Swipe and Tell: Using Implicit Feedback to Predict User Engagement on Tablets

KLAAS NELISSEN, KU Leuven
MONIQUE SNOECK, KU Leuven
SEPPE VANDEN BROUCKE, KU Leuven
BART BAESENS, KU Leuven and University of Southampton

When content consumers explicitly judge content positively, we consider them to be engaged. Unfortunately, explicit user evaluations are difficult to collect, as they require user effort. Therefore, we propose to use device interactions as implicit feedback to detect engagement.

We assess the usefulness of swipe interactions on tablets for predicting engagement, and make the comparison with using traditional features based on time spent.

We gathered two unique datasets of more than 250,000 swipes, 100,000 unique article visits, and over 35,000 explicitly judged news articles, by modifying two commonly used tablet apps of two newspapers. We tracked all device interactions of 407 experiment participants during one month of habitual news reading.

We employed a behavioral metric as a proxy for engagement, because our analysis needed to be scalable to many users, and scanning behavior required us to allow users to indicate engagement quickly.

We point out the importance of taking into account content ordering, report the most predictive features, zoom in on briefly read content and on the most frequently read articles.

Our findings demonstrate that fine-grained tablet interactions are useful indicators of engagement for newsreaders on tablets. The best features successfully combine both time-based aspects and swipe interactions.

CCS Concepts: • **Computing methodologies** → *Learning from implicit feedback*; • **Applied computing** → *Publishing*; • **Human-centered computing** → *Tablet computers*;

Additional Key Words and Phrases: User Engagement; Implicit Feedback; Tablets; Dwell Time; Touch Interactions; Newspaper; Online News, Content Ordering; Position Bias; Briefly Read Content; Frequently Read Content.

ACM Reference format:
Klaas Nelissen, Monique Snoeck, Seppe vanden Broucke, and Bart Baesens. 2017. Swipe and Tell: Using Implicit Feedback to Predict User Engagement on Tablets. *ACM Transactions on Information Systems* 1, 1, Article 1 (June 2017), 27 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

User engagement is defined as the quality of the user experience that emphasizes the positive aspects of interacting with an online application [Lalmas et al. 2014]. Users are engaged when they appreciate the content to which they have given their attention. Identifying when users are engaged is interesting because it provides content creators with insights on how their products are used, and so it can be used to improve the offering towards users. At a small scale, we could just ask users to judge the content they consume and thus get accurate explicit user evaluations. And although explicit user judgments are the best measures for assessing relevance, it requires a high

The authors would like to thank Twipe for their cooperation with this research. This work was facilitated by iMinds and funded by VLAIO (grant number 140655).
Correspondence concerning this article should be addressed to Klaas Nelissen (e-mail: Klaas.Nelissen@kuleuven.be).
2017. 1046-8188/2017/6-ART1 $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

cognitive effort [O'Brien and Toms 2010]. Moreover, in online applications, this explicit feedback is always given voluntarily, and thus not scalable to a large numbers of users. This study fits in the research looking for better proxies for explicit user evaluations which can be used to improve large-scale measurements of user engagement.

One way user engagement has been measured at large scale is by tracking how much time users spend with content. But the time spent (i.e., dwell time) does not necessarily indicate appreciation. A user may spend 30 seconds reading the first half of a text attentively, or may be skimming through the whole text, scanning for relevant information. So there is a need for finer measures for user engagement, and this has been proven successful in web search on computers where mouse interactions and scrolling behavior could be used for identifying document relevance [Guo and Agichtein 2012] (taking into account that document relevance is not the same as user engagement). A better web search result ranking could be achieved on mobile devices by taking into account fine-grained swipe interactions [Guo et al. 2013b; Huang et al. 2011]. Recent research has shown that users' experiences are different on different devices, and earlier gained insights might not be transferable across devices [Huang and Diriye 2012]. While other studies have primarily focused on web search on computers, this study extends the current research to the context of news reading on tablets.

In web search, the order of presenting the results to a query has a very large impact on the click through rate [Agichtein et al. 2006b]. In a newspaper, content is also presented chronologically, in an order chosen by the editors. This decision about when to present which content to the reader is a key aspect of the newspaper creation process, which the editors spend a lot of time and effort on. It is therefore interesting to look into whether the ordering of the content in the context of a tablet app for a digital newspaper has an impact on engagement.

Another interesting question to ask is how to detect engagement for content which is read for only a short period of time. In general, when considering interactions or experiences of a short duration, the approach of using *time spent* will not work anymore and alternative approaches are required. Different interaction features might play a different role in this use case. Also, for each piece of content a reader comes across, the reader makes an (unconscious) decision about whether to spend more time with it or not. Editors are especially interested in those situations where a reader only interacts briefly with some content, but still judges that content positively. Because editors optimize for engaging content, it is interesting to investigate which interaction behaviors lead to readers judging content positively which they have only read briefly.

The final question we study concerns the difference between articles which are frequently read and those which are not. We repeat the analysis for the 25% most frequently read articles. From discussions with the newspaper editors, we learned that they spend relatively more time analyzing and discussing these more popular articles, trying to find out why these articles work so well. The most frequently read articles also function as a common divisor across the whole user population, thereby giving editors insight into the preferences of their reader base. The most important features for predicting engagement with these most frequently read articles might also be different.

The research questions of this paper are:

> **R.Q. 1:** How do fine-grained swipe interactions (as implicit feedback features) compare to time-based features in terms of performance for predicting user engagement in the context of news reading on tablets, and which are the most important features?
> **R.Q. 2:** What is the effect of the order in which the content is presented?
> **R.Q. 3:** How useful are fine-grained swipe interactions for predicting engagement with briefly read content, thereby taking into account that time-based features are probably not useful anymore?

R.Q. 4: To what extent do the results change when we consider only the most frequently read articles?

To find an answer to these research questions, we did experiments with people who read the digital newspaper on a tablet app. We instrumented two apps to track every user interaction, added an in-app feedback mechanism, and asked users to give feedback when they found certain content engaging. For each article in the newspaper, users could give a thumbs up or down, so we obtained a large set of explicitly judged articles.

We consider a user to be engaged with an article when she gives a thumbs up on that article. Admittedly, this is a simplistic behavioral measure which functions as a proxy for user engagement and which does not capture the holistic nature of user engagement as discussed by O'Brien and Toms [2008; 2010]. However, this metric does satisfy our requirements of allowing large-scale measurements of user engagement which are scalable to all users, and it demands almost no user effort, so the metric allows users to quickly give a thumbs up to newspaper articles they were just scanning over. Furthermore, a behavioral metric is also easily embeddable in other apps. We further address in the methodology section why choosing for a simplistic behavioral metric is the best option for this study and why alternative methods based on a lengthy survey are not feasible.

We created a large number of interaction features to capture the user behavior while reading, and used these features in logistic regression models to predict whether a user will judge an article positively or not.

In summary, we make the following contributions:

- We extend the current research on scalable measurements for user engagement to the context of news reading on tablets.
- We contrast the usefulness of device interactions as implicit features versus time-based features for predicting user engagement.
- We illustrate that the order in which content is presented has an impact on user engagement predictions.
- We discuss user engagement predictions for briefly read content and for the most frequently read articles, showing that different types of features perform differently in each of these two specific settings (which have not been analyzed separately before).

2 RELATED WORK

Song et al. [2013a] make the point that user behavior on tablets is not only different from user behavior on computers, but also from user behavior on smartphones. They suggest that each device should be treated differently, and that insights are not necessarily transferable across devices. Content which causes engagement on computers or smartphones is not necessarily also engaging on tablets [Lu et al. 2014]. Huang and Diriye [2012] argue in a position paper that touch events have a different meaning than cursor events but that they have great potential in helping to better understand user experiences. They propose to focus on tracking the viewport. This is the part of the page the user is currently seeing, and is more useful on smaller screens such as smartphones and tablets. Several features included in our analysis are based on the viewport.

2.1 The usefulness of device interactions

There is little empirical research which specifically focuses on touch interactions for detecting user engagement, or more generally, for identifying positive aspects of the user experience. Most closely related to our work is the study by Guo et al. [2013a], which shows that web search result rankings can be significantly improved by taking into account touch interactions. The authors conducted an experiment where users were asked to answer a number of questions by searching the web, and while every touch interaction during the search was captured, the users also explicitly rated the relevance of every page they visited. Two of the most useful features in their study are the swipe frequency (which is the number of swipes on a page divided by the dwell time on that page) and the maximum inactive time between two touch interactions. They find that more and faster swipes are negatively correlated with document relevance, as they indicate scanning behavior. In contrast, slow swiping and long periods of inactive time suggest that users are paying attention and actively reading the current web page.

There are more studies which do not specifically focus on touch interactions, but show the value of implicit interactions for estimating appreciation of content, document relevance, or user engagement. Several studies in both the domains of information retrieval [Agichtein et al. 2006a; Fox et al. 2005; Guo and Agichtein 2012; White et al. 2005] and recommender systems [Konstan et al. 1997; Lee and Park 2007; Liu et al. 2010] have shown that implicit interactions are useful for distinguishing document relevance. Based on implicit feedback, Guo and Agichtein [2008] could in one of their earlier studies identify whether a searcher had an intent to purchase or was just browsing for information. In another study by the same authors, they prove that incorporating post-click searcher behavior (such as scrolling and cursor movements) in addition to dwell time and clickthrough statistics can improve estimates of document relevance [Guo and Agichtein 2012]. Their analysis asserts that slow gestures might be indicative of reading, while faster mouse gestures might characterise a navigational pattern to locate certain information of interest in the text. Agichtein et al. [2006b] show that implicit feedback can be of even more value if the features are modeled as deviations from expected user behavior. We also include deviational features in our current study.

Other studies show that using fine-grained mouse interactions offer a scalable way to infer user attention on web pages [Claypool et al. 2001b; Huang et al. 2011]. Huang et al. [2011] did a study where they correlate cursor movements on web pages with explicit relevance judgments of users. They show that incorporating these fine-grained cursor interactions can improve estimates of document relevance. In their experiments, the mouse hover rate is the feature which correlates best with human relevance judgments. In contrast, duration of mouse hovers correlates negatively with relevance here, while in other studies such as the one by Claypool et al. [2001b], cursor travel time is a positive indicator of web page relevance. Unfortunately, these features do not have their equivalent in terms of tablet interactions.

Navalpakkam and Churchill [2012] use mouse cursor interactions to predict whether the reading experience of the user is pleasant or not significantly better than normal. They report that long and frequent mouse visits on text are strong predictors of an unpleasant experience. Speicher and Gaedke [2013] do a similar study which results in their end-to-end system TellMyRelevance. The system learns relevance models by automatically tracking and analyzing client cursor interactions.

Arapakis et al. [2014a] model a large set of features based on mouse interactions with the goal of developing a taxonomy of mouse patterns for determining interestingness of web pages. They include more than 60 features describing how the mouse was used. Only features based on speed, and minimum, average, and total distance are significant. The already mentioned study by Guo and Agichtein [2012] finds similar results, where frequency and speed correlate with document relevance. Lagun et al. [2014] recently took this a step further, using dynamic time warping to automatically identify cursor motifs (frequent subsequences) which could then be used as features for more accurate estimations of relevance. Shapira et al. [2006] find that mouse travel distance is a worse indicator than the ratio of mouse movement to reading time for document relevance.

The evidence of using only page dwell time for inferring relevance shows mixed conclusions [Fox et al. 2005; Guo et al. 2013a; Lagun and Lalmas 2016; Yi et al. 2014]. The correlation between time spent and relevance is often, but not always, significantly positive [Liu et al. 2016]. Early research shows that there is a strong tendency that users spend more time on interesting rather than uninteresting news articles [Claypool et al. 2001b; Morita and Shinoda 1994]. However, dwell time is for example not the best indicator for page quality in the study done by Shapira et al. [2006].

In summary, past research suggests that using fine-grained interactions in addition to features based on *time spent* proves to be useful for explaining document relevance on computers. However, none of these studies which use implicit feedback make the difference between briefly or long read content, or focus on the most frequently accessed items. Most previous experiments took place on computers.

2.2 Defining and measuring user engagement

O'Brien and Toms [2008; 2010] did the fundamental work of constructing a good definition for user engagement as well as developing a valid and reliable 31-item survey. They identified six distinct attributes of engagement: perceived usability, aesthetics, focused attention, felt involvement, novelty, and endurability. Their findings indicate that these attributes are highly intertwined, and that engagement is both a process and a product of interaction which can vary in intensity over the course of an experience. O'Brien also situates these findings in the context of mobile devices in a different study [O'Brien et al. 2013].

Other research suggests that there is not one best approach to measure user engagement, but that the most suitable measurement method depends on the online experience which is being studied [Lehmann et al. 2012]. The overview by Lalmas, O'Brien, and Yom-Tov [2014] describes three different measurement methods, each with its own advantages and drawbacks: self-reports, physiological signals, and behavioral metrics.

Surveys suffer from subjectivity and are hard to administer at massive scale. Physiological signals such as EEG or eye-trackers offer the most objective measurement method, but the need for specialized equipment limits their practical use outside research [Lalmas et al. 2014]. Only behavioral metrics allow researchers to collect data from all users of a service with almost no user effort, which is one of the requirements for our current study. These behavioral metrics are unable to explain *why* users find something engaging, they can only act as a proxy for user engagement [Lehmann et al. 2012].

Using behavioral metrics as proxies for user engagement is also done by Song et al. [2013b] and Drutsa and Serdyukov [2015]. In Song et al. [2013b], the authors develop a machine learning model which can predict drops in user engagement (as measured by behavioral metrics) on the long term by having previously purposefully degraded the relevance of returned web search results. The starting point of another study by Lagun and Lalmas [2016] is the acknowledgement of the limitations of dwell time as a metric for user engagement, specifically because dwell time can not tell whether a user is paying attention or not. Using viewport data from a computer they come up with four scalable behavioral metrics which capture different levels of intensity of engagement: bounce, shallow engagement, deep engagement and complete engagement. Their unit of analysis is one news article, but there is no ground truth of engagement provided by a user. Another recently proposed behavioral metric by Dupret and Lalmas [2013] is absence time, which is defined as the time between two user visits. While the results of this study are promising, this metric is not relevant for our current research because we do not consider engagement levels over different reading sessions.

Arapakis et al. [2014b] investigate user engagement in online news on computers. They do not use any behavioral metrics based on user interactions, but instead use eye tracking as the objective measure for user engagement, and use surveys to determine the interestingness of news articles, among other things. They find that the level of focused attention is determined by the perceived interestingness of the news article. This finding is corroborated in a study by McCay-Peet et al. [2012], where a user's self-reported level of interest in a topic is found to be a good predictor for self-reported focused attention.

In summary, simple behavioral metrics are frequently employed as proxies for user engagement. In fact, when the measurement method for user engagement is required to be scalable to all users, behavioral metrics are the only viable method. In the domain of information retrieval, identifying document relevance can be done by asking the user only one question - whether the presented result was deemed relevant or not. User engagement is harder to measure, as it covers several distinct aspects of the user experience, is formed in the long run, and often does not follow from a goal-oriented experience, which also makes it harder to evaluate [Lalmas et al. 2014].

The most advanced studies in measuring user engagement try to combine different measurement methods to better measure engagement. O'Brien and Lebow [2013] were among the first to set up a study which employed this mixed-methods approach by including both surveys, behavioral metrics, and physiological signals. Mathur, Lane and Kawsar [2016] also combine EEG signals, self-reported perceived engagement scores, and eventually also contextual features automatically derived from smartphones to successfully develop a machine learning model which can detect different levels of engagement.

Our work builds on previous research connecting explicitly expressed user engagement with device interaction behavior. To the best of our knowledge, it is the first to consider mining touch interaction data on tablets in the context of news reading, to take into account the ordering of the content, and to investigate engagement on briefly read articles and on frequently read articles.

3 METHODOLOGY

As an operationalization of user engagement, we use the presence of a user's explicit feedback on an article as the positive outcome of a binary feature. By giving a thumbs up, the user indicates appreciation and relevance. One observation in our dataset is one article visit by one user. The binary dependent feature then says whether the users judged the article positively, or not. How we obtained the explicit judgments in the app is further explained in the section on the experimental set-up. Of course, this is a coarse and short term operationalization, which can only function as a proxy for user engagement.

However, behavioral metrics are the only measurement method which are easily scalable to all users. A simple behavioral metric also allows readers to indicate that they found an article engaging in a matter of seconds, without disrupting the regular reading experience. Alternative methods are neither scalable nor fast. Filling in a survey with even a small number of questions would already interrupt the reading experience too much.

We use logistic regression models to predict whether a user was going to be engaged with an article. We also tried random forests, but this method did not improve the results. We chose logistic regression because it is fast, easy to integrate in internal company tools, and the coefficients of the model offer an intuitive interpretation for feature importance. This makes it easier to communicate the results of the models to a non-technical audience such as editors and journalists. As the dataset is large enough, we could evaluate the predictive performance of the model by doing out-of-time-validation, which is the strongest way to test predictive models [Baesens et al. 2015]. We keep the last 25% of the data separate for testing. As a new newspaper gets released every day of the week except on Sunday, the test set includes only articles which were not seen by any user before. Although in some of the models there is a clear class imbalance, using the SMOTE resampling technique [Chawla et al. 2002] did not significantly improve the results.

Table 1. Time-based features and strictly implicit features used for modeling.

Feature Name	Description
Time-based features	
timeOnArticle	The time in seconds a user spent on the article.
timeOnPage	The total time in seconds spent on the current page where the article is situated.
timeSpentNextPage	The total time in seconds a user spent on the next page.
timeSpentPrevPage	The total time in seconds a user spent on the previous page.
isNextPageRead	Whether the next page is read, taking into account the number of words on that next page.
isPrevPageRead	Whether the next page is read, taking into account the number of words on that previous page.
Strictly implicit features	
articleCompleteness	A % giving the proportion of an article the user has *seen* by scrolling down vertically.
weekend	Whether the session took place during the weekend or not.
nrSwipesArticle	The number of swipes on an article.
nrSwipesPage	The total number of swipes on the current page where the article is situated.
timeToFirstInterPage	The time in seconds it took until the user first interacted with the current page.
timesViewnThisPage	The number of times the user visited this page.
tappedTeaser	Whether the user tapped on a teaser to jump to this article, or not.
nrSessionsNewspaper	The total number of distinct reading sessions on this newspaper.
sessTimeOfDay	Categorical feature saying when the session was taking place; possible values: morning (until 10AM), day (until 5PM), evening.
daysSincePrevSess	The number of days since the user's previous session.
isImageOpened	Whether the user tapped on an image in this article, or not.

The independent features are listed in table 1, 2 and 3. We built five models, each with a different set of independent features: **(1)** only time-based features; **(2)** only strictly implicit features; **(3)** only features based on content ordering; **(4)** a combination of strictly implicit features, features based on content ordering and some additional features which combine implicit feedback and content ordering information (see table 2); **(5)** all features combined (this includes again some additional features which combine implicit feedback and dwell time information, see table 3).

We evaluate the predictive power of the models by calculating the AUC on the test set. We show ROC curves which plot the true positive rate against the false positive rate, and use the DeLong et al. [1988] test to assess whether the AUC of two models is statistically different. We report the sensitivity and specificity for that threshold which yields the largest value for the Kolmogorov-Smirnov statistic between the predictive model and a random prediction model. The Kolmogorov-Smirnov test is a non-parametric test which measures the maximum difference between two cumulative distribution functions [Lilliefors 1967].

Showing the ROC curves, the AUC scores, and the sensitivity and specificity allows us to compare the predictive performance of the five different groups of independent features, thereby answering the first part of the first research question.

Table 2. Content features describing the structure of the newspaper and additional features originating from combining the strictly implicit & content features.

Feature Name	Description
Content features	
pageNumber	The page number of the current page.
isFirstPage	Whether the current article is on the first page of the newspaper, or not.
isLastPage	Whether the current article is on the last page of the newspaper, or not.
category	Each article has an associated category.
catPageNumber	The sequential order of presenting the page, calculated by category.
nrWordsArticle	The number of words of the article.
nrWordsPage	The total number of words on the page.
articleIsAd	Whether the article is an advertisement, or not.
nrImgsOnPage	The number of images on the page.
nrArtsOnPage	The number of articles on the page.
pageHasTeaser	Whether the current page has a teaser to another page, or not.
articleIsTeaser	Whether the article is a short teaser which links to another article, or not.
isTeasedArticle	Whether the current article was teased earlier in the newspaper, or not.
isTeasedPage	Whether the current page was teased earlier in the newspaper, or not.
nextPageNrWords	The number of words on the next page.
prevPageNrWords	The number of words on the previous page.
Implicit & content features combined	
swipeFreqArtWords	Swipe frequency by every 100 words of the article (100×nrSwipesArticle/nrWordsArticle).
swipeFreqPageWords	Swipe frequency on the page by every 100 words on the page (100×nrSwipesPage/nrWordsPage).
swipeDevArticle	The deviation in number of swipes from the average number of swipes on an article for this user.
swipeDevPage	The deviation in number of swipes on the page from the average number of swipes on a page for this user.
swipeDevPageNr	The deviation in number of swipes on the page from the average number of swipes on a page with this page number for this user.

To answer the second part of the first research question, we report the top five most important features of each model by ranking each of the features on the p-value of the Wald statistic, the logistic pseudo partial correlation, the adequacy and the c-statistic (calculated over the whole dataset), and then taking the average of these four rankings to produce a final importance ranking for each feature, as in [Harrell 2015].

Besides showing the five highest ranking features for each model, we also calculate the odds ratio *ceteris paribus* of the top five features of each model. In logistic regression, the odds ratio of an independent feature describes the multiplicative increase in the odds of the dependent feature given a one-unit increase in that independent feature. It is calculated by exponentiating the logistic model coefficients. In this study, the odds ratio of a feature in one of the models describes the change in

Table 3. Additional features originating from combining of time-based and strictly implicit features.

Feature Name	Description
Implicit & time-based features combined	
swipeFreqArticleTime	Swipe frequency by each minute spent on the article (60×nrSwipesArticle/timeOnArticle).
swipeFreqPageTime	Swipe frequency by each minute spent on the page (60×nrSwipesPage/timeOnPage).
pageNrReadProb	The probability (calculated over all users) saying whether the page with this page number will be read or not.
persPageNrReadProb	The probability for this user with which the page with this page number will be read or not.
catReadProb	The probability (calculated over all users) saying whether the category to which the current article belongs, will be read or not.
persCatReadProb	The probability for this user saying whether the category to which the current article belongs, will be read or not.
catPageNrReadProb	The probability (calculated over all users) with which the category to which the current article belongs, will be read or not, taking into account the sequential order of presenting the page, within a category.
persCPgNrReadProb	The probability for this user saying whether the category to which the current article belongs, will be read or not, taking into account the sequential order in which the page is presented, within a category.
devMeanTimeOnPage	The deviation of the average time on a page for this user.
devMeanTOPageNr	The deviation of the average time on a page for this user, taking into account the current page number.

the odds of a user being engaged with an article, given a one-unit increase in the feature value. Odds ratios can be used to compare the magnitude of the effect of different independent features. An odds ratio larger than one is associated with higher odds of a user being engaged, while an odds ratio smaller than one is associated with lower odds of engagement occurring [Baesens et al. 2015]. However, sometimes we have to be careful when interpreting these odds ratios, because we observe some correlation between the independent features, which makes interpreting the odds ratios *ceteris paribus* harder. When we present and discuss the most important features in the results, we always mention when a feature is highly correlated with another feature.

The second research question is also answered by showing the predictive performance of the models which include features based on content ordering as independent features and contrasting their performance with the other models.

To answer the third research question, about briefly read content, we restrict the observations to only keep those user-article pairs on which at most 15 seconds were spent. This threshold was chosen together with the newspaper editors. For this subset of observations, we also report the AUC, the ROC curves and the odds ratios of the most important features. This allows us to contrast the usefulness of the different groups of features when we limit the observations to only briefly read content.

For the final research question, concerning the most frequently read articles, we subset the data on the top 25% of articles which were read by the highest number of users. Again, we also report the AUC, the ROC curves and the odds ratios of the most important features.

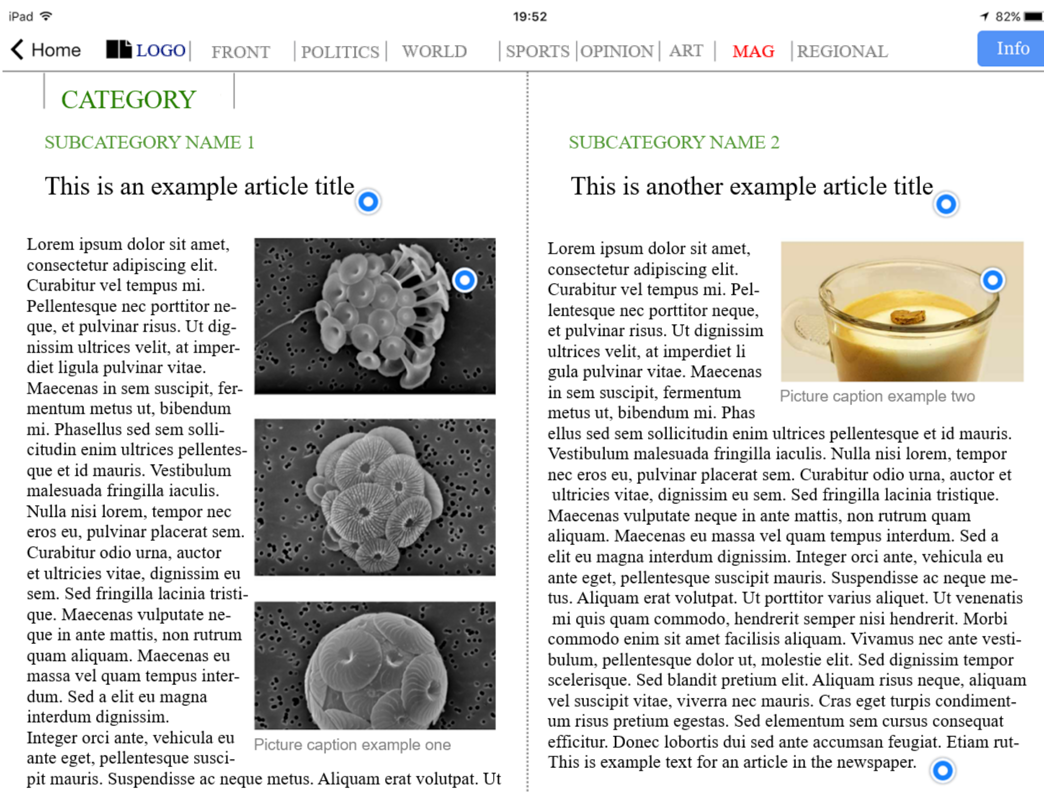

Fig. 1. In-app screenshot (anonymized) of a random page of one of the two digital newspapers. The blue dots are present next to every title, image and at the end of each article.

4 EXPERIMENTAL SET-UP

We did experiments with people who use a tablet app for reading a digital newspaper. We included two separate newspaper brands which are published by the same mother company. Respectively 198 and 209 paying subscribers of each newspaper who used the app regularly participated in the experiment. Each of the two experiments had a duration of one month. The newspapers' brand names can not be mentioned due to confidentiality reasons, but they exist already for several years and have each more than 10,000 active users. Both newspaper brands are among the five most popular in a Western-European country.

A particular aspect of the digital newspaper reading experience is that the content is presented and consumed in a linear way. Users start at the first page of the newspaper and most of them swipe through the pages sequentially until they end their reading session (e.g., they swipe through page 1, 2, 3, ..., 10 and then stop reading). This linear reading aspect of newspapers implies that content ordering might be very important for predicting engagement, and we investigate this effect in the results.

We worked together with Twipe (`www.twipemobile.com`), the company which developed the apps, to modify the apps for the experiment to include an in-app feedback mechanism. An example of an anonymized screenshot of the app can be seen in figure 1. We added small blue dots next to the title, images, and at the end of each article. When a user taps these dots, a pop-up appears

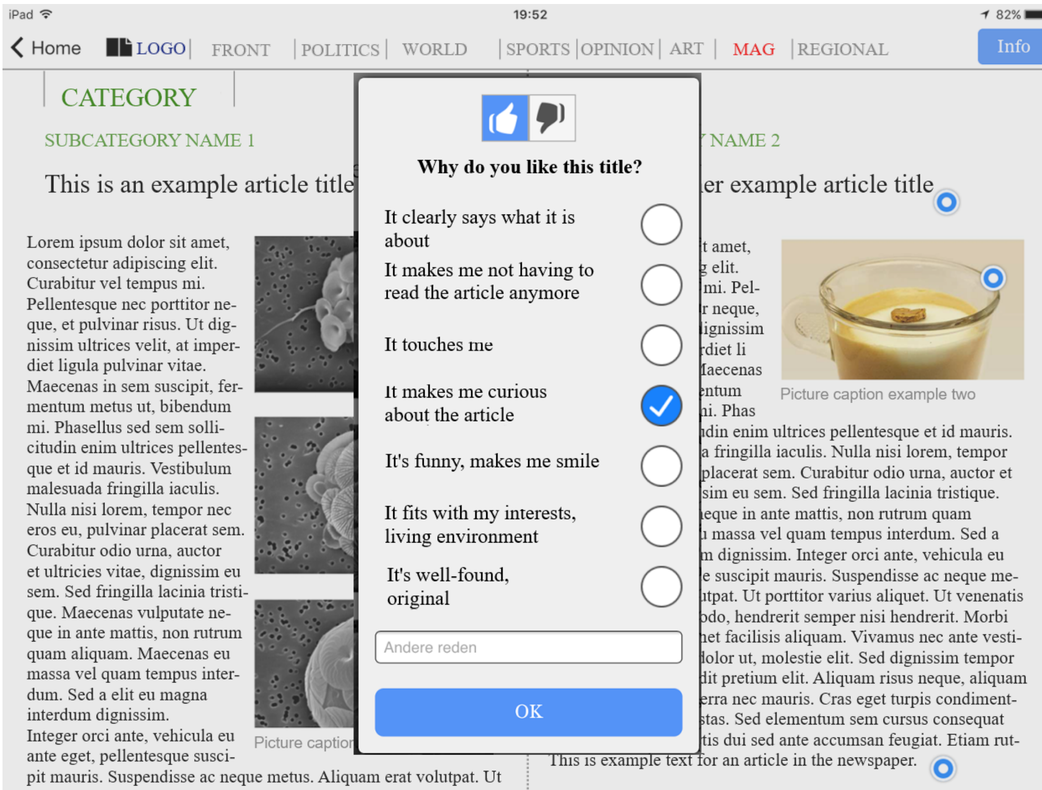

Fig. 2. Example of the pop-up that appears when the blue dot of the title is tapped. The reasons that users gave were not included in this study.

where a thumbs up or thumbs down can be given, and a number of reasons for giving this explicit feedback can be selected, as can be seen in figure 2. The multiple reasons which can be selected when indicating appreciation of an article, are not included in the analysis done for this paper, but could be included in future research. To get more accurate measurements for the time a user spent on an article, the time spent between tapping a blue dot to give feedback and tapping *OK* which signified the end of giving feedback, was subtracted from the total time spent on the article. Users can also give a thumbs-down on an article, but this occurred very infrequently and primarily happened on advertisements in the newspapers. By consequence, we excluded all observations where the article was rated negatively by the users.

We consider a user to be engaged with an article when she gives a thumbs up on any item of that article (can be image, title, or text). We added the blue dots on images, titles, and article texts to give readers plenty of opportunities to give feedback and remove as much barriers as possible. The goal is to minimize user effort. Before the experiment started, we explained to users that all feedback on an article counts equally, irrespective of where they tapped the blue dot. More concretely, if one user gives a thumbs up to an article by tapping the blue dot next to an image of that article, and another user gives a thumbs up to the same article by tapping the blue dot next to the title of that same article, we count both users as having been engaged with that same article.

Table 4. Contingency table of all observations used for the main models, describing whether the article is considered to be engaging or not. Total number of observations is 59875 for newspaper A, and 48659 for newspaper B.

Newspaper A

	not engaging	engaging
nr. observations	42673	17202
% of total	71.27%	28.73%

Newspaper B

	not engaging	engaging
nr. observations	28475	20184
% of total	58.52%	41.48%

Sometimes users give feedback on more than one aspect of the article (e.g. give a thumbs up to both the title and the image of the same article), but we take these interactions together as one observation and consider a user to be engaged when she gives positive feedback about at least one aspect of the article. This was explained to the users before the experiment.

There were also a number of observations for which the calculated time spent on the article was very low or almost zero. As can be seen in figure 1, the situation could occur where multiple articles are visible in the viewport of a user at the same time. The user could be interacting with the article on the left, swiping up and down, and then all of a sudden swipe once on the article on the right. If this happens, the calculated total time spent on the article on the right is very low. This situation also occurs frequently in practice with other apps and websites: there are often links to other content, with corresponding images and multiple sentence captions next to or under the current article. This is an aspect of the experience which we could not control or mitigate.

We emailed a selection of paying subscribers of each of the newspapers with an invitation to fill in a recruitment survey, which assessed eligibility for participation in the experiment. The survey consisted of sociodemographic questions and questions concerning the user's typical reading behavior. Based on the answers to this survey, a sample of candidate participants was drawn which was representative for each newspaper's population of subscribers. All of our candidate participants were acquainted with the app and used it regularly (at least weekly, often more frequently). This set of candidate participants received a personal invitation to download and use the modified version of the app during the next month. Users were explicitly asked to frequently judge those articles they found engaging while they were reading. Eventually, we collected useful data for 407 experiment participants in total, and ended up with over 100,000 unique article visits by users (see table 4).

5 RESULTS AND DISCUSSION

We now show the results for our models. We first show the results for the most general main models. Next, we discuss the results of the models for the briefly read content, and finally we analyze the models for the most frequently read articles.

In each of the next three sections, we report each time for both newspapers a contingency table of the observations used for constructing the models, the ROC curves, the AUC scores, specificity, and sensitivity of each model, and a table with for each model the top five most important features and their associated odds ratios.

In general, the specificity is the proportion of true negatives which are correctly identified by the model. In this study, the specificity is the proportion of articles on which a user spent time

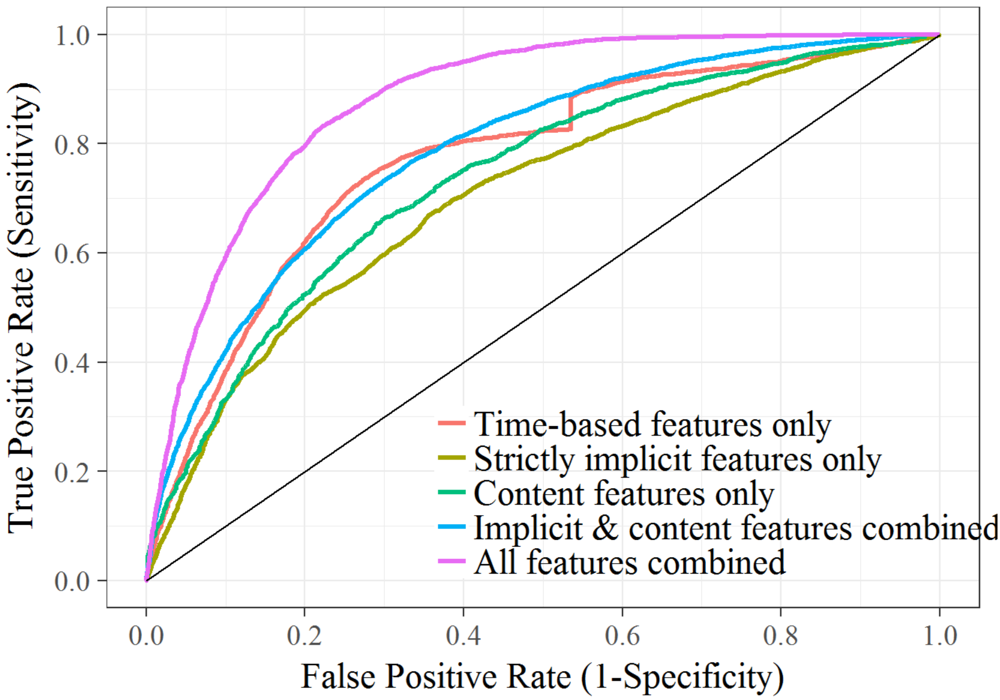

Fig. 3. ROC curves for the main models for newspaper A.

but did not find engaging and which were correctly predicted by the model. The sensitivity is the proportion of observations for which the model correctly predicts that there would be engagement, i.e. that a user would indicate her appreciation of the article. We evaluate the models on their AUC as it is a good summary measure of predictive performance [Baesens et al. 2015].

When discussing the most important features and their associated odds ratios (OR), we only call attention to the particularly interesting or unexpected results. Generally, the differences in feature importance which determine the rankings are minuscule. Note that the OR always need to be interpreted *ceteris paribus*. When the features we discuss are highly correlated with other features and consequentially make the interpretation more difficult, we mention this in the text. The full correlation matrix of all the features for each of the models is available upon request.

5.1 Main models

The **ROC curves** in figure 3 and figure 4 immediately show that using all features yields the best predictive performance on our test set, generating an AUC of 87.96% and 81.63%. We notice a jump in the curve for the model which uses time-based features. This happens because there are a number of observations which have a very small amount of time spent on the article, as explained more thoroughly in the experimental set-up.

The **AUCs** are reported in table 5, together with the specificity and sensitivity of the predictions. At first sight, it seems like the combination of implicit & content features yields an almost equally powerful model as the model that uses only time-based features, with an AUC of 78.64% vs. 77.15% for newspaper A. With newspaper B, the difference is a bit more pronounced, with an AUC of

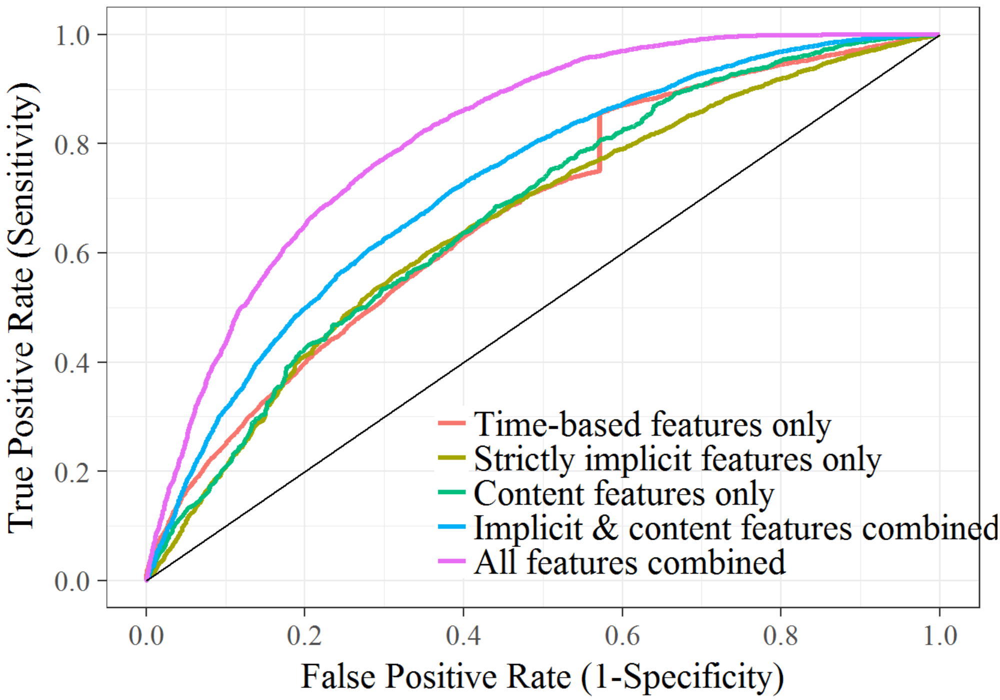

Fig. 4. ROC curves for the main models for newspaper B.

Table 5. AUC, Specificity and Sensitivity of the main models.

Newspaper A

Model Name	AUC	Specificity	Sensitivity
Time features only	77.15%	72.67%	73.51%
Implicit features only	70.4%	64.35%	66.99%
Content features only	73.63%	70.11%	66.32%
Implicit & content features combined	78.64%	70%	73.37%
All features combined	87.96%	77.41%	83.61%

Newspaper B

Model Name	AUC	Specificity	Sensitivity
Time features only	67.2%	42.85%	85.6%
Implicit features only	65.66%	68.86%	55.8%
Content features only	67.24%	56.04%	68.55%
Implicit & content features combined	72.6%	61.31%	71.69%
All features combined	81.63%	71.4%	76.1%

72.6% vs. 67.2%. However, the DeLong et al. [1988] test for comparing two ROC curves shows that the model with time-based features and the model with implicit & content features are for both newspapers significantly different (A: z = -2.7166, p-value = 0.0066; B: z = -8.6598, p-value < 0.0001). This result shows that by using *only* implicit feedback and content ordering features, we can better predict user engagement compared to using *only* time-based features.

The AUC of the model which uses only content ordering features is 73.63% for newspaper A and 67.24% for newspaper B. This model, which uses only static newspaper structure characteristics (see table 2), stands firm between the other models in its performance. It is remarkable that we can achieve this performance without even taking user interactions into account. This confirms that the editors' decisions about where to put which content and accounting for a linear reading pattern is crucial, because it can have a substantial impact on reader engagement occurring. We conclude there is a content ordering effect present with news reading on tablets, similar to the position bias found in web search result ranking [Claypool et al. 2001a].

By combining the implicit & content features, the AUC increases to 78.64% for newspaper A and 72.6% for newspaper B, boosting the performance compared to using these features separately. The implicit & content features seem to complement each other, each giving information about different aspects of the user's experience.

The final model which includes all features shows that combining fine-grained swipe behavior with time spent on content gives the most additional value in terms of predictive power, as shown by the dominating ROC curve and AUC scores. It is the combination of these separate aspects of a user's experience which yields the highest predictive power for both newspapers.

We now examine the results from table 6, where we report the top five **most important features** and their corresponding odds ratios (OR).

For the model which uses only implicit features, we discuss the three features out of the top five which are for both newspapers related to swiping behavior (nrSwipesArticle, articleCompleteness, isImageOpened).

For newspaper A, for each extra swipe on an article, the odds of being engaged with that article increase by 17% (OR nrSwipesArticle: 1.17). This shows that swiping on an article is a positive indication of user interest. This confirms our intuition that more swipes in absolute numbers have a positive impact on the occurrence of engagement.

The context of the user is also important, as each day that has passed by since the user's last reading session (the feature daysSincePrevSess), the odds of judging an article positively increase by 13% for newspaper A and 12% for newspaper B. We suspect there is participant bias in play here, because we explicitly asked users to give feedback on many articles during the experiment.

The feature articleCompleteness is 100% when the user scrolled down to the end of the article. Surprisingly, for both newspapers this feature has an OR of 0.99. This means that for every percentage that a user scrolls further down, the odds of judging that content positively decrease by 1%. A possible explanation is that this feature captures scanning behavior.

The most surprising feature in this model is isImageOpened. When the user taps on any image of the article to open the image and see it more clearly, the odds of being engaged with that article increase by 187% for newspaper A or 56% for newspaper B (OR isImageOpened: 2.87 and 1.56). We can conclude from this that opening on an image is a behavioral action that shows clear user interest.

For the model which uses only content features (second column of table 6), the category feature relates to the importance of the ordering of the articles in the app. Note that for the feature *category*, the odds ratio is given for each possible level of *category* versus the base level *Front Page*. The categories are shown in the tables in the order that they appear in the app.

Table 6. This table shows for each of the main models the top five most important features and their corresponding odds ratios (OR) ceteris paribus in the model.

Newspaper A

	Implicit features only	*OR*	**Content features only**	*OR*	**Implicit & content features combined**	*OR*	**All features combined**	*OR*
1	nrSwipesArticle	1.17	nrArtsOnPage	0.85	nrArtsOnPage	0.87	swipeFreqArticleTime	0.8
2	daysSincePrevSess	1.13	category: Extra	0.19	category: Extra	0.17	nrArtsOnPage	0.83
			category: Regional	0.2	category: Regional	0.2		
			category: Sports	0.35	category: Sports	0.35		
3	articleCompleteness	0.99	catPageNumber	0.95	swipeDevArticle	0.8	category: Extra	0.11
							category: Regional	0.17
							category: Sports	0.25
4	sessTimeOfDay	1.77	nrWordsArticle	1.001	catPageNumber	0.95	swipeDevArticle	0.73
5	isImageOpened	2.87	articleIsTeaser	4.65	nrSwipesArticle	1.37	nrSwipesArticle	1.5

Newspaper B

	Implicit features only	*OR*	**Content features only**	*OR*	**Implicit & content features combined**	*OR*	**All features combined**	*OR*
1	articleCompleteness	0.99	nrWordsArticle	1.001	articleCompleteness	0.99	swipeFreqArticleTime	0.85
2	daysSincePrevSess	1.12	nrArtsOnPage	0.91	nrWordsArticle	1.001	timeOnArticle	1.006
3	isImageOpened	1.56	category: News	0.82	daysSincePrevSess	1.09	devMeanTimeOnPage	1.009
			category: Econ.	0.5				
			category: Sports	0.41				
			category: Culture	0.35				
			category: Regional	0.26				
			category: Opinions	0.46				
4	nrSwipesArticle	1.15	articleIsTeaser	1.92	nrArtsOnPage	0.93	timeOnPage	0.99
5	nrSessions	0.89	isFirstPage	0.43	isImageOpened	1.42	articleCompleteness	0.996

The feature catPageNumber for newspaper A also points in the direction of the effect that when swiping further through the newspaper, it becomes less likely to encounter engaging content.

For newspaper B, the OR of isFirstPage is low (0.43) because in the design of this app, the first page is a front cover which does not have any content which can be judged.

This content ordering effect we observe in the feature rankings is logical because editors put the most important content in the beginning of the newspaper. This insight is similar to that found in web search: the first pieces of content presented to the user (the highest ranked results in search) are the most relevant [Agichtein et al. 2006b].

For each extra article on a page (nrArtsOnPage), the odds of being engaged with one of those articles decreases by 15% for newspaper A or 9% for newspaper B. We believe that when there are more articles visible, the user's attention is more spread out over all these different articles.

Another feature in the top five of most important features which is not related to content ordering is nrWordsArticle. For every extra 100 words in an article, the odds of being engaged increase by 10% (OR nrWordsArticle: 1.001, for both newspapers). It is a stretch to generalize this to saying that longer articles will always be more relevant to users. However, we can state that very short articles have lower odds of being considered engaging.

When considering the combination of both implicit & content features, we see that the top five of these models (third column of table 6) are also present in the top five of the models with both sets of features considered separately, for both newspapers. The effect of content ordering persists with newspaper A, represented by the features category and catPageNumber.

The exception is feature swipeDevArticle for newspaper A, which is a new feature introduced by combining implicit & content features (see table 2). The OR of swipeDevArticle is 0.8, so the odds of liking an article surprisingly decrease when an article is swiped more than average. Luckily, this effect is compensated by nrSwipesArticle: just like in the model with only implicit features, for each extra swipe on the article the odds of liking that article of newspaper A increase by 37%. However, the correlation between the features swipeDevArticle and nrSwipesArt is 73%, so we can not give additional meaning to these features here. The top five features for newspaper B deliver no new insights, they were all already encountered in the previous models, with OR's pointing in the same direction and of similar magnitude.

Finally, in the last model where all features are included, the most important feature is swipeFreqArticleTime and its OR is 0.8 for newspaper A and 0.85 for newspaper B (last column of table 6). This confirms the findings of Guo et al. [2013b] as this feature combines swipe behavior with dwell time information.

This means that when the number of swipes for each minute spent on an article increase by one, the odds of being engaged with that article decrease by 20% or 15%. When swipeFreqArticleTime is larger, there are either more swipes for the same time spent on the content or the same number of swipes for a shorter time spent. In both cases, for larger values of the swipe frequency by each minute spent on the article, the time between swipes decreases, which makes it more likely that the user was scanning the article. When a user scans an article, she is less likely to be engaged by that content. Conversely, if the values for swipeFreqArticleTime are smaller, the time between swipes increases, which means that the user was more actively reading the article. Active reading thus makes a user more likely to be engaged with the content. By combining swipe behavior with dwell time in this feature, we can infer engagement more accurately.

Both newspaper A and B achieve excellent performance for predicting when a user is engaged with the content she is reading. When we consider the different groups of independent features separately, we achieve comparable predictive performance, even if we only use static newspaper content features which do not depend on user interactions at all. The performance increases when we combine the different types of features. The results show that it is not so that there exists one group of features which is always performing better than another. It is rather the combination of different types of features that makes these models perform so well.

5.2 Briefly read articles

When we subset our data to keep only briefly read articles, time-based features are not really useful anymore for determining whether the user is engaged with the content or not. The goal here is to assess the usefulness of alternative user interactions for predicting engagement, despite the fact that she spent only less than 15 seconds with it. The proportion of engaging articles changes, as can be seen in table 7. Although the class imbalance becomes stronger, this had no significant impact on the predictive results. We repeated the modeling exercise by using the SMOTE resampling technique [Chawla et al. 2002] to account for class imbalance, but the results did not differ much.

The **ROC curves** are visually shown in figure 5 and figure 6. One thing that immediately stands out is the defective performance of the model which uses time-based features, especially for newspaper B. The predictions are almost as bad as a random model. Fortunately, this confirms what we expected to see. The jump in the curve is very pronounced and can again be explained by the fact that there are a number of articles for which very little time spent on the article was registered (as more thoroughly explained earlier in the section on the experimental set-up).

Table 8 shows the **AUCs**. The sensitivity of this model is much lower compared to the other models for the briefly read articles. This means that the time-based features are not useful for

Table 7. Contingency table of the observations used in the models for the briefly read articles, describing whether the article is considered to be engaging or not. Total number of observations which are read less than 15 seconds is 35451 for newspaper A, and 18904 for newspaper B.

Newspaper A

	not engaging	engaging
nr. observations	30615	4836
% of total	86.36%	13.64%

Newspaper B

	not engaging	engaging
nr. observations	14051	4853
% of total	74.33%	25.67%

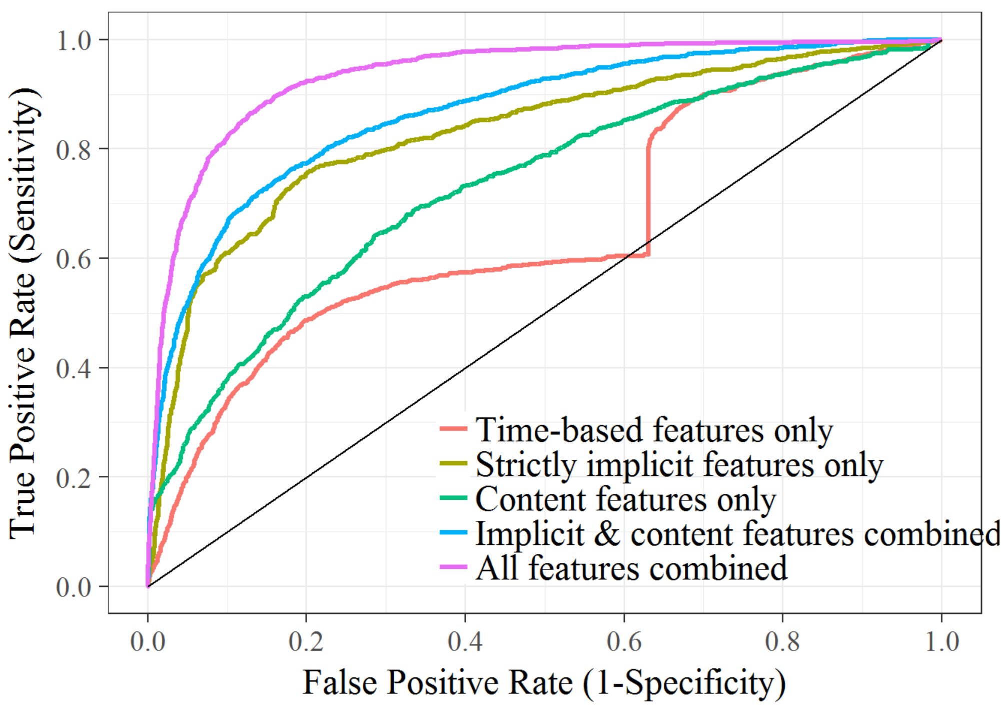

Fig. 5. ROC curves for the models for the briefly read articles for newspaper A.

distinguishing the engaging articles, as we expected. It is an interesting insight that including swipe interactions or content characteristics is necessary for identifying engaging content when we consider only brief interactions.

Here, the model with only implicit features performs really well with an AUC of 82.94% for newspaper A and 83.51% for newspaper B. The difference in model performance between using only implicit features and using only content features is larger compared to the models analyzed in the previous section which included all observations, and the difference in model performance

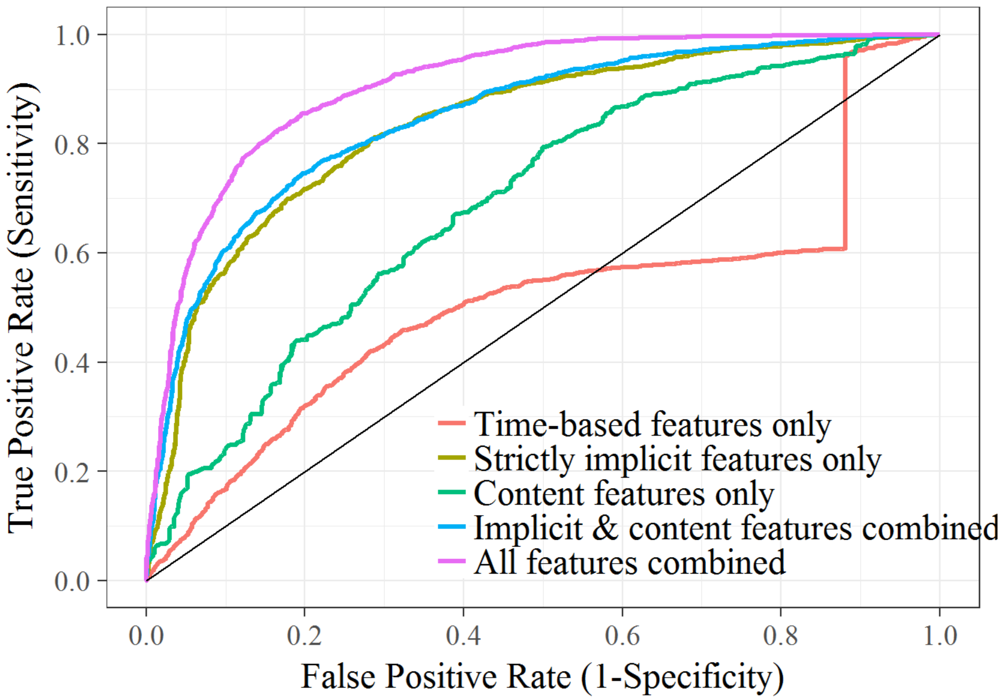

Fig. 6. ROC curves for the models for the briefly read articles for newspaper B.

Table 8. AUC, Specificity and Sensitivity of the models for the briefly read articles.

Newspaper A

Model Name	AUC	Specificity	Sensitivity
Time features only	65.1%	80.3%	48.57%
Implicit features only	82.94%	79.77%	75.91%
Content features only	73.06%	71.49%	64.18%
Implicit & content features combined	86.73%	82.24%	76.07%
All features combined	93.57%	85.19%	88.51%

Newspaper B

Model Name	AUC	Specificity	Sensitivity
Time features only	51.05%	71.83%	41.94%
Implicit features only	83.51%	71.33%	81.15%
Content features only	69.16%	50.11%	79.42%
Implicit & content features combined	84.83%	80.41%	74.51%
All features combined	90.67%	83.89%	82.1%

between using only implicit features and using implicit & content features combined is smaller compared to the models from the previous section which included all observations. The ROC curves of the models which use only implicit features and the models which use both implicit & content features lie closest to each other. However, the DeLong et al. [1988] test shows that these ROC curves are significantly different from each other (A: z = 18.0007, p-value < 0.0001; B: z = 22.665, p-value < 0.0001).

The last two models which combine different types of features both perform very well. If we combine all the features described in table 1, 2, and 3, we achieve an excellent AUC of 93.57% for newspaper A and 90.67% for Newspaper B.

Based on the results of the ROC curves and the AUC scores, we can conclude that even if content was only briefly interacted with, we are able to identify engaging content by using implicit features.

Table 9 shows the top five **most important features** for each model for the briefly read articles. For the model which uses only implicit features, the top five features are exactly the same for newspaper A and B. Furthermore, for those main models from the previous section which also use only implicit features, the features articleCompleteness, nrSwipesArticle and daysSincePrevSess also appear as most important features with odds ratios pointing in a similar direction. For example, here too, articleCompleteness has an OR smaller than one, and sessTimeOfDay has a high OR. Note that there is a high correlation between the features nrSwipesArticle and nrSwipesPage, for both newspapers. We conclude that the most important features are similar to those of the corresponding model from the previous section. However, the performance of this model for briefly read articles which uses only implicit features is relatively higher compared to the best performing model which included all features. So exactly the same implicit features yield better predictive performance if we consider only briefly read articles. Those implicit features are able to compensate for the decrease in predictive performance due to the loss of usefulness of the time-based features. This model which uses only implicit features can already make accurate predictions from interactions that happen in only a short period of time.

The models which use only features based on content ordering, have four out of five top features in common and with similar odds ratios as the models from the previous section which did not restrict the reading times. Comparing newspaper A and B in the second column of table 9 shows that three out of their top five features are identical. The effect when an article is a teaser article linking to another article, is extreme for newspaper A (OR articleIsTeaser: 11.09). The model performance is almost the same relative to the corresponding main model from the previous section, so the content ordering effect is not different for briefly read articles compared to all articles. We do not observe an effect like in the previous paragraph when using only implicit features, where the same features became more useful when considering only briefly read articles instead of all articles.

When we look at the most important features for the model that combines both implicit & content features (third column of table 9), again all the features in the top five are also present with the models where both sets of features are only considered separately. The exception is swipeFreqPageWords, which is a new feature resulting from combining implicit & content features (see table 2). We have to be careful in interpreting the odds ratios here, as for example nrSwipesPage and nrSwipesArticle have a correlation of 41%. The feature articleIsTeaser has again a high OR for newspaper A, but swipeFreqPageWords has also an OR of 6.43 for newspaper B. This means that for each additional swipe for every 100 words on a page, the odds of finding the current content engaging increases by 543%. Of course, this should be nuanced when the number of words on a page is low. In this case, very little swipes are needed to achieve a high value for this feature. The majority of the top five features for both newspaper A and B relate to swiping behavior, indicating

Table 9. This table shows for each model for the briefly read articles the top five most important features and their corresponding odds ratios (OR) ceteris paribus in the model.

Newspaper A

	Implicit features only	*OR*	**Content features only**		*OR*	**Implicit & content features combined**	*OR*	**All features combined**		*OR*
1	articleCompleteness	0.97	articleIsTeaser		11.09	articleCompleteness	0.97	swipeFreqArticleTime		0.89
2	nrSwipesPage	0.96	catPageNumber		0.95	nrSwipesPage	0.82	articleCompleteness		0.98
3	sessTimeOfDay	1.92		Extra	0.23	sessTimeOfDay	1.7	swipeFreqPageTime		0.93
			category	Regional	0.17					
				Sports	0.46					
4	daysSincePrevSess	1.3	nrArtsOnPage		0.84	articleIsTeaser	6.12	articleIsTeaser		9.42
5	nrSwipesArticle	1.31	pageHasTeaser		4.58	catPageNumber	0.95		Extra	0.1
								category	Regional	0.13
									Sports	0.22

Newspaper B

	Implicit features only	*OR*	**Content features only**		*OR*	**Implicit & content features combined**	*OR*	**All features combined**	*OR*
1	articleCompleteness	0.96	nrArtsOnPage		0.88	articleCompleteness	0.97	swipeFreqArticleTime	0.82
2	nrSwipesArticle	1.25		News	0.66	nrSwipesPage	0.76	articleCompleteness	0.98
				Econ.	0.32				
				Sports	0.32				
			category	Culture	0.16				
				Regional	0.09				
				Opinions	0.24				
3	daysSincePrevSess	1.27	articleIsTeaser		4.29	daysSincePrevSess	1.23	swipeFreqPageTime	0.91
4	nrSwipesPage	0.92	isFirstPage		0.22	swipeFreqPageWords	6.43	timeOnArticle	1.1
5	sessTimeOfDay	1.32	nrImgsOnPage		0.89	nrWordsPage	0.99	daysSincePrevSess	1.29

that for briefly read articles, implicit features are of more value for predicting engagement than content ordering features.

The final model combines all features and has a high AUC of 93.57% for newspaper A and 90.67% for newspaper B. The top three features for the model with all features combined are the same for newspaper A and B. If we look at the features that are most important in contributing to that predictive power (last column of table 9), we find again that combining the time aspect with the swiping behavior yields the two most important features, swipeFreqArticleTime and swipeFreqPageTime. These features describe the swipe frequency by time spent on the article and page, and tell us more about whether a user is scanning or actively reading (as explained earlier in the previous section).

Also notice that including time-based features in addition to implicit & content features still boosts the predictive performance a bit higher. This is surprising because we need to take into account that there is a lot less variation in the time-based features now. It is probably not the addition of the simple time-based features which causes the performance boost, but the inclusion of exceptional features such as swipeFreqArticleTime, which succeed in combining swipe interactions with dwell time in one feature.

Finally, the set of most important features for the models which use only implicit features and the models which use all available features does not vary a lot between the main models from the previous section and the models for briefly read articles. Fine-grained swipe interactions as implicit features are of great value for predicting engagement when users only briefly interact with some content.

5.3 Frequently read articles

It is interesting to look into the subset of top 25% most frequently read articles for several reasons. Newspaper editors spend a lot of time with the best performing content, analyzing why it works well and trying to replicate it with other stories. There might also be specific user interactions which are indicative of content which appeals to many subscribers of the newspaper. These interactions would help to identify engaging articles which function as a greatest common divisor across the whole reader base of the newspaper.

Table 10 shows the subset of observations of people spending time on and interacting with the top 25% most read articles. Although we retain only 25% of all unique articles present in the full dataset, these observations represent 71% of all observations for newspaper A, and 56.7% for newspaper B. There is again some class imbalance but adapting the modeling approach by employing a resampling scheme did not significantly improve the results.

The **ROC curves** in figure 7 and figure 8 show visually the predictive performance of the models for these most frequently read articles, and table 11 reports its performance in terms of **AUCs**. Here, the models with only time-based features outperform the models with implicit & content features combined, for both newspapers. For newspaper A, the model with time-based features achieves an AUC of 77.49% compared to 74.79% for the model which uses implicit & content features combined, and with newspaper B the difference is 68.43% against 65.95%. This means that for the most frequently read articles, specific fine-grained swipe interactions do not increase predictive power additionally to time-based features. This is in contrast to the results from the two previous sections, where the combination of implicit & content features in both sections outperformed the models which used only time-based features.

The model based on content features alone does not perform well with an AUC of 68.61% for newspaper A and 59.26% for newspaper B. However, for this model, this is to be expected. We selected the observations in this section based on how frequently the article was read, and those articles which are most frequently read share the same characteristics. The articles in this subset of the data are on general topics which many people find interesting, are typically very newsworthy, and are situated on the first pages of the newspaper. There is no content ordering effect with the most frequently read articles.

The best model, which uses all features, performs about 9% better in AUC compared to the second-best model, which uses only time-based features, for both newspapers. This shows that

Table 10. Contingency table of the observations used in the models for the most frequently read articles, describing whether the article is considered to be engaging or not. The top 25% most frequently read articles account for 42685 observations for newspaper A and 27605 observations for newspaper B.

Newspaper A

	not engaging	engaging
nr. observations	28902	13783
% of total	67.71%	32.29%

Newspaper B

	not engaging	engaging
nr. observations	13869	13736
% of total	50.24%	49.76%

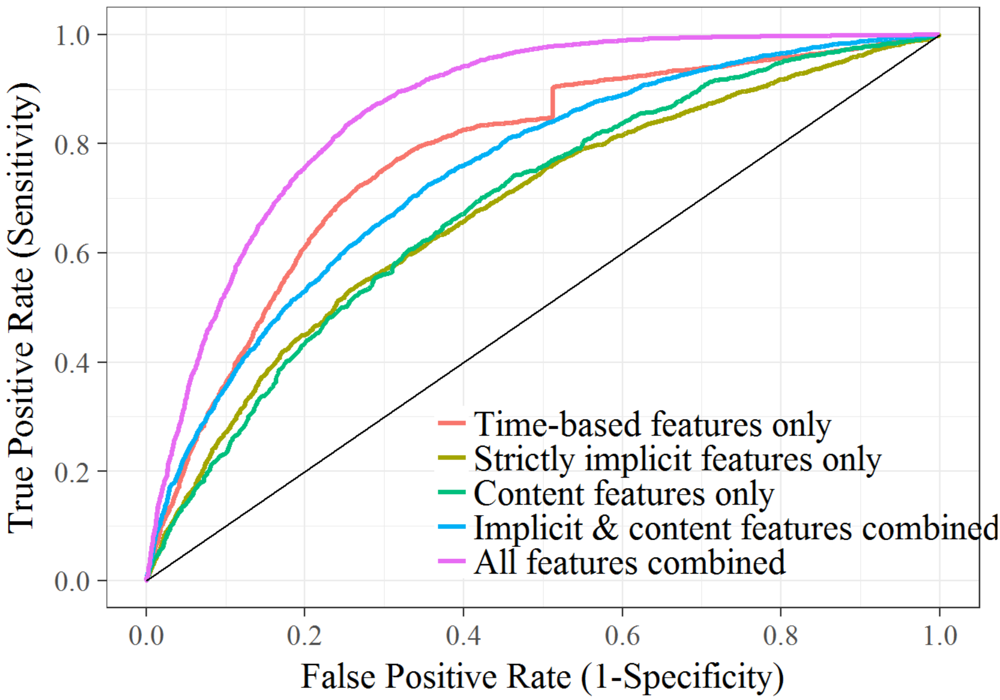

Fig. 7. ROC curves for the models for the most frequently read articles for newspaper A.

Table 11. AUC, Specificity and Sensitivity of the models for the most frequently read articles.

Newspaper A

Model Name	AUC	Specificity	Sensitivity
Time features only	77.49%	68.47%	77.11%
Implicit features only	67.91%	73.27%	54.24%
Content features only	68.61%	53.81%	74.07%
Implicit & content features combined	74.79%	64.51%	72.46%
All features combined	86.36%	72.33%	86.1%

Newspaper B

Model Name	AUC	Specificity	Sensitivity
Time features only	68.43%	54.15%	78.93%
Implicit features only	62.74%	65.84%	55.16%
Content features only	59.26%	58.07%	59.41%
Implicit & content features combined	65.95%	64.73%	59.91%
All features combined	77.33%	76.33%	73.38%

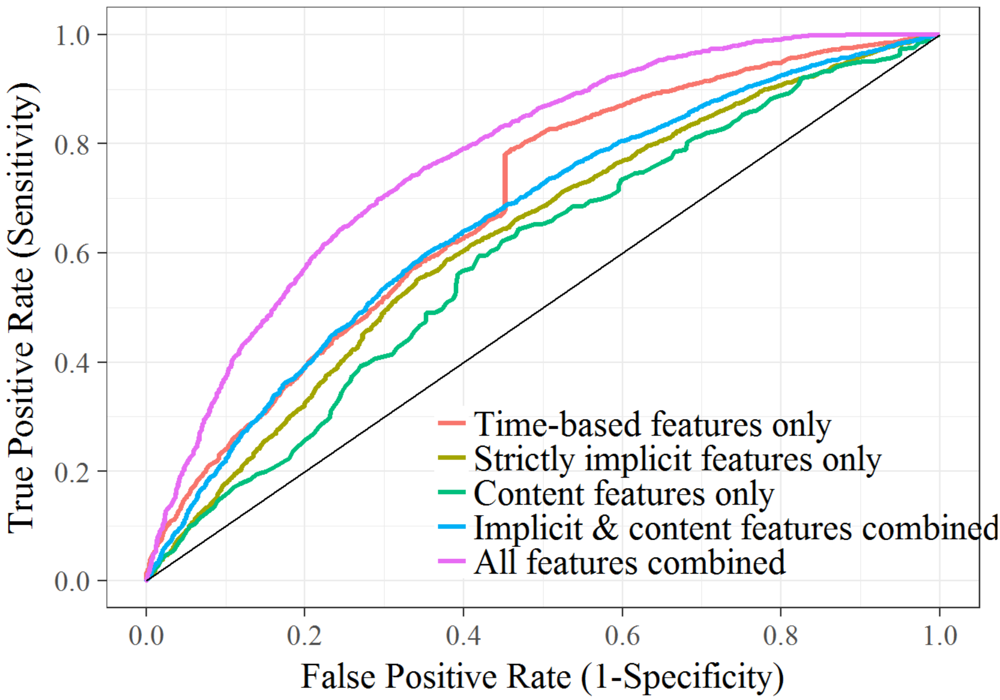

Fig. 8. ROC curves for the models for the most frequently read articles for newspaper B.

both types of features are useful in predicting engagement, and that these types of features should be used complementary. We found the same result in the two previous sections.

We now highlight some results from table 12, which summarizes the **most important features** with their odds ratios. The most important features of the model which uses only implicit features are similar to those found in the previous sections. However, the most important features of this model are not as interesting to discuss compared to the previous sections, because here, the models which use only implicit features perform weakly. The most important features of the model which uses only content-based features (second column of table 12) are almost the same as for the well-performing models with briefly read content which also used only content-based features, as discussed in the previous section. These models have weak predictive performance when we consider only the most frequently read articles. Just like subsetting on only briefly read articles in the previous section eliminated variation in the time-based features, it seems like there is now also less variation in the content-based features because we only consider the most frequently accessed articles.

The complementarity of time-based features and implicit features also shows itself here in the most important feature of the best model which uses all features: swipeFreqArticleTime. This feature integrates both a time-aspect and a swipe interaction aspect of the user experience and comes back as a key feature in each of the three settings we discussed.

Table 12. This table shows for each model for the most frequently read articles the top five most important features and their corresponding odds ratios (OR) ceteris paribus in the model.

	Implicit features only	*OR*	Content features only		*OR*	Implicit & content features combined	*OR*	All features combined	*OR*
	Newspaper A								
1	nrSwipesArticle	1.14	nrArtsOnPage		0.88	nrArtsOnPage	0.19	swipeFreqArticleTime	0.8
2	isImageOpened	3.15	catPageNumber		0.95	swipeDevArticle	0.78	nrArtsOnPage	0.83
3	daysSincePrevSess	1.10	nrWordsArticle		1.001	catPageNumber	0.95	swipeDevArticle	0.74
4	sessTimeOfDay	1.69	category	Extra	0.14	nrSwipesArticle	1.41	nrSwipesArticle	1.46
				Regional	0.43				
				Sports	0.38				
5	nrSwipesPage	0.97	isTeasedArticle		1.67	swipeDevPage	1.33	catPageNumber	0.94
	Newspaper B								
1	daysSincePrevSess	1.10	nrArtsOnPage		0.89	nrArtsOnPage	0.91	swipeFreqArticleTime	0.86
2	isImageOpened	1.53	nrWordsArticle		1.001	nrWordsArticle	1.001	timeOnArticle	1
3	nrSessions	0.9	category	News	0.78	daysSincePrevSess	1.07	devMeanTimeOnPage	1.007
				Econ.	0.65				
				Sports	0.37				
				Culture	0.25				
				Opinions	0.43				
4	nrSwipesArticle	1.14	nrWordsPage		0.99	isImageOpened	1.52	timeOnPage	0.99
5	articleCompleteness	0.99	isTeasedArticle		2.1	articleCompleteness	0.99	nrArtsOnPage	0.9

6 CONCLUSION

This paper proposed a solution to enable better large scale measurement and prediction of user engagement in the context of digital newspaper reading on tablets. We used the behavioral metric of positive in-app feedback on news articles as a proxy for engagement.

Although on a small scale users can be asked to give explicit feedback about content and that explicit feedback is an accurate measure for user engagement, it requires high cognitive effort and is not scalable. Traditionally, dwell time is used as a proxy for this explicit feedback which is usable on large scale.

We showed that by incorporating implicit feedback in the form of swiping interactions and taking into account the order of presenting the content we can in general achieve better user engagement predictions. We did an out-of-time validation of each of the predictive logistic regression models, for each model varying the set of independent features and assessing the performance on the AUC, specificity and sensitivity.

To evaluate the most important predictive features, we calculated the odds ratios after ranking the features of each model. The best features take into account the complementarity of time-based and implicit features. Features that can combine both are the most important features, such as swipeFreqArticleTime, which is the swipe frequency by each minute that a user spends on an article.

Finally, we also zoomed in on briefly read articles and the 25% most frequently read articles. We redid the analysis for the subset of observations of articles on which users spent maximally 15 seconds, and also redid the analysis by only taking into account the top 25% most frequently read articles. The briefly read articles could still be engaging for users, but time-based features were not useful anymore. Our results showed that we can predict user engagement for briefly read articles more accurately, specifically because we use the information present in the swipe interactions. If the swipe interaction information would not be available, the user engagement predictions would be worse. In contrast, for the 25% most read articles, the model which uses only time-based features performs better than models which use a combination of implicit features and content features.

In summary, we have presented the case for better large-scale predictions of user engagement by exploiting implicit feedback. In general, for the three settings we evaluated, leveraging features which succeed in combining time-based aspects and swipe interactions as implicit feedback into a single feature, always improved the performance of the predictions for user engagement.

REFERENCES

E. Agichtein, E. Brill, and S. Dumais. 2006a. Improving web search ranking by incorporating user behavior information. *SIGIR* (2006), 19–26. https://doi.org/10.1145/1148170.1148177

E. Agichtein, E. Brill, S. Dumais, and R. Ragno. 2006b. Learning User Interaction Models for Predicting Web Search Result Preferences. *SIGIR* (2006), 3–10. https://doi.org/10.1145/1148170.1148175

Ioannis Arapakis, M. Lalmas, B. Barla Cambazoglu, Mari-Carmen Marcos, and Joemon M. Jose. 2014b. User engagement in online news: under the scope of sentiment, interest, affect, and gaze. *Journal of the Association for Information Science and Technology (JASIST)* 65, 10 (10 2014), 1988–2005. https://doi.org/10.1002/asi.23096

Ioannis Arapakis, M. Lalmas, and George Valkanas. 2014a. Understanding within-content engagement through pattern analysis of mouse gestures. *CIKM* (2014), 1439–1448. https://doi.org/10.1145/2661829.2661909

Bart Baesens, Veronique Van Vlasselaer, and Wouter Verbeke. 2015. *Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection.* Wiley. 400 pages.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16 (2002), 321–357. https://doi.org/10.1613/jair.953

M. Claypool, D. Brown, P. Le, and M. Waseda. 2001a. Inferring user interest. *IEEE Internet Computing* 5, 6 (2001), 32–39. https://doi.org/10.1109/4236.968829

M. Claypool, P. Le, M. Wased, and D. Brown. 2001b. Implicit interest indicators. *IUI* (2001), 33–40. https://doi.org/10.1145/359784.359836

Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* (1988), 837–845.

A. Drutsa and P. Serdyukov. 2015. Future user engagement prediction and its application to improve the sensitivity of online experiments. *WWW* (2015), 256–266. https://doi.org/10.1145/2736277.2741116

Georges Dupret and M. Lalmas. 2013. Absence Time and User Engagement: Evaluating Ranking Functions. *WSDM* (2013), 173–182.

Steve Fox, Kuldeep Karnawat, Mark Mydland, Susan Dumais, and Thomas White. 2005. Evaluating Implicit Measures to Improve Web Search. *ACM Transactions on Information Systems* 23, 2 (April 2005), 147–168. https://doi.org/10.1145/1059981.1059982

Q. Guo and E. Agichtein. 2008. Exploring Mouse Movements for Inferring Query Intent. *SIGIR* (2008), 707–708. https://doi.org/10.1145/1390334.1390462

Q. Guo and E. Agichtein. 2012. Beyond dwell time: estimating document relevance from cursor movements and other post-click searcher behavior. *WWW* (2012), 569–578. https://doi.org/10.1145/2187836.2187914

Q. Guo, H. Jin, D. Lagun, S. Yuan, and E. Agichtein. 2013a. Mining Touch Interaction Data on Mobile Devices to Predict Web Search Result Relevance. *SIGIR* (2013), 153–162. https://doi.org/10.1145/2484028.2484100

Q. Guo, H. Jin, D. Lagun, S. Yuan, and E. Agichtein. 2013b. Towards estimating web search result relevance from touch interactions on mobile devices. *CHI Extended Abstracts* (2013), 1821–1826. https://doi.org/10.1145/2468356.2468683

Frank E. Harrell. 2015. *Regression Modeling Strategies.* Springer International Publishing. https://doi.org/10.1007/978-1-4757-3462-1

Jeff Huang and Abdigani Diriye. 2012. Web User Interaction Mining from Touch-Enabled Mobile Devices. *HCIR* (2012).

Jeff Huang, Ryen W. White, and Susan Dumais. 2011. No Clicks, No Problem: Using Cursor Movements to Understand and Improve Search. *CHI* (2011), 1225–1234. https://doi.org/10.1145/1978942.1979125

J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl. 1997. GroupLens: applying collaborative filtering to Usenet news. *Commun. ACM* 40, 3 (1997), 77–87.

D. Lagun, M. Ageev, Q. Guo, and E. Agichtein. 2014. Discovering common motifs in cursor movement data for improving web search. *WSDM* (2014), 183–192. https://doi.org/10.1145/2556195.2556265

D. Lagun and M. Lalmas. 2016. Understanding and Measuring User Engagement and Attention in Online News Reading. *WSDM* (2016), 113–122. https://doi.org/10.1145/2835776.2835833

M. Lalmas, H. L. O'Brien, and E. Yom-Tov. 2014. Measuring user engagement. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 6, 4 (2014), 1–132.

H. J. Lee and Sung Joo Park. 2007. MONERS: A news recommender for the mobile web. *ESWA* 32, 1 (2007), 143–150. https://doi.org/10.1016/j.eswa.2005.11.010

Janette Lehmann, M. Lalmas, Elad Yom-tov, and Georges Dupret. 2012. Models of User Engagement. *UMAP* (2012), 164–175.

H. W. Lilliefors. 1967. On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. *J. Amer. Statist. Assoc.* 62, 318 (1967), 399–402.

Jiahui Liu, Peter Dolan, and Er Pedersen. 2010. Personalized news recommendation based on click behavior. *IUI* (2010), 31–40. https://doi.org/10.1145/1719970.1719976

Yiqun Liu, Xiaohui Xie, Chao Wang, Jian-Yun Nie, Min Zhang, and Shaoping Ma. 2016. Time-Aware Click Model. *ACM Trans. Inf. Syst.* 35, 3, Article 16 (Dec. 2016), 24 pages. https://doi.org/10.1145/2988230

Shiyang Lu, Tao Mei, Jingdong Wang, Jian Zhang, Zhiyong Wang, and Shipeng Li. 2014. Browse-to-Search: Interactive Exploratory Search with Visual Entities. *ACM Trans. Inf. Syst.* 32, 4, Article 18 (Oct. 2014), 27 pages. https://doi.org/10.1145/2630420

Akhil Mathur, Nicholas D. Lane, and Fahim Kawsar. 2016. Engagement-aware computing: Modelling User Engagement from Mobile Contexts. *UbiComp* (2016), 622–633. https://doi.org/10.1145/2971648.2971760

Lori McCay-Peet, M. Lalmas, and Vidhya Navalpakkam. 2012. On Saliency, Affect and Focused Attention. *CHI* (2012), 541–551. https://doi.org/10.1145/2207676.2207751

Masahiro Morita and Yoichi Shinoda. 1994. Information Filtering Based on User Behavior Analysis and Best Match Text Retrieval. *SIGIR* (1994), 272–281. http://dl.acm.org/citation.cfm?id=188490.188583

Vidhya Navalpakkam and Elizabeth Churchill. 2012. Mouse tracking: measuring and predicting users' experience of web-based content. *CHI* (2012), 2963–2972. https://doi.org/10.1145/2207676.2208705

H. L. O'Brien, R. Absar, and H. Halbert. 2013. Toward a Model of Mobile User Engagement. *HCIR* (3 2013). http://circle.ubc.ca/handle/2429/45340

H. L. O'Brien and M. Lebow. 2013. Mixed-Methods Approach to Measuring User Experience in Online News Interactions. *Journal of the American Society for Information Science and Technology (JASIST)* 64, 8 (2013), 1543–1556. https://doi.org/10.1002/asi.22871

H. L. O'Brien and E. G. Toms. 2008. What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology (JASIST)* 59, 6 (2008), 938–955. https://doi.org/10.1002/asi.20801

H. L. O'Brien and E. G. Toms. 2010. The Development and Evaluation of a Survey to Measure User Engagement. *Journal of the American Society for Information Science and Technology (JASIST)* 61, 1 (2010), 50–69. https://doi.org/10.1002/asi.21229

Bracha Shapira, Meirav Taieb-Maimon, and Anny Moskowitz. 2006. Study of the Usefulness of Known and New Implicit Indicators and Their Optimal Combination for Accurate Inference of Users Interests. *SAC* (2006), 1118–1119. https://doi.org/10.1145/1141277.1141542

Y. Song, Hao Ma, Hongning Wang, and Kuansan Wang. 2013a. Exploring and Exploiting User Search Behavior on Mobile and Tablet Devices to Improve Search Relevance. *WWW* (2013), 1201–1212. https://doi.org/10.1145/2488388.2488493

Y. Song, X. Shi, and X. Fu. 2013b. Evaluating and Predicting User Engagement Change with Degraded Search Relevance. *WWW* (2013), 1213–1223.

Maximilian Speicher and Martin Gaedke. 2013. TellMyRelevance! Predicting the Relevance of Web Search Results from Cursor Interactions. *CIKM* (2013), 1281–1290. https://doi.org/10.1145/2505515.2505703

Ryen W. White, Ian Ruthven, Joemon M. Jose, and C. J. Van Rijsbergen. 2005. Evaluating Implicit Feedback Models Using Searcher Simulations. *ACM Trans. Inf. Syst.* 23, 3 (July 2005), 325–361. https://doi.org/10.1145/1080343.1080347

Xing Yi, Liangjie Hong, Erheng Zhong, Nathan Nan, and Liu Suju. 2014. Beyond Clicks: Dwell Time for Personalization. *RecSys* (2014), 113–120. https://doi.org/10.1145/2505515.2505682