Predicting Pre-click Quality for Native Advertisements

Ke Zhou, Miriam Redi, Andy Haines, Mounia Lalmas
Yahoo Labs, London
{kezhou,redi,haines,mounia}@yahoo-inc.com

ABSTRACT

Native advertising is a specific form of online advertising where ads replicate the look-and-feel of their serving platform. In such context, providing a good user experience with the served ads is crucial to ensure long-term user engagement. In this work, we explore the notion of ad quality, namely the effectiveness of advertising from a user experience perspective. We design a learning framework to predict the pre-click quality of native ads. More specifically, we look at detecting *offensive* native ads, showing that, to quantify ad quality, ad offensive user feedback rates are more reliable than the commonly used click-through rate metrics. We then conduct a crowd-sourcing study to identify which criteria drive user preferences in native advertising. We translate these criteria into a set of ad quality features that we extract from the ad text, image and advertiser, and then use them to train a model able to identify offensive ads. We show that our model is very effective in detecting offensive ads, and provide in-depth insights on how different features affect ad quality. Finally, we deploy a preliminary version of such model and show its effectiveness in the reduction of the offensive ad feedback rate.

1. INTRODUCTION

In online services, native advertising has become a very popular form of online advertising [18], where the ads served reproduce the look-and-feel of the platform in which they appear. Online native ads[1] are served as suggested posts on Facebook, promoted tweets on Twitter, or sponsored contents on Yahoo news stream (see example in Figure 1). Native ads tend to be more effective than traditional display ads in terms of user attention and purchase intent [16], and cause much less prominent ad blindness effect [2].

To improve the effectiveness of native advertising, ad serving systems should provide ads that satisfy users's need according to two aspects, *relevance* and *quality*. *Relevance* is the extent to which an ad matches a user interest: ads are indeed often personalized according to the target user preferences, browsing patterns, search behavior, etc. *Quality* is a characteristic of the ad itself, independent of the users targeted by the platform. The *quality* of an ad reflects the nature of the advertised product and the design decision of the advertiser, and affects the experience of any user exposed to the ad. The quality of an ad depends on, for example, the visual composition of the ad creative, the clarity and trustworthiness of the text in the ad copy or the landing page, or the adultness of the ad content.

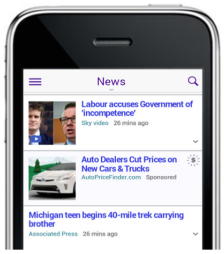


Figure 1: Example of a native ad (the second item with the "dollar" sign) in a news stream on a mobile device.

Promoting relevant *and* quality ads to users is crucial to maximize long-term user engagement with the platform [8]. In particular, low-quality advertising (the promotion of low quality ads) has been shown to have detrimental effect on long-term user engagement [11, 39]. In display advertising, several studies [10, 11] suggest that excessive animation or high level of intrusiveness can have an undesirable impact on the ad effectiveness. In addition, disturbing ads cause various issues beyond mere annoyance, as users might get distracted, or unable to consume the actual content of the page where the ad is displayed [11].

Low quality advertising can have even more severe consequences in the context of native advertising, since native advertisement forms an integrated part of the user experience of the product. For example, as shown in [18], a bad *post-click quality* (quantified by short dwell time on the *ad landing page*) in native ads can result in weaker long-term engagement (e.g. fewer clicks).

Given these observations, in this work we especially aim at countering low quality native advertising. We are only interested in the perceived *quality* of the ads served, independent of their relevance to the user, or the targeting algorithm used for ad serving. As a first step towards the full understanding of native ad quality, we focus on the *pre-click* user experi-

[1]We use *advertisement* and *ad* interchangeably.

WWW 2016, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4143-1/16/04.
http://dx.doi.org/10.1145/2872427.2883053.

ence of the native ad, i.e. the user experience induced by the ad creative[2] *before* the user decides (or not) to click.

Due to the low variability in terms of ad formats in native advertising, the content and the presentation of the ad *creative* are extremely important to determine the quality of the ad. To tackle the problem of predicting low quality *pre-click user experience* of native ads, we therefore design a learning approach that analyzes various attributes of native ad creative. Such framework is based on two main elements: a learning target and a set of features extracted from the ad creatives.

Learning Target. How to define the learning target for pre-click ad quality? One may think of *click-through rate (CTR)* as a natural metric (learning target) to predict pre-click quality. However, CTR only reflects short-term user engagement. Although CTR is somehow related to the ad quality, high CTR may not imply good ad quality, as shown in Section 2. Serving ads predicted to have high CTR focuses on short-term revenue, and does not guarantee long-term user engagement.

To quantify a bad pre-click user experience, we therefore use an alternative quality metric, namely ad *offensiveness*. To collect *offensiveness* annotations, we exploit the Yahoo ad feedback mechanism. Such mechanism allows users to choose to hide the ads they are exposed to, and further select one option motivating the reason of their choice, one of them being "It is offensive to me". We collect these judgements and use ad *offensive feedback rate (OFR)* as our ground truth metric of pre-click quality.

Pre-Click Ad Creative Features. We also design a set of ad features that allow our model to predict the offensiveness of the ads. We first conduct a crowd-sourcing study to understand what makes ads more preferred by users. We find that, for example, the aesthetic appeal of the ad creative image, the brand of the advertiser and the trustworthiness of the ad creative are important factors of user preferences. Based on these results, we engineer a set of features specifically reflecting those factors. We derive features from the ad copy (text and image), and the advertiser properties. Such features include text readability, brand quality and image composition. We also include a set of features that characterize user behavior after the ads are served (e.g. dwell time on the landing page).

To summarize, our contributions lie in:

- We use offensive ad feedback as our proxy of pre-click ad quality and analyze its relationship with CTR.
- We conduct a crowd-sourcing study with hundreds of users to understand the underlying reasons of pre-click ad quality preferences.
- We design and analyze a large set of features that characterize various aspects of pre-click native ad quality.
- We learn an effective prediction pre-click quality model, reaching an AUC of 0.77 with cold-start features, as well as providing a thorough understanding of the predictive power of the features.
- We deploy a model based on a subset of the features on Yahoo news streams, which yields a reduction of ad offensive feedback rate of 17.6% on mobile and 8.7% on desktop.

[2]The ad creative is the ad impression shown within the stream, and includes text, visuals, and layout.

2. MEASURING BAD ADS

The detection of low quality ads is a challenging task. The first step is to define how to measure quality on a large scale. One may argue that low click-through rate could be a good indicator of poor ad quality. However, click-through rate is a compounding factor that may be affected by several dimensions, including ad relevance (whether the ads match users' interests), the nature of the advertiser (e.g. its popularity or seasonality) and, certainly, the ad quality (e.g. visuals, trustworthiness). In addition, high CTR may not necessarily mean high quality. In our data, many ads labeled as offensive could be seen as "provocative", attracting clicks.

Offensive Ad Feedback. To monitor ad quality, Internet companies have put in place *ad feedback mechanisms*, which give the users the possibility to provide negative feedback on the ads served, allowing them to hide an ad, and to provide a reason for doing so. Ad feedback tools have been launched by Facebook,[3] Yahoo,[4] and Twitter.[5]

In this work, we exploit the information provided by the Yahoo ad feedback tool, collecting a large-scale dataset of users negative responses to ads. An example of such feedback tool is shown in Figure 2(a). The user can choose to hide the ads they are exposed to, and further select one of the following options as the reason for doing so: (a) *It is offensive to me*; (b) *I keep seeing this*; (c) *It is not relevant to me*; (d) *Something else*. Among the many reasons why users may prefer to hide ads, marking one ad as *offensive* seems the most explicit indication of the *quality* of the ad. Therefore, we propose to use the *offensive* feedback to label the ad quality, with the assumption that the worse the ad, the higher the number of offensive feedback given to it.[6]

Offensive Feedback Metric and Data Collection. We collect ad feedback data over a two-month period, and only select the subset of ads that received a number of feedbacks greater than a given threshold (we selected 5 in our case) to eliminate random or unintentional feedback. We found a long tail of ads that receive no offensive ad feedback or even no negative feedback at all.

From the collected data, we calculate for each ad its *Offensive Feedback Rate (OFR)*:

$$OFR = \frac{freq_{off}}{freq_{impr}}$$

where $freq_{off}$ represents the number of offensive ad feedback registered by the tool, and $freq_{impr}$ denotes the number of ad impressions (the number of times users saw the ad) within the time period. Therefore, OFR quantifies the percentage of the ad impressions that offended the users.

In Figure 2(b), we plot the log-based OFR distribution of all ads with at least 5 offensive ad feedback within the period. The shape of log(OFR) looks similar to a normal distribution with most of the ads having quite small OFR and a few relatively high OFR.

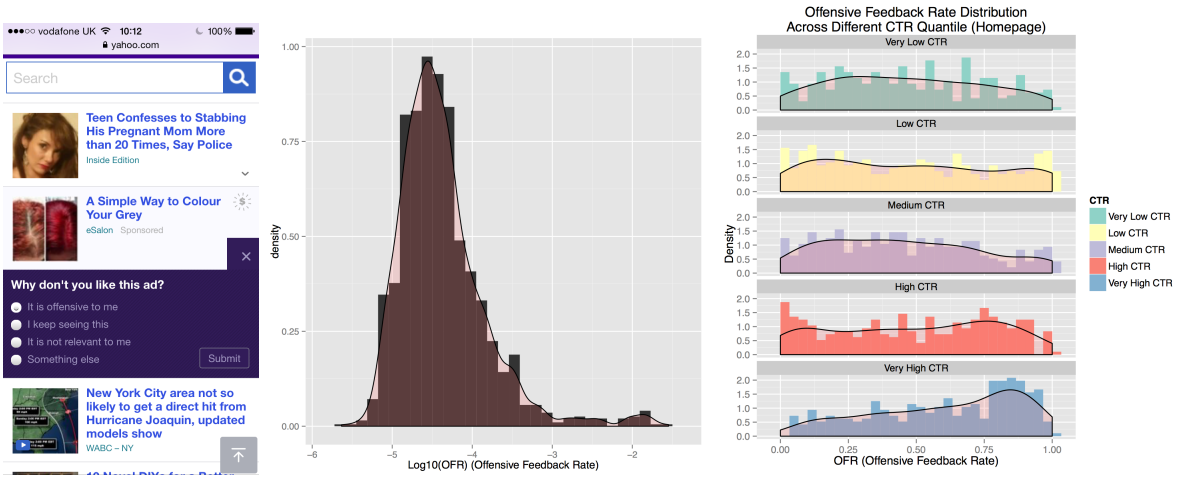
Offensive Feedback Rate vs CTR. We analyze the relationship between OFR and CTR, the commonly used metric for pre-click ad effectiveness. We compute the Spearman

[3]https://www.facebook.com/help/1440106149571479

[4]https://help.yahoo.com/kb/SLN25244.html

[5]https://business.twitter.com/help/what-are-promoted-tweets

[6]While studying the impact of other criteria (e.g. relevance) on ad effectiveness is out of the scope of this paper, the methodology described in this paper can be used to evaluate other dimensions than offensiveness.



(a) Feedback options. (b) Offensive Feedback Rate distribution. (c) Offensive Feedback Rate distribution across five equally sized bins of CTR.

Figure 2: Offensive ad feedback to measure native ad pre-click quality.

and Pearson correlations between CTR and OFR, and found that they are almost uncorrelated, with Spearman correlation standing at 0.155 and Pearson at -0.043: ads with high OFR do not necessarily have a low CTR, and vice versa.

Next, to gain more insights about this relation, we perform a quantile analysis. We first normalize OFR values using the Min-Max method. We split the ads into five equally sized bins (quantiles) according to their CTR:"Very low", "Low", "Medium", "High", and "Very high". We then plot the distribution of normalized OFR of those ads falling in each quantile in Figure 2(c). We find ads with high OFR in all different CTR bins (quantiles), which reaffirms that offensive ads do not necessarily have low CTR. When comparing OFR distributions, we see higher OFR for high CTR quantiles, which means that many offensive ads are click-prone. We manually inspected few examples of high OFR/high CTR ads, and noticed that in many cases, these were click-bait ads, namely ads describing controversial topics, which tend to attract clicks.

As this analysis is performed after removing the long tail of ads with insufficient ad feedback or impressions, the shape of the resulting OFR distributions might be slightly biased. However, ads with high CTR are the ones we care most. Given their high popularity, these ads can be further promoted and therefore shown more frequently, hence possibly offend more users.

3. USER PREFERENCES

So far, we defined from a *metric* perspective how to quantify bad quality ads, using ad *offensive feedback rate*. In this section, we want to understand ad quality from a *user* perspective, thus inferring the underlying criteria that users assess when choosing between ads. To this end, we design a crowd-sourcing study to spot what drives users' quality preferences in the native advertising domain.

Although the ultimate aim of this paper is to detect low quality ads, we do not restrict our user study to bad (offensive) ads only. We want to understand quality perception from a general perspective, and obtain insights about the factors triggering users preferences for *any* ad. This would allow us to reliably discover which elements grasp users' attention when evaluating ad quality, and therefore engineer quality-specific features that model such criteria using a computational approach. For this purpose, we must collect a set of ads representing the general ad population in the marketplace. Hence, we carefully design the data collection for this task to account for ad diversity, selecting ads sampled from the whole ad quality spectrum.

3.1 Methodology

Data Collection. We extract a sample of ads impressed on Yahoo mobile news stream.[7] We consider a period of three months. To ensure diversity and the representativeness of our data in terms of subjects and quality ranges, we uniformly sample a subset of those ads from different CTR quantiles. In addition, we choose to focus on ads from five different popular topical categories: "travel", "automotive", "personal finance", "education" and "beauty and personal care". In total, we sample a set of 80 ads to be shown for assessments.

Task Design. We show users pairs of native ads, and ask them to indicate which ad they prefer in the pair, and the underlying reasons for their choice. By doing so, we elicit pair-wise preference assessments, a widely used methodology to evaluate user experience [30]. Adopting the pair-wise preference methodology is particularly useful in our context, given that people's reactions to ads are generally hostile. Without pre-examining a large number of ads with diverse quality, users might encounter difficulties in objectively giving absolute assessments regarding the quality of an ad, since the sequence of absolute assessments can affect significantly the final user judgements [31]. On the other hand, pair-wise preference assessment enables users to make relative comparison between any two ads, thus easing the task of ad quality assessment and perceptual criteria explanation.

Moreover, individual personal interests can lead to different quality judgements for the same two ads. A user may prefer automotive ads over beauty ads simply because he or

[7]We chose the mobile context for no particular reason, apart for our general interest in mobile native advertising.

Table 1: Underlying reasons of users' preference of ad pairs based on pre-click quality.

Verticals	Brand	Product/Service	Trustworthiness	Clarity	Layout	Aesthetic Appeal
All	0.359	0.429	0.393	0.259	0.153	0.724
Automotive	0.383	0.200	0.333	0.192	0.025	0.800
Beauty and Personal Care	0.036	0.600	0.055	0.182	0.291	0.836
Education	0.179	0.571	0.179	0.250	0.250	0.857
Personal Finance	0.015	0.333	0.333	0.472	0.389	0.667
Travel	0.633	0.575	0.675	0.300	0.125	0.583

she is more interested in cars. To eliminate the effect of ad relevance, we present the users with topically-coherent ads (i.e. ads from the same subject category, such as "beauty"), assuming that, for example, when users are comparing two beauty ads, the preference only depends on the ad quality.

Once chosen their preferred ad, we ask users to express the reasons *why* they chose the selected ad. Users are asked: "Why do you prefer that ad?", and then allowed to evaluate a set of pre-defined possible underlying reasons.

To define such options, we resort to existing user experience/perception research literature. Among others, we were inspired by the UES (User Engagement Scale) framework [24], an evaluation scale for user experience capturing a wide range of hedonic and cognitive aspects of perception, such as aesthetic appeal, novelty, involvement, focused attention, perceived usability, and endurability. UES is partly applicable to our work, since we want to understand the subjective reasons that drive user preferences towards ads. Moreover, previous studies in the context of native advertising [5] investigated user perceptions of native ads with dimensions such as "annoying", "design", "trust" and "familiar brand". Similarly, researchers have studied the amount of ad "annoyingness" in the context of display advertising [11], showing that users tend to relate ad annoyance with factors such as advertiser reputation, ad aesthetic composition and ad logic.

Based on these studies, we provide users with the following options as underlying reasons of their choice: the *brand* displayed, the *product/service* offered, the *trustworthiness*, the *clarity* of the description, the *layout* and the *aesthetic appeal*), all to be rated on a five-grade scale: 1 (strongly disagree), to 5 (strongly agree) or NA (not available).

Experimental Setup. We use Amazon Mechanical Turk to conduct our study. Assessors can choose to perform as many tasks (1 task=1 ad pair) as they wish, and each task is paid $0.05. To avoid learning effect, we ensure that each assessor is not shown the same task more than once. To avoid position bias, we randomly position the two ads (left or right) for each pair-wise assessment.

To guarantee the quality of the collected annotations, we employ various quality control mechanisms. To ensure cultural consistency, we restrict the assessors' provenience to be US only. Moreover, we select assessors with average task acceptance rate over 90%, with a history of at least 1000 assessments. For each task, we display the ad pair, then, after 5 seconds, we show the preference options. This time gap enforces users to first read the two ads carefully (at least for five seconds). For each ad pair, we collect three judgements, from three independent assessors.

We employ two additional mechanisms to further ensure annotation quality: gold standard check (asking users to assess trap ad pairs for which we know which ad should be preferred) and redundancy check (checking that the assessors make similar assessments on an ad pair they previously assessed). To create the gold standard check, we generate trap ad creatives, made of fake text crawled from academic paper content and randomly selected pictures, and rendered with the same format as normal ads. We employ such mechanisms for 1 out of 5 tasks. Assessors are deemed as *untrusted* if failing a certain number of quality checks and their assessments are discarded.

3.2 Analysis

We collect 2250 judgements (including traps) from 154 non-malicious assessors for the 600 ad pairs. 45 assessors (29%) performed 80% of the assessments. Our quality control mechanisms filtered out 136 assessors. To analyze the importance of different factors, we report the percentage of judgements that, for each factor, is assigned to grades 4 or 5 (the assessor highly agrees this factor affects his or her ad preference choice).[8]

Table 1 summarizes the results per ad category. The most important factors are, in order of importance: Aesthetic appeal > Product, Brand, Trustworthiness > Clarity > Layout, where ">" represents a significant improvement on a paired t-test with p-value < 0.05 between factors.[9] To investigate the extent to which the factor importance is consistent across five categories, we perform the ANOVA test since it tests the significance of the differences between the means of different groups of judgements. For all factors impacting user preferences, apart from the *brand* factor, we did not find any significant differences (p-value > 0.05). This suggests that our findings generalize across ad categories.

For different ad categories, compared to the general pattern, however, we can still observe few small differences. Aesthetic appeal is more important for Automotive, Beauty and Education, than Personal Finance and Travel. As a matter of fact, for the Travel category, where most ad images are beautiful, aesthetics does not affect much compared to others. For Beauty and Education categories, the product advertised is the most important factor (other than aesthetic appeal) affecting user assessments; for Automotive, the brand is crucial. For Personal Finance category, the clarity of the description has a big impact on the user perception of the pre-click ad quality.

This study provides important insights into how users perceive the native ads. Next, we map these insights to engineered features used to predict the pre-click experience.

4. PRE-CLICK AD QUALITY FEATURES

We design two sets of features. The first set is inspired by the results of our crowd-sourcing study. We refer to this set as *cold-start features*, since they do not require prior knowledge about how users interact with the ad. We collect features mined from the ad creatives, including the ad copy, the image and the advertiser characteristics. An overview of the features together with their mapping to the preference

[8] We find similar results using grade 5 as our threshold.

[9] In general, pairwise statistical significance is not transitive. However, our results do not violate transitivity.

reasons is shown in Table 2. We then collect a second set of features, based on the *user behavior* (e.g. dwell time) after the ads were served, as shown in Table 3.

4.1 Cold-start Ad Features

4.1.1 Clarity

The clarity of the ad reflects the ease with which the ad text (title or description) can be understood by a reader. To describe this aspect, we measure the *readability* of the ad copy text with several readability metrics. From both the ad title and description, we compute Flesch's reading ease test, Flesch-Kincaid grade level, the Gunning fog index, the Coleman-Liau index, the Laesbarheds index and RIX index.[10] These metrics are defined according to a set of low-level text features, such as the number of words, the percentage of complex words, the number of sentences, number of acronyms, number of capitalized words and syllables per words. For completeness, we retain these low-level statistics as additional clarity features.

4.1.2 Trustworthiness

Another important aspect of ad quality is its *trustworthiness*, namely the extent to which users perceive the ad as reliable. We represent this dimension by analyzing different psychological reactions that users might experience when reading the ad creative. We mine information about the sentiment value of the text, its psychological incentives, and the language style and usage in the ad copy.

Sentiment and Psychological Incentives. Sentiment analysis tools automatically detect the attitude of a speaker or a writer with respect to a topic, or the overall contextual polarity of a text. To determine the polarity (positive, negative) of the ad sentiment, we analyze the ad title and description with SentiStrength [34], an open source sentiment analysis tool. For a sentence, SentiStrength reports two values, the probabilities (on a 5-scale grade) of the text sentiment being *positive* and *negative*, respectively.

The words used in the ad copy could have different psychological effects on the users [36]. To capture these, we resort to the LIWC 2007 dictionary [33], which associates psychological attributes to common words. For our purpose, we look at words categorized as *social*, *affective*, *cognitive*, *perceptual*, *biological*, *personal concerns* and *relativity*. For both the ad title and the description, we retain the frequency of the words that the LIWC dictionary associates with each of these seven categories as Psychological Incentives features.

Content Coherence. The consistency between ad title and ad description may also affect the ad trustworthiness. We capture this by calculating the cosine similarity between the bag of word vectors of the ad title and the ad description.

Language Style. To reflect the stylistic dimension of the ad text, we analyze the degree of *formality* of the language in the ad, using a linguistic formality measure [13] and a proprietary learned formality classifier. The linguistic formality weights different types of words, with nouns, adjectives, articles and prepositions as positive elements, and adverbs, verbs and interjections as negative. The in-house classifier is based on linguistic features designed on top of the SpaCy NLP toolkit,[11] such as text readability, n-gram counts, constituency, part-of-speech, lexical features, casing and punctuation, entity, subjectivity (TextBlob NLP)[12] and Word2Vec[13] features. We also include low-level features, such as the frequency of punctuation, numbers, "5W1H" words, superlative adjectives and adverbs.

Language Usage. To understand the language usage of an ad textual content, we parse the text using a proprietary content analysis platform (CAP). The CAP underlying classifiers are based on natural language processing techniques, modeling the general usage of the language. We are interested in two classifiers: *spam* and *hate speech*. The spam score [32] reflects the likelihood of a text to be of a spamming nature and utilizes a set of content and style based features. The hate speech score [9] captures the extent to which any speech may suggest violence, intimidation or prejudicial action against/by an individual or a group.

The ad title is written to grasp users attention. Advertisers often choose catchy word combinations, to persuade users to click on the ad creative. To measure the attractiveness of the ad title, we extract a set of features originally used to train a proprietary learned *click bait* classifier,[14] including a set of low-level features (e.g. whether the text contains slang or profane words), sentiment values and Bag-of-words. We also retain the frequency counts of words relating to slang[15] and profanity[16] as trustworthiness features.

4.1.3 Product/Service

Although quality is independent to relevance, some ad categories might be considered lower quality (offensive) than others, and features may be more important for some types of product/service, as indicated in Section 3.2.

Text. To capture the topical categories of the *product or service* provided by the ad, we use a proprietary text-based classifier (YCT) that computes, given a text, a set of category scores (e.g. sports, entertainment) according to a topic taxonomy (only top-level categories) [21]. We also add to this group the *adult* score as extracted from the CAP, that suggests whether the product advertised is adult-related such as dating websites [12].

Image. To understand the content of the ad creative from a visual perspective, we tag the ad image with the Flickr *machine tags*,[17] namely deep-learning based computer vision classifiers that automatically recognize the objects depicted in a picture (a person, or a flower). For each of the detectable objects, the Flickr classifiers output a confidence score corresponding to the probability that the object is represented in the image. Since tag scores are very sparse (an image shows few objects), we group semantically similar tags into topically-coherent tag clusters (e.g. dog, cat will fall in the *animal* cluster), and aggregate the raw tag confidence scores at a cluster level. Examples of the clusters include "plants", "animals". We also run a deep-learning proprietary *adult image detector*, and retain the output confidence score as an indicator of the adultness of the ad creative [29].

[10] Formulas for those readability tests: http://bit.ly/1MEgXJW.

[11] https://github.com/honnibal/spaCy

[12] https://textblob.readthedocs.org/en/dev/

[13] https://code.google.com/p/word2vec/

[14] Others: https://github.com/peterldowns/clickbait-classifier

[15] http://www.personal.psu.edu/users/p/x/pxb5080/slang.txt

[16] http://www.bannedwordlist.com/lists/swearWords.txt; http://www.cs.cmu.edu/~biglou/resources/bad-words.txt

[17] http://www.fastcolabs.com/3037882/how-flickrs-deep-learning-algorithms-see-whats-in-your-photos

User Reasons	Feature Type	Feature	Dim	Description	Feature Source
Clarity	**Readability**	*Flesch's reading ease test*	2	Combination of number of words per sentence and syllables per words	Ad Copy
		Flesch-Kincaid grade level	2	Combination of number of words per sentence and syllables per words	
		Gunning fog index	2	Combination of number of words per sentence and percentage of complex words	
		Coleman-Liau index	2	Combination of number of letters per words and average number of sentences per words	
		Laesbarheds index	2	Combination of number of words per sentence and number of long words (words over six characters)	
		RIX index	2	number of long words (words over six characters) per sentences	
		number of capitalized words	4	number of capitalized words, and whether text contains at least one capitalized words	
		number of acronyms	4	number of acronyms, and whether text contains at least one acronyms	
		words per sentence	2	number of words per sentence	
		percentage of complex words	2	complex words contain three or more syllables	
		syllables per words	2	Number of syllables per words	
Trustworthiness	**Psychology**	*Positive Polarity* [34]	2	Sentistrength positive polarity classification based on 298 positive terms in the sentiment word strength list	
		Negative Polarity [34]	2	Sentistrength negative polarity classification based on 465 negative terms in the sentiment word strength list	
		Aggregated Polarity [34]	2	Sum of Sentistrength positive and negative polarity for the overall polarity	
		Psychological Incentives [33]	14	Frequency of words relating to social, affective, cognitive, perceptual, biological, relativity, personal concerns in the LIWC dictionary	
	Content Coherence	*title-description similarity*	1	Similarity between texts of ad title and description	
	Language Style	*Formality* [13]	2	formality f-score based on the frequencies of different word classes (part-of-speech) and machine learning based formality classifier trained on various features	
		Punctuation	6	number of different punctuation marks, including exclaim point '!', question mark '?' and quotes	
		start with number	2	whether text starts with number	
		contain non-starting number	2	whether text contains number that does not start with the text	
		start with 5W1H	1	whether text starts with "what", "where", "when", "why", "who" and "how"	
		contain superlative	1	whether text contains a superlative adverb or adjective	
	Language Usage	*Spam* [32]	1	Likelihood of text to be classified as spam from CAP (trained on HTML web documents)	
		Hate speech [9]	1	Likelihood of text to contain abusive speech targeting specific group characteristics, such as ethnicity, religion, or gender, from CAP	
		Click bait	3	likelihood of text to be classified as click bait, exploiting a learned prediction model based on a set of low-level, sentiment and bag-of-words features	
		number of slang words	2	number of slang words used (defined in a word list)	
		number of profane words	2	number of profane words used (defined in a word list)	
Product/Service	**Content**	*YCT (text)* [21]	21	Likelihood of the most top level YCT (Yahoo Category Taxonomy, e.g. sports) the text to be classified from CAP	
		Adult (text) [12]	1	Likelihood of text to contain adult contents from CAP	
		Adult (image) [29]	1	Likelihood of image to contain adult related images (e.g. too much skin)	
		Image Object Taxonomy	1	Likelihood of image to contain objects within a given topical category (such as plant, man-made objects)	
		Image CNN classifier	50	Likelihood of image to contain deep learning based objects based on the second last layer of the Convolutional Neural Networks (CNN)	
Layout	**Readability**	*number of sentences*	2	Number of sentences	Ad Copy and Image
		number of words	2	Number of words	
	Composition	*Presence of Objects [27]*	9	Amount of saliency [15] in 9 image quadrants	
		Uniqueness [27]	1	Difference between the image spectral signal and the average spectrum of natural images	
		Symmetry [27]	1	Difference between the HOG [6] feature vectors of the image left-half and right-half	
		Depth Of Field [7]	12	Low DOF indicators based on haar wavelets	
		Image text detector [38]	1	Likelihood of image to contain text	
Aesthetic Appeal	**Colors**	*Contrast* [27]	1	Ratio between the sum of max and min luminance values and the average luminance	Ad Image
		H,S,V [22]	3	Average *Hue, Saturation, Brightness* computed on the the *whole* image	
		H,S,V (Central Quadrant) [22]	3	Average *Hue, Saturation, Brightness* computed on the *central quadrant*	
		H,S,V Color Histograms [22]	20	Histograms of H, S and V values quantized over 12, 3, and 5 bins	
		H,S,V Contrasts [22]	3	Standard deviation of the HSV Color Histograms distributions	
		Pleasure, Arousal, Dominance [22]	3	Based on average HSV combinations	
	Textures	*GLCM Properties [22]*	4	Entropy, Energy, Contrast, and Homogeneity of the Gray-Level Co-Occurrence Matrix	
	Photographic Quality	*Contrast Balance* [27]	1	Distance between original and contrast-normalized images	
		Exposure Balance [27]	1	Absolute value of the luminance histogram skewness	
		JPEG Quality[27]	1	No-reference quality estimation algorithm in [37]	
		JPEG Blockiness[27]	1	JPEG artifacts detection based on image re-compression.	
		Sharpness [27]	1	Sum of the image pixels after applying horizontal/vertical Sobel masks	
		Foreground Sharpness	1	Sum of the image pixels after applying horizontal/vertical Sobel masks on the salient image zones	
Brand	**Brand Quality**	*Advertiser Domain Pagerank* [26]	1	the WCC pagerank score of the top level domain of the ad landing page	Advertiser
		Advertiser Search Volume	2	the number of Yahoo search query volume given the advertiser name or the sponsored by label	

Table 2: Pre-click ad quality: summary of the features based on the ad creative (cold-start features).

User Behavior	Feature Type	Feature	Dim	Description	Feature Source
Engagement	**Pre-click**	*Click-through rate (CTR)*	1	the number of ad clicks divided by the number of ad impressions	User Behavior
	Post-click	*Dwell Time* [18]	1	the average dwell time of the ad landing page	
		Bounce Rate [18]	1	the fraction of sessions with ad landing page dwell time shorter than five seconds	

Table 3: Pre-click ad quality: summary of user engagement features of the advertisement.

To further capture the underlying semantics of the image, we get richer visual descriptions from the CNN-based Flickr classifiers. We extract a 4096-dimensional feature vector corresponding to the outputs of the 4096 neurons of the second last layer of the deep learning network generating the Flickr *machine tags*. To reduce dimensionality, we run feed-forward feature selection, and retain the top-50 discriminative *CNN features* for ad offensiveness detection.

4.1.4 Layout

Text. Since the ad format of the native ads served on a given platform is fixed, we capture the textual *layout* of the ad creative by looking at the length of the ad creative copy text (e.g. number of sentences or words).

Image. To quantify the composition of the ad image, we analyze the spatial layout in the scene using compositional visual features inspired by computational aesthetics research [27]. We resize the image to a squared matrix, and compute a *Symmetry* descriptor based on the gradient difference between the left half of the image and its flipped right half. We then analyze whether the image follows the photographic Rule of Thirds, according to which important compositional elements of the picture should lie on four ideal lines (two horizontal and two vertical) that divide it into nine equal parts, using saliency distribution counts to detect the *Object Presence* as in [27]. Finally, we look at the *Depth of Field*, which measures the ranges of distances from the observer that appear acceptably sharp in a scene, using wavelet coefficients as in [7]. We also include an image text detector to capture whether the image contains text in it [38].

4.1.5 Aesthetic Appeal

To explore the contribution of *visual aesthetics* for ad quality, we resort to computational aesthetics, a branch of computer vision that studies ways to automatically predict the beauty degree of images and videos. Computational aesthetics uses compositional visual features to train "beauty" classifiers. Similar to computational aesthetic studies [7, 27], we extract 43 compositional features from the ad images.

Color. Color patterns are important cues to understand the aesthetic value of a picture. To describe the color palette, we first compute a luminance-based *Contrast* metric, that reflects the distinguishability of the image colors. We then extract the average *Hue, Saturation, Brightness (H,S,V)*, by averaging HSV channels of the whole image and HSV values of the inner image quadrant, similar to [22, 7]. We then linearly combine average Saturation ($\bar{S}$) and Brightness ($\bar{V}$) values, and obtain three indicators of emotional responses, *Pleasure, Arousal and Dominance*, as suggested by [22]. In addition, we quantize the HSV values into 12 Hue bins, 5 Saturation bins, and 3 Brightness bins and collect the pixel occurrences in the HSV *Itten Color Histograms* [22]. Finally, we compute *Itten Color Contrasts* [22] as the standard deviation of H, S and V Itten Color Histograms.

Texture. To describe the overall complexity and homogeneity of the image texture, we extract the Haralick's features from the Gray-Level Co-occurrence Matrices, namely the *Entropy, Energy, Homogeneity, Contrast*, similar to [22].

Photographic Quality. These features describe the image quality and integrity. High-quality photographs are images where the degradation due to image post-processing or registration is not highly perceivable. To determine the perceived image degradation, we extract a set of simple image metrics originally designed for computational portrait aesthetics [27], independent of the composition, the content, or its artistic value. These are:

- *Contrast Balance*: This is the distance between the original image and its contrast-equalized version.
- *Exposure Balance*: To capture over/under exposure, we compute the luminance histogram skewness.
- *JPEG Quality*: When too strong, JPEG compression can cause disturbing blockiness effects. We compute here the objective quality measure for JPEG images from [37].
- *JPEG Blockiness*: This detects the amount of 'blockiness' based on the difference between the image and its compressed version at low quality factor.
- *Sharpness*: We detect the image sharpness by aggregating the edge strength after applying horizontal or vertical Sobel masks (Teengrad's method).
- *Foreground Sharpness*: We compute the Sharpness metric on salient image zones only.

4.1.6 Brand

These features reflect the advertiser characteristics. Following the findings of our user study, we hypothesize that the intrinsic properties of the advertiser (such as the brand) have an effect on the user perception of ad quality. We extract two features: *domain pagerank* and *search volume*. The *domain pagerank* is the pagerank score [26] of the advertiser domain for a given ad landing page. This is obtained by mining the web crawl cache (WCC) data, which contains the pagerank score for any given URLs crawled. The *search volume* reflects the raw search volume of the advertiser within a big commercial search engine. This represents the overall popularity of the advertiser and its product/service.

4.2 User Behavior Features

All the above features are cold-start, i.e. they do not consider the interactions ad-users after the ad is consumed. However, after serving the ad, very informative user behavior signals can be collected. We explore whether these signals contribute in determining the pre-click quality of ads. To this end, we collect user behavior features related to the pre-click experience (click-through rate on the ad creative). Moreover, we look at the user behavior with respect to the post-click experience, using bounce rate and average dwell time, which are good proxy of the quality of ad landing pages [18]. Our intuition is that a bad ad creative is likely to have a bad landing page.

5. EXPERIMENTS

In this section, we build a model for ad offensiveness prediction based on the metrics (learning target) defined in Section 2 and the features described in Section 4.

5.1 Dataset

We use a sample of 28,664 ads served over a 1.5 month period on aggregator news streams operated by a major Internet company, Yahoo. We select ads with at least 10 clicks (as we use dwell time and bounce rate as features). To counter the sparsity problem (many ads receive few feedbacks), we collect and aggregate ad feedbacks from different streams. We run a cross-platform consistency check, and saw that

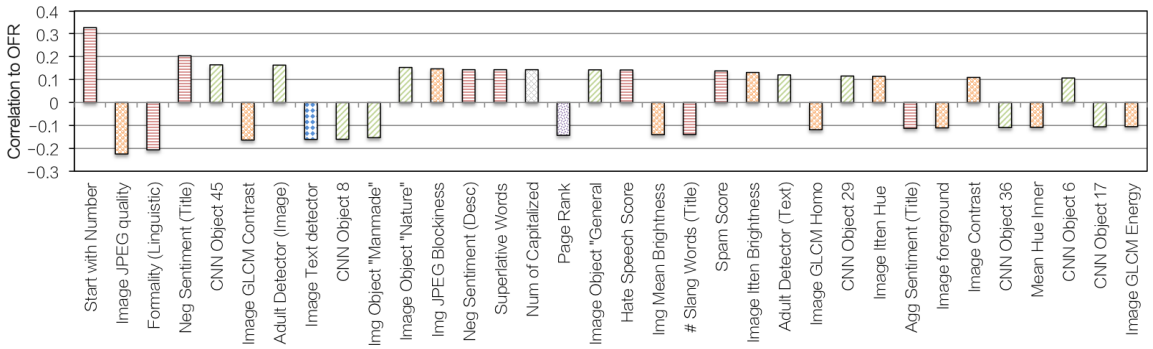


Figure 3: Correlation between ad quality features and offensive feedback rate.

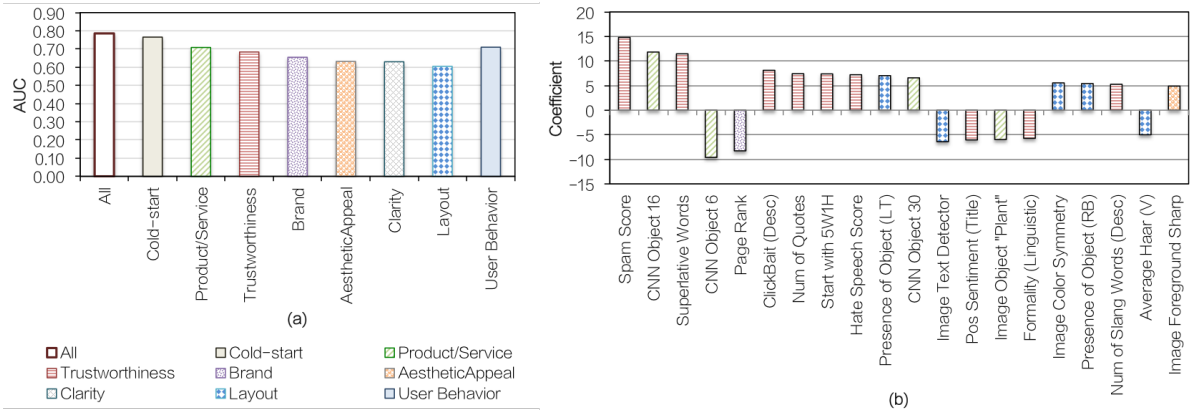


Figure 4: Predicting Pre-click Ad Quality: (a) Feature Subset; (b) Top Significant Feature Coefficients.

the ad feedback rate was comparable across streams (Spearman correlations over 0.70); if an ad was bad, it was so everywhere.

5.2 Feature Analysis

To better understand our dataset, we analyze the extent to which each feature individually correlates with offensive feedback rate (OFR). We report the Spearman correlation between each feature and OFR in Figure 3 (top-correlated features only). Many features, such as visual features (e.g. JPEG compression artifacts), text features (e.g. whether the title contains negative sentiments), and advertiser features (e.g. advertiser landing page domain page rank) correlate with OFR. Interestingly, we see that an ad title starting with a number is likely to belong to an offensive ad. Through manual inspection, we found that many offensive ads' titles indeed tend to start with numbers, for example "10 most hated...".

Overall, the correlations between each single feature and OFR are relatively weak, thus making it difficult to predict ad offensiveness using individual features. Therefore, we propose next a learning framework that combines ad quality features to predict the OFR of native ads.

5.3 Ad Quality Model

Prediction Model. We use the offensive feedback rate as our proxy to determine the "low quality" ads. We treat the offensiveness prediction as a binary classification task and consider as *positive* all ads that fall within the fourth quartile of the OFR distribution (the offensive ads). For *negative* examples, we randomly sample the remaining ads. To ensure reliable OFR within the positive training examples, we select ads marked as offensive at least five times to eliminate random or unintentional feedback.

We use logistic regression to learn the offensiveness model. Logistic regression is parameterized through a weight vector w. We assume that the posterior pre-click quality probabilities can be estimated through a linear combination of the input features x, passed through a sigmoidal function:

$$P(y=1|x) = f(x,w) = \frac{1}{1+exp(-x^T w)}$$

To estimate the parameters w, we minimize the loss function

$$\min_{w \in R^K} \frac{1}{N} \sum_{i=1}^{N} m_i (y_i - f(x_i, w))^2 + \lambda ||w||_1$$

where the hyper-parameter λ controls the L1-regularization, introduced to induce sparsity in the parameter vector, thus reducing the feature space to a subset of discriminative features. To overcome the problem of imbalanced training set (there are more "non-offensive" than offensive ads), we use the SMOTE [3] method. To over-sample the minority class (offensive), we generate synthetic examples in the neighborhood of the observed offensive ads, by interpolating between examples of the same class.

Given the trained logistic regression model, we estimate the posterior pre-click ad quality probabilities as $f(x_i, w) \in [0,1]$, and obtain the predicted class y_i (offensive, not offensive) by thresholding the obtained probabilities: $y_i = sign(f(x_i, w) - \theta)$, where threshold θ is usually set to 0.5.

However, θ can be chosen anywhere between 0 and 1 to ensure desired precision. We use 5-fold cross-validation to train and test the model, and report AUC (area under the ROC curve) as our performance metric.

Results The performances of our framework expressed in terms of AUC values are shown in Figure 4(a). Using all features, we reach an AUC of 0.79. Using the cold-start features, this value is 0.77, whereas using only behavior features, we reach an AUC of 0.70.

To understand how each feature type performs, we group the features according to the categorization described in Section 4 (see Table 2), and run a model based on each feature category. All feature types are helpful, and in particular those related to product/service, trustworthiness, advertiser brand and aesthetic appeal. Results are generally consistent with the findings of our user study results reported in Section 3, if we exclude the weaker importance given by our framework to the image aesthetic appeal. This might be due to the fact that, although aesthetic appeal is crucial for ad quality in general, it may be less crucial in terms of "offensiveness" of the ads, compared to the effect of product and trustworthiness (including language usage and style).

To further understand which cold-start features contribute most to the model, we plot the regression coefficients for the subset of top features that are significant (p-value < 0.01) in Figure 4(b). The top features relate to trustworthiness (language usage, style and sentiment), the product/service provided, the brand (advertisers' page rank) and the layout (composition of the ad creative image). In particular, among the highest indicator of offensiveness, we can find the *Spam Score* and the use of superlative words.

5.4 Quantile Regression Analysis

To evaluate how the features react when predicting different quantiles of OFR, we perform an in-depth study using quantile regression analysis [17]. This provides insights into how the regression coefficient of each feature (its importance) varies according to the different quantiles of OFR.

To explore the importance of various features, we use as input to the quantile model the different features. We estimate the regression coefficients (black dashed lines in Figure 5) for quantiles 0.1 to 0.9 (in steps of 0.1) using the bootstrap method [17]. We present the results for those feature subsets exhibiting significant changes in Figure 5. All features are normalized to compare their effects.

Several trends emerge. First, feature coefficients tend to vary more significantly at higher quantiles of OFR. This is particularly visible for those features that are most discriminative for predicting ad offensiveness, suggesting that those features are more useful in identifying the "most offensive" ads. For example, the *spam* score and the *adult* detector are among the most discriminative features to predict ad offensiveness for ads within high quantile (0.6-1.0) of OFR. This is somehow expected, since such features quantify the likelihood of the ad to be spam or to contain adult contents, often triggering high levels of offensiveness.

It is also interesting to observe that the sentiment of the text (positive or negative) and the image characteristics behave differently when predicting different OFR quantiles. The negativity of the sentiment expressed in the ad description and hate speech language usage becomes less important as the ads lie within the more offensive quantile. On the other hand, positive text sentiment or visual features such as Image GLCM Entropy (indicating highly textured images) or the Average Haar (indicating the sharpness of the image foreground) are very strong indicators of highly offensive ads. Again, we also observe that as ads become more offensive, it is more likely that their title starts with a number. Finally, we can see that some feature coefficient tend to stay stable across OFR quantiles, such as linguistic formality feature as shown in Figure 5.

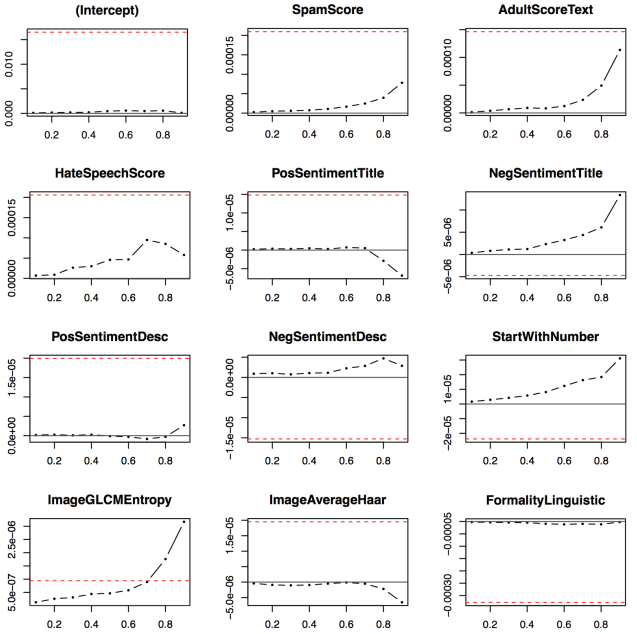


Figure 5: Quantile regression analysis of ad quality features to predict OFR.

5.5 Online A/B Testing Evaluation

A version of the pre-click model with a subset of the features investigated in this paper was deployed in an A/B test on a 1% traffic on several Yahoo news aggregator streams. Yahoo offers users news content in the form of streams, with native ads slotted at various positions within the stream. The features included a subset of readability features (related to *clarity*), spamscore (related to *trustworthiness*) and adultscore (related to *product/service*), as our aim was to test step-by-step various sets of increasingly rich features.

The model was deployed as a filtering task, and as such was used to annotate ads into "good" versus "bad". The predicted pre-click ad quality probability scores output from the logistic regression were therefore converted into a Boolean score that discriminates between "bad" ads (above threshold) and "good" ads (below threshold). We used a threshold (probability of being marked as offensive) based on the analysis of the ad distributions. Only ads annotated as good were allowed to be served.[18]

We show in Table 4 the performance, i.e. difference in ad offensive feedback rates, of the bucket during a period of 6 weeks. We report the results for both the mobile and the desktop cases. We observe that, by adding to the system our pre-click ad quality model, we have significantly (paired t-test with $p < 0.01$) reduced the offensive feedback rate. The reduction is even higher in the mobile context, likely to do with the fact that, in that context, the experience is

[18]For confidentiality, we do not share the percentage of ads annotated as bad versus good, nor the threshold used.

Table 4: Online bucket performance based on a pre-click ad quality model with a subset of features.

Evaluation	Mobile	Desktop
δOFR	-17.6%	-8.7%

restricted to the stream, whereas in the desktop case, the stream is only one part of the experience.

6. RELATED WORK

Our work belongs to the field of computational advertising, which studies ways to promote relevant and quality ads using a computational approach. Many effective algorithms have been developed for this task , especially in the context of display ads and sponsored search [8]. Our contributions bring two fundamental different elements to this body of work: the learning target and the ad features.

Learning target. Many existing works build systems to improve ad *relevance* in sponsored search, aiming to maximize CTR [14], dwell time [1] and ultimately revenue [40], or design systems for display ads aiming to optimize CTR [4], viewability [35] and conversion rate [19]. These works mainly focus on ad *relevance*, and aim to maximize ad CTR or similar compounding metrics (e.g. conversion rate) that measure both relevance and quality.

Our work focuses on advertising *quality*, explicitly taking user ad feedback into account to estimate ad quality, here quantified in terms of offensiveness. Although some user studies [5, 11, 25] have looked at the effect of ad quality on user perception or engagement, little is known on how to measure and optimize ad quality in large-scale, especially in the context of native advertising. For example, as shown in Section 2, the commonly used CTR-based metrics might not be good indicators of quality. A recent work [18] utilizes dwell time and bounce rate as learning targets of post-click quality prediction. However, to our knowledge, this paper is the first work focusing on predicting bad quality ads using pre-click metrics, focussing solely on quality.

Ad Features. Previous works in computational advertising have addressed the problem of designing ad-specific features for various tasks. For example, [28] propose to use a variety of textual features to capture appearance, attention, reputation and relevance for CTR prediction. In our work, we go beyond text-only features, using visual features extracted from the ad creative image.

Similar to our work, to predict CTR for display ads, [4] and [23] propose to exploit a set of hand-crafted image and motion features and deep learning based visual features, respectively. Our work differs to these in two ways. Our work focuses on native advertising, intrinsically very different from display advertising. Unlike display ads, native ads follow a standard format dictated by the platform, thus constraining the diversity of computable visual features, making predicting pre-click quality harder. In addition, the *multimodality* component is different. Unlike native ads, display ads do not contain textual description surrounding the main image, thus directing previous works towards the analysis of the ad visuals only. In our work, we design a wide set of textual features and combine them with a set of visual features in a complete, multimodal model for ad quality prediction.

Finally, our multimodal model for ad quality prediction also includes various textual and visual features (such as text formality or image foreground sharpness) that have not been evaluated before in computational advertising.

7. CONCLUSIONS

We presented an approach that aimed at identifying low quality ads, from the pre-click experience. Our focus was native advertising, where the ads served reproduce the look-and-feel of the platform in which they appear. As our proxy of low pre-click quality, we used the ad offensiveness annotations extracted from an ad feedback mechanism used by a large Internet company. We showed that ad offensiveness feedback rate, which we use as our learning target, is different to CTR, and that it is important to deploy computational approaches to detect offensive ads, independently of whether they are clicked or not.

We then carried a crowd-sourced user study to understand how users perceive the quality of ads. This allowed us to identify several reasons, for example related to aesthetics, brand and clarity. From these, we engineered a large set of features ranging from page rank to capture brand awareness, to visual features to incorporate aesthetics, to readability to characterize clarity. We also discussed the importance of these features in identifying offensive ads. Using these features and ad offensiveness feedback rate as our learning target, we developed a learning framework able to predict with high accuracy the probability of an ad being offensive. We deployed the framework with a subset of the features, and already saw significant positive effects on reducing offensive the feedback rate on both mobile and desktop. This paves the way to experiment with additional features, in particular those related to visual aesthetics (e.g. image quality).

In the future we will focus on broadening our study on advertising effectiveness from a user perspective, developing metrics and learning models able to embrace a more complex and multifaceted notion of ad quality, thus going beyond the simple but significative dimension of offensiveness. For example, we could apply ad offensiveness predictors for large-scale filtering tasks, by re-visiting our dataset and embedding our approach in a learning to rank framework [20]. Similarly, we could flip the perspective of our learning framework, and, rather than demoting low-quality, build systems to promote high quality native ads (e.g. native ads in magazines). Moreover, we would like to better exploit the precious data collected through ad feedback tools, incorporating, for example, the credibility of each feedback, or studying the relation between ad quality perception and the various hiding options. On a longer term, our investigation will focus on how to jointly optimize ad quality and relevance for a complete, pleasant, and effective user experience.

Finally, in this study, we did not consider the impact of our approach (filtering out low quality pre-click ads) on CTR and revenue. Our goal was solely to reduce offensive rates. In the future, it will be important to account for the effect on both short- and long-term revenue. Initial analysis of the data shows that many of the ads identified as of low quality are associated with low cost-per-click values, which looks promising.

Acknowledgments We thank Adi Unger and Tamar Lavee for the deployment of the model for the online A/B testing evaluation, the NLP/Discourse and Dialogue team at Yahoo Labs for providing several NLP features (e.g. formality classifier) and Yahoo Flickr Vision team for providing some of the visual features.

8. REFERENCES

[1] J. Attenberg, S. Pandey, and T. Suel. Modeling and predicting user behavior in sponsored search. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1067–1076. ACM, 2009.

[2] J. P. Benway. Banner blindness: The irony of attention grabbing on the world wide web. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 42, pages 463–467. SAGE Publications, 1998.

[3] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, pages 321–357, 2002.

[4] H. Cheng, R. v. Zwol, J. Azimi, E. Manavoglu, R. Zhang, Y. Zhou, and V. Navalpakkam. Multimedia features for click prediction of new ads in display advertising. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 777–785. ACM, 2012.

[5] H. Cramer. Effects of ad quality and content-relevance on perceived content quality. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 2231–2234, 2015.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[7] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *European Conference on Computer Vision*, volume 3953 of *Lecture Notes in Computer Science*, pages 288–301. Springer Berlin Heidelberg, 2006.

[8] K. Dave and V. Varma. *Computational Advertising: Techniques for Targeting Relevant Ads*. now Publishers, 2014.

[9] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati. Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web Companion*, pages 29–30. International World Wide Web Conferences Steering Committee, 2015.

[10] A. Goldfarb and C. Tucker. Online display advertising: Targeting and obtrusiveness. *Marketing Science*, 30(3):389–404, 2011.

[11] D. G. Goldstein, R. P. McAfee, and S. Suri. The cost of annoying ads. In *Proceedings of the 22nd international conference on World Wide Web*, pages 459–470. International World Wide Web Conferences Steering Committee, 2013.

[12] M. Hammami, Y. Chahir, and L. Chen. Webguard: Web based adult content detection and filtering system. In *Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on*, pages 574–578. IEEE, 2003.

[13] F. Heylighen and J.-M. Dewaele. Formality of language: definition, measurement and behavioral determinants. *Interner Bericht, Center ?Leo Apostel?, Vrije Universiteit Brüssel*, 1999.

[14] D. Hillard, S. Schroedl, E. Manavoglu, H. Raghavan, and C. Leggetter. Improving ad relevance in sponsored search. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 361–370. ACM, 2010.

[15] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[16] S. IPG. Native ad research from ipg & sharethrough reveals that in-feed beats banners. In *Industrial report, http://bit.ly/1QvB387*, 2015.

[17] R. Koenker. *Quantile regression*. Number 38. Cambridge university press, 2005.

[18] M. Lalmas, J. Lehmann, G. Shaked, F. Silvestri, and G. Tolomei. Promoting positive post-click experience for in-stream yahoo gemini users. In *KDD'15 Industry Track*. ACM, 2015.

[19] K.-c. Lee, B. Orten, A. Dasdan, and W. Li. Estimating conversion rate in display advertising from past erformance data. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 768–776. ACM, 2012.

[20] T.-Y. Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009.

[21] T.-Y. Liu, Y. Yang, H. Wan, H.-J. Zeng, Z. Chen, and W.-Y. Ma. Support vector machines classification with a very large-scale taxonomy. *ACM SIGKDD Explorations Newsletter*, 7(1):36–43, 2005.

[22] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the international conference on Multimedia*, pages 83–92. ACM, 2010.

[23] K. Mo, B. Liu, L. Xiao, Y. Li, and J. Jiang. Image feature learning for cold start problem in display advertising. In *IJCAI'15 Machine Learning Track*, 2015.

[24] H. L. O'Brien and E. G. Toms. The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology*, 61(1):50–69, 2010.

[25] K. O'Donnell and H. Cramer. People's perceptions of personalized ads. In *Proceedings of the 24th International Conference on World Wide Web Companion*, pages 1293–1298. International World Wide Web Conferences Steering Committee, 2015.

[26] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: bringing order to the web. 1999.

[27] M. Redi, N. Rasiwasia, G. Aggarwal, and A. Jaimes. The beauty of capturing faces: Rating the quality of digital portraits. In *IEEE International Conference on Automatic Face and Gesture Recognition 2015*. IEEE, 2015.

[28] M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: estimating the click-through rate for new ads. In *Proceedings of the 16th international conference on World Wide Web*, pages 521–530. ACM, 2007.

[29] C. X. Ries and R. Lienhart. A survey on visual adult image recognition. *Multimedia tools and applications*, 69(3):661–688, 2014.

[30] M. Sanderson, M. L. Paramita, P. Clough, and E. Kanoulas. Do user preferences and evaluation measures line up? In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 555–562. ACM, 2010.

[31] F. Scholer, D. Kelly, W.-C. Wu, H. S. Lee, and W. Webber. The effect of threshold priming and need for cognition on relevance calibration and assessment. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 623–632. ACM, 2013.

[32] N. Spirin and J. Han. Survey on web spam detection: principles and algorithms. *ACM SIGKDD Explorations Newsletter*, 13(2):50–64, 2012.

[33] Y. R. Tausczik and J. W. Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.

[34] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, 2010.

[35] C. Wang, A. Kalra, C. Borcea, and Y. Chen. Viewability prediction for online display ads. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 413–422. ACM, 2015.

[36] T. Wang, J. Bian, S. Liu, Y. Zhang, and T.-Y. Liu. Psychological advertising: exploring user psychology for click prediction in sponsored search. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 563–571. ACM, 2013.

[37] Z. Wang, H. R. Sheikh, and A. C. Bovik. No-reference perceptual quality assessment of jpeg compressed images. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 1, pages I–477. IEEE, 2002.

[38] V. Wu, R. Manmatha, and E. M. Riseman. Textfinder: An automatic system to detect and recognize text in images. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (11):1224–1229, 1999.

[39] C. Yun Yoo and K. Kim. Processing of animation in online banner advertising: The roles of cognitive and emotional responses. *Journal of Interactive Marketing*, 19(4):18–34, 2005.

[40] Y. Zhu, G. Wang, J. Yang, D. Wang, J. Yan, J. Hu, and Z. Chen. Optimizing search engine revenue in sponsored search. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 588–595. ACM, 2009.