To assess the validity of our prediction models, we run a traditional offline evaluation based on historical data.

Experimental Setup. For each definition of high quality ads, we report the performance of predictors using three standard metrics, Area Under the ROC Curve (AUC), F_1, and the Matthews Correlation Coefficient (MCC) [17]. The latter is a correlation measure between predictions and labels taking into account the popularity of each class.

The dataset used to run the experiments is a uniformly generated sample of our ad set. As a training set we extract a sample of 1,500 ads shown in March 2014 to users of the system. The test set contains a sample of 550 ads shown in April 2014. In all the tests we conduct we experiment with several thresholds for dwell time (t_δ) and bounce rate (τ_β). Finally, we test several combinations of features: content-based or C features; similarity-based or S features; and history-based or H features.

Tables 2 and 3 report the results for predicting the probability of high dwell time and low bounce rate. Only for high dwell time, we report the results for the three classification methods (LogReg, SVM, and GBDT). For bounce rate, we report the results using LogReg.

First, the various classification methods perform similarly (with a slight advantage of SVM over the others for high dwell time). Similarity-based features perform bad as they

[10] http://en.wikipedia.org/wiki/Responsive_web_design

[11] http://scikit-learn.org

never increase (and in some case are detrimental to) both metrics. Finally, history-based features are very important as they boost, for instance, AUC above 0.8 when combined with content. Content-based features alone are already achieving high values with both metrics.

We should note that history-based features are very sparse for ads and, in particular, are non-existing for newly inserted ones. Content-only classifiers can always be used as they provide the perfect solution to the "item cold-start problem" that will be often experienced with many ads.

Features	Method	t_δ	AUC	F$_1$	MCC
C	LogReg	35	0.71	0.65	0.44
C-S	LogReg	35	0.70	0.64	0.42
C-H	LogReg	35	0.82	0.81	0.64
C-S-H	**LogReg**	**35**	**0.84**	**0.83**	**0.67**
C	SVM	35	0.82	0.81	0.65
C-S	SVM	35	0.82	0.81	0.64
C-H	SVM	35	0.83	0.82	0.66
C-S-H	SVM	35	0.83	0.82	0.66
C	GBDT	35	0.77	0.73	0.55
C-S	GBDT	35	0.77	0.74	0.56
C-H	GBDT	35	0.83	0.82	0.66
C-S-H	GBDT	35	0.83	0.82	0.66
C	LogReg	40	0.70	0.60	0.47
C-S	LogReg	40	0.72	0.63	0.49
C-H	LogReg	40	0.83	0.79	0.66
C-S-H	LogReg	40	0.83	0.79	0.66
C	SVM	40	0.83	0.79	0.67
C-S	SVM	40	0.83	0.79	0.67
C-H	**SVM**	**40**	**0.83**	**0.80**	**0.68**
C-S-H	**SVM**	**40**	**0.83**	**0.80**	**0.68**
C	GBDT	40	0.82	0.77	0.70
C-S	GBDT	40	0.81	0.77	0.68
C-H	GBDT	40	0.83	0.80	0.68
C-S-H	GBDT	40	0.83	0.80	0.68
C	LogReg	45	0.69	0.57	0.48
C-S	LogReg	45	0.70	0.57	0.50
C-H	LogReg	45	0.79	0.72	0.60
C-S-H	LogReg	45	0.79	0.72	0.60
C	SVM	45	0.82	0.76	0.67
C-S	SVM	45	0.82	0.76	0.67
C-H	SVM	45	0.80	0.73	0.61
C-S-H	SVM	45	0.80	0.73	0.61
C	GBDT	45	0.72	0.62	0.56
C-S	GBDT	45	0.71	0.60	0.54
C-H	GBDT	45	0.80	0.73	0.61
C-S-H	GBDT	45	0.80	0.73	0.61

Table 2: Dwell Time prediction performance on models built on ads data from March 2014 and tested on April 2014. We vary t_δ to evaluate the impact of the threshold chosen on the prediction ability of the model (best results in bold).

4.6 Feature Ranking

We show the importance of each set of features, namely C, C-S and C-S-H. We limit our analysis to the GBDT classifier and use the technique implemented in the scikit-learn toolkit. This technique can naturally be used to induce a ranking of the "importance" of features in a regression or classification problem. Table 4 shows the top-15 features ranked according to their importance scores as output by GBDT when trained for predicting ad quality on the basis of dwell time being above a fixed threshold[12]. Each column refers to the ranking of a specific set of features: C, C-S and C-S-H. The i-th row contains the i-th ranked feature, along with its category (C, S or H) and importance score, for each set.

Features	Method	τ_β	AUC	F$_1$	MCC
C	LogReg	0.2	0.51	0.78	0.06
C-S	LogReg	0.2	0.59	0.8	0.24
C-H	LogReg	0.2	0.79	0.85	0.58
C-S-H	LogReg	0.2	0.78	0.85	0.57
C	LogReg	0.22	0.61	0.74	0.27
C-S	LogReg	0.22	0.61	0.72	0.23
C-H	LogReg	0.22	0.86	0.86	0.71
C-S-H	**LogReg**	**0.22**	**0.85**	**0.86**	**0.70**
C	LogReg	0.25	0.57	0.63	0.15
C-S	LogReg	0.25	0.63	0.63	0.26
C-H	LogReg	0.25	0.83	0.82	0.67
C-S-H	LogReg	0.25	0.83	0.81	0.66

Table 3: Bounce rate prediction performance on models built on ads data from March 2014 and tested on April 2014. We vary τ_β to evaluate the impact of the threshold chosen on the prediction ability of the model (best result in bold).

When the classifier is trained using only content features (C) or both content and similarity features (C-S), *clickToCall* is the most important signal. The first similarity feature (*similarityNoun*) is ranked 11-th when using the C-S set. This means that similarity features do not provide significant insights to discriminate between high and low quality ads. This is even more evident when using the C-S-H set, where no similarity features appear in the top-15 list. We also observe that features related to the functionality of the landing page (*numDropdown*, *numInputRadio*) impact more than those related to the content itself (*tokenCount*, *nounsSumOfScores*) and aesthetic (*isMobileOptimised*, *media*). Only when considering the similarity features (C-S) as well, two content–related features become more important (*summarisabilityScore* and *numConceptAnnotation*). Finally, for the set C-S-H, the top-3 most important signals come all from historical features; *avgdwelltime* and *avgdwelltimenonshort* play a crucial role. This is not surprising as our classifier uses dwell time as the proxy for ad quality.

Using LogReg, we saw that the functionality features have negative coefficients, suggesting that they affect negatively the post-click experience. These features relate to the existence of a form on the landing page, and as such are a strong deterrent to users. The forms when displayed on a mobile device may not be user-friendly, or users are simply not willing to share private information.

In Section 3, we showed that dwell time was a good proxy of an ad post-click experience. The LogReg solution showed very good performance in predicting high dwell time, i.e. dwell time being above a given threshold. Although SVM performed slightly better, LogReg supports quick update operations, which is important when deployed in production. We therefore decide to deploy such a model using only content-based features to allow full coverage of the ads in the database, and not just those for which we have historical data. We use 40 seconds as our threshold, as in our dataset it corresponds to the *median* of the overall dwell times distribution. We also choose to deploy the version predicting high dwell time as we wanted to serve ads on which users spend time. This means that not only they are of high quality but also have the potential of leading to a "conversion".

[12] As explained later in this section, we select 40 seconds as our dwell time threshold.

Rank	C	C-S	C-S-H
1.	[C] *clickToCall* (0.331)	[C] *clickToCall* (0.313)	[H] *avgdwelltime* (0.480)
2.	[C] *windowSize* (0.125)	[C] *summarizabilityScore* (0.109)	[H] *avgdwelltimenonshort* (0.314)
3.	[C] *numClickable* (0.113)	[C] *numConceptAnnotation* (0.086)	[H] *bouncerate* (0.059)
4.	[C] *numInputRadio* (0.089)	[C] *numInputRadio* (0.086)	[C] *clickToCall* (0.047)
5.	[C] *numDropdown* (0.084)	[C] *numDropdown* (0.077)	[C] *numClickable* (0.019)
6.	[C] *numImages* (0.064)	[C] *numClickable* (0.075)	[C] *numImages* (0.012)
7.	[C] *numInputCheckbox* (0.031)	[C] *numImages* (0.048)	[C] *summarizabilityScore* (0.011)
8.	[C] *imageHeight* (0.029)	[C] *windowSize* (0.036)	[C] *numInputRadio* (0.008)
9.	[C] *nounsSumOfScores* (0.026)	[C] *imageHeight* (0.029)	[C] *numDropdown* (0.008)
10.	[C] *numConceptAnnotation* (0.024)	[C] *nounsSumOfScores* (0.028)	[H] *ctr* (0.007)
11.	[C] *viewPort* (0.021)	[S] *similarityNoun* (0.023)	[C] *numConceptAnnotation* (0.006)
12.	[C] *media* (0.021)	[C] *numInputCheckbox* (0.023)	[H] *clicks* (0.006)
13.	[C] *tokenCount* (0.014)	[C] *viewPort* (0.016)	[H] *cpx* (0.004)
14.	[C] *numInputString* (0.011)	[C] *spaceball_score* (0.014)	[H] *impressions* (0.003)
15.	[C] *imageWidth* (0.010)	[C] *media* (0.012)	[C] *numInputCheckbox* (0.003)

Table 4: Top-15 ranked features using GBDT with three sets of features: C, C-S, and C-S-H.

5. ONLINE BUCKETING EVALUATION

To measure the impact that our ad quality prediction model has on users we conduct an *online* evaluation through A/B testing. We implement an ad ranking scorer on Yahoo *Gemini* based on the Logistic Regression (LogReg) prediction model, and assess its performance on the mobile news stream app running on iOS.

We split the incoming traffic into two *buckets*, i.e. *baseline* and *ad quality*. In the first bucket, ads are served using the existing ranking scheme, i.e. the expected cost per click, whereas in the second bucket ads are served according to the newly proposed ranking scorer that accounts for the ad post-click experience (i.e. the probability that users do not return to the stream within the next 40 seconds).

We measure the user post-click experience with dwell time and bounce rate. Specifically, we compute the *median* of the former to deal with the high variance of dwell times. For the bounce rate we report, instead, the *average* of the values. Bounce rate is already a normalised score in the $[0, 1]$ range and also exhibits small variance.

Experimental Setup. We consider two distinct datasets of ad clicks, randomly sampled from May to June 2014. The first dataset is drawn from the baseline bucket and contains only clicks on ads ranked by the baseline scorer. The second dataset is drawn from the ad quality bucket and contains all the clicks of ads served by the ad quality ranking scorer.

We conduct three analyses, at the *(ad-)click-level*, at the *ad-level*, and at the *user-level*. For all three, we evaluate the two buckets performance. First, we compare the performance accounting for all ads (users) as they appear in the two datasets; we refer to this experimental setting as All. Then, we measure the performance limited to only those ads (users) common to the two datasets; we call this Shared. Finally, we focus only on those ads (users) that appear in only one of the two datasets; we refer to this as Unique.

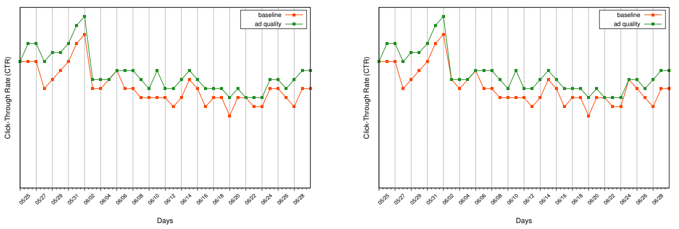
Click-level Analysis. We discuss how the daily click-through rate behaves on the two buckets. When assessing the effect of any change, e.g., in a ranking algorithm, it is important to do so over a long period of time, as an increased performance shortly after the change may not translate in the long run to better user experience [13].

We collect around 14,500 ads from the dataset relating to the baseline bucket, and about 12,500 ads from the ad quality bucket (All). More than 11,000 ads are shared between the two buckets (Shared), and therefore are very likely high quality ads. Looking at these two settings (All and Shared) we observe that the overall click-through rate (the whole month) increases by about 18% and 13% on the ad quality bucket. Figure 5 shows that for both settings, the click-through rate is always higher for the ad quality bucket. This means that the probability of a user clicking on an ad increases in the ad quality bucket. Interestingly, the two time series are perfectly correlated (Pearson's $r = 1.0$). The *paired t-test* applied to each pair of time series samples shows that the differences between the two samples, though correlated, are statistically significant (p-value $= 0.01$).

If we consider only the (high quality) ads shared between the two buckets the chance of clicking on an ad seems to depend on which bucket the ad is served. Intuitively, if a high quality ad is served together with other, lower quality ads (i.e. baseline bucket) then users may not perceive it as valuable as it is, and thus the probability of clicking on it decreases. However, if the *same* high quality ad is served with other high quality ads (i.e. ad quality bucket) the users may be more likely to click on it because they have been exposed to ads leading to positive experience. Therefore we conjecture that the perception of the ad quality is influenced by the ads to which a user has been "exposed" and by how those have been experienced (either positive or not), which is what Figure 5(b) suggests. This further confirms our results relating ad post-click experience and future ad clicks discussed in Section 3.

Ad-level Analysis. The second analysis shows how dwell time and bounce rate behave on the two buckets, from an "ad perspective". We first remove all the ad clicks with dwell time greater than 10 minutes (following from Section 3). Afterwards, we consider only those ads that received at least 10 clicks. This is done to avoid the effect of outliers and ensure that we have enough clicks to calculate bounce rates. Finally, a click is considered a bounce if its dwell time is less than or equal to 5 seconds. This resulted in around 1,000 ads in the baseline bucket and 700 in the ad quality bucket (All), with around 600 ads common to both buckets (Shared). We should note the ads used in this analysis form a subset of those used in the first click-level analysis.

The results for the All, Shared, and Unique datasets are reported in Table 5. Unique is the dataset comprised of ads that are only in one of the two buckets. Each cell of the table refers to the relative difference (in percentage) between



(a) All

(b) Shared

Figure 5: Daily Click-Through Rate (CTR): *baseline* vs. *ad quality*.

	All	Shared	Unique
Dwell Time (*median*)	+30.0	+20.0	+35.7
Bounce Rate (*avg*)	-6.7	0.0	-25.0

Table 5: Differences (%) in Dwell Time and Bounce Rate between *ad quality* and *baseline* ads.

	All	Shared	Unique
Dwell Time (*median*)	+25.0	+20.0	+30.0
Bounce Rate (*avg*)	-12.0	-12.5	-12.0

Table 6: Differences (%) in Dwell Time and Bounce Rate between *ad quality* and *baseline* users.

the statistics as computed from the ad quality and baseline buckets, respectively. Note also that while a positive difference is desirable in the case of dwell time (i.e. showing the ad quality bucket exhibits more time spent on the ad landing page), for the bounce rate we aim at reducing the number of short clicks, and so we prefer a negative difference.

In all the experimental settings, the ad quality bucket outperforms the baseline bucket. This is particularly visible when considering Unique ads. Interestingly, when looking at the Shared ads, the median dwell time is still higher when the ads are served as part of the ad quality bucket, compared to the baseline bucket. In Figure 5(b), we can see that the click-through rate for the Shared ads is higher in the ad quality bucket. Therefore, not only serving high quality ads together (i.e. the ad quality bucket) attract more clicks, the post-click experience, as measured with dwell time, is also positively affected: *users engage more with the ads.*

Similarly, the average bounce rate is lower with the ad quality bucket, which implies a lower probability of bouncing back once users click on ads that have been deemed to be of high quality. Interestingly, for Shared ads, there is no difference in bounce rate. Bounce rate is more a reflection of an ad being annoying. This suggests that the property of "being annoying" is a characteristic of the ad itself (i.e. its landing page as experienced on average by users), and does not depend on which other ads are served during the session.

We compare the distribution of dwell time and bounce rate, as observed in the two buckets. We run a Kolmogorov-Smirnov test for two samples on each pair of observations to test if both samples might have been drawn from the same underlying probability distribution. For all cases except one, this is not the case with p-value lower than 0.01. The exception is with the bounce rate on the Shared setting. This however complies with the results in Table 5, where no difference exists between the two buckets for the average bounce rate for Shared, further confirming that "being an annoying" ad or being a *bad ad* comes from the ad "itself".

User-level Analysis. The aim of the last analysis is close to that of the previous one yet from a "user perspective". We remove all the ad clicks having dwell time larger than 10 minutes and a click is considered a bounce as long as its dwell time is at most 5 seconds. Furthermore, we take into account only those users who clicked on at least 2 unique ads. This results into around 16,000 users in the baseline bucket and 11,000 in the ad quality bucket, with about 2,700 individuals shared between the two buckets. Table 6 shows the relative differences (in percentage) of dwell time and bounce rate as computed from the two buckets.

The median dwell time is higher for the ad quality bucket, which means that when users are served ads deemed of high quality, they spend time on the ad landing page before returning to the stream. In particular, looking at the Shared setting (i.e. users appearing in both buckets), serving high quality ads indeed promotes a positive post-click experience. This is further accentuated when users experience only high quality ads (as seen with Unique).

Concerning the average bounce rate, this is computed as the average fraction of bounce clicks for each user. This is different from the actual definition of bounce rate, which instead is formulated at the ad-level. Still, we observe a decrease of the average bounce rate in the ad quality bucket. This relates to the fact that more high quality ads are served in the ad quality bucket, which leads to fewer users bouncing back after clicking on them. Finally, the difference between dwell time and bounce rate distributions is statistically significant (p-value $\ll 0.01$) using Kolmogorov-Smirnov test.

This section shows that returning high quality ads, as measured in terms of the ad post-click experience, is important. Not only this increases CTR, and as a likely consequence revenue in the long-term, it has a positive effect on users, as seen by the increase in dwell time and decrease in bounce rate. Interestingly, when users are served ads deemed of high quality together, their engagement with the ads, in terms of time spent on the ad site, is positively affected. In addition, so called *bad ads* are so because of themselves, and this independently of whether they are served or not with ads deemed of high quality.

6. CONCLUSIONS

We proposed a method to identify high quality ads served on one of Yahoo mobile stream platforms. In this paper, quality refers to the ad post-click experience. We therefore related the ad landing page to the ad post-click experience, which we measured through well known engagement metrics, dwell time and bounce rate. We first showed that these measures were appropriate proxies of ad quality in our context. We also showed that users clicking on ads that promote a positive post-click experience are more likely to click on ads in the future, and their long-term engagement is positively affected. We also compared these measures in the context of mobile versus desktop devices, and found that a positive ad post-click experience is not just about serving ads with mobile-optimised landing pages; other aspects of an landing page affect the post-click experience.

We then put forward an approach that analyses ad landing pages, and shows how these can affect dwell time and bounce rate. We experimented with three types of features, related to the actual content and organization of the ad landing page, the similarity between the creative and the landing page, and ad past performance. We saw that the later type of features were best at predicting dwell time and bounce rate; we also show that content and organization features perform well, and have the advantages to be applicable for all ads, not only for those that have been served. Overall, our offline evaluation showed that, using dwell time as a proxy of post-click experience, we could predict the quality of an ad with good accuracy.

Finally, we deployed our prediction model for ad quality based on dwell time on Yahoo *Gemini*, and validated its performance on the mobile news stream app running on iOS. This was conducted through A/B testing to measure the impact of our approach in a real-world scenario. Finally, dwell time and bounce rate exhibited statistically significant differences, when the ads served took into account their predicted dwell time, using the landing pages features, compared to when they did not. Dwell time increased by 30% whereas bounce rate decreased by 6.7%.

As future work, we plan to train our model on new features, such as readability, PageRank, sentimentality level of the landing pages. We are currently carrying out user studies to understand how users perceive the quality of the landing pages, and how this can be translated into additional features. Finally, we will be studying alternative metrics, for example combining dwell time and bounce rate, but also others, to fully capture the ad post-click experience.

7. ACKNOWLEDGEMENTS

We would like to thank John Agapiou and Andy Haines for their help in obtaining and polishing the data we used in our experiments and the feedbacks on our results.

8. REFERENCES

[1] J. Azimi, R. Zhang, Y. Zhou, V. Navalpakkam, J. Mao, and X. Fern. Visual appearance of display ads and its effect on click through rate. In *CIKM*, 2012.

[2] H. Becker, A. Broder, E. Gabrilovich, V. Josifovski, and B. Pang. Context transfer in search advertising. In *SIGIR*, 2009.

[3] H. Becker, A. Broder, E. Gabrilovich, V. Josifovski, and B. Pang. What happens after an ad click?: Quantifying the impact of landing pages in web advertising. In *CIKM*, 2009.

[4] K. E. Boudreau. Mobile advertising and its acceptance by american consumers. Bachelor thesis, 2013.

[5] A. Z. Broder, M. Fontoura, V. Josifovski, and L. Riedel. A semantic approach to contextual advertising. In *SIGIR*, 2007.

[6] Y. Choi, M. Fontoura, E. Gabrilovich, V. Josifovski, M. Mediano, and B. Pang. Using landing pages for sponsored search ad selection. In *WWW*, 2010.

[7] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 1995.

[8] M. de Sa, V. Navalpakkam, and E. F. Churchill. Mobile advertising: evaluating the effects of animation, user and content relevance. In *CHI*, 2013.

[9] T. Foran. Native advertising strategies for mobile devices. `http://www.forbes.com/sites/ciocentral/2013/03/14/native-advertising-strategies-for-mobile-devices/`, 2013.

[10] J. H. Friedman. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 2002.

[11] A. Kae, K. Kan, V. K. Narayanan, and D. Yankov. Categorization of display ads using image and landing page features. In *LDMTA*, 2011.

[12] Y. Kim, A. Hassan, R. W. White, and I. Zitouni. Modeling dwell time to predict click-level satisfaction. In *WSDM*, 2014.

[13] R. Kohavi, A. Deng, B. Frasca, R. Longbotham, T. Walker, and Y. Xu. Trustworthy online controlled experiments: Five puzzling outcomes explained. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 786–794. ACM, 2012.

[14] J.-H. Lee, J. Ha, J.-Y. Jung, and S. Lee. Semantic contextual advertising based on the open directory project. *ACM TWEB*, 2013.

[15] M. Levene. *An Introduction to Search Engines and Web Navigation.* Addison Wesley Publishing Company, 2005.

[16] H. Liu, W.-C. Kim, and D. Lee. Characterizing landing pages in sponsored search. In *LA-WEB*, 2012.

[17] B. W. Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysical Acta*, 1975.

[18] V. Murdock, M. Ciaramita, and V. Plachouras. A noisy-channel approach to contextual advertising. In *ADKDD*, 2007.

[19] R. J. Oentaryo, E.-P. Lim, J.-W. Low, D. Lo, and M. Finegold. Predicting response in mobile advertising with hierarchical importance-aware factorization machine. In *WSDM*, 2014.

[20] A. Penev and R. K. Wong. Framework for timely and accurate ads on mobile devices. In *CIKM*, 2009.

[21] L. Ritzel, C. V. der Schaar, and S. Goodman. *Native Advertising Mobil.* GRIN Verlag GmbH, 2013.

[22] R. Rosales, H. Cheng, and E. Manavoglu. Post-click conversion modeling and analysis for non-guaranteed delivery display advertising. In *WSDM*, 2012.

[23] D. Sculley, R. G. Malkin, S. Basu, and R. J. Bayardo. Predicting bounce rates in sponsored search advertisements. In *KDD*, 2009.

[24] E. Sodomka, S. Lahaie, and D. Hillard. A predictive model for advertiser value-per-click in sponsored search. In *WWW*, 2013.

[25] Y. Song, H. Ma, H. Wang, and K. Wang. Exploring and exploiting user search behavior on mobile and tablet devices to improve search relevance. In *WWW*, 2013.

[26] X. Yi, L. Hong, E. Zhong, N. Liu, and S. Rajan. Beyond clicks: Dwell time for personalization. In *RecSys*, 2014.

[27] P. Yin, P. Luo, W.-C. Lee, and M. Wang. Silence is also evidence: interpreting dwell time for recommendation from psychological perspective. In *KDD*, 2013.

[28] H.-F. Yu, F.-L. Huang, and C.-J. Lin. Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning*, 2011.