Uncovering Personal and Context-Dependent Display Preferences in Mobile Newsreader App

Emir Hasanbegović
eh7718@student.uni-lj.si
Faculty of Computer and Information Science,
University of Ljubljana
Slovenia

Veljko Pejović
veljko.pejovic@fri.uni-lj.si
Faculty of Computer and Information Science,
University of Ljubljana
Slovenia

ABSTRACT

The smartphone has revolutionised the way we receive news, enabling on-demand, personalised content to be viewed in a range of different situations. Yet, while the content of the news is often adapted to the user's preferences and the current environment (e.g. location), the actual interface of a mobile newsreader app often remains the same across users and contexts of use. In this work we first collect and examine real-world mobile news reading data to uncover the way contextual factors affect the perception of different aspects of the newsreader app interface, and then develop a method for modelling personalised context-dependent viewing preferences. Through a four-week long user study we demonstrate that our reinforcement and active learning-based personalisation approach leads to 26% higher user acceptance as compared to a generic context-aware mobile newsreader interface adaptation model.

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile computing design and evaluation methods**; *User models.*

KEYWORDS

mobile news, context-aware computing, mixed-effects modelling, reinforcement learning

ACM Reference Format:
Emir Hasanbegović and Veljko Pejović. 2021. Uncovering Personal and Context-Dependent Display Preferences in Mobile Newsreader App. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '21), June 21–25, 2021, Utrecht, Netherlands.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3450613.3456808

1 INTRODUCTION

With more than half of the world's population owning a mobile device, and with the penetration reaching into the most remote of the world's regions, where conventional landline and cable connectivity has never had, the smartphone represents the most ubiquitous source of information on the planet. Besides the conventional communication capabilities, such as voice and text messaging, the smartphone brings instant access to news articles via the mobile Web and dedicated news reading applications.

Mobile news reading has gained tremendous traction in the last decade, contrasting a sharp drop in readership experienced by traditional media [20]. The convenience of obtaining information through a multipurpose device that is carried around and available at all times, the fact that the news consumption is not constrained to "mediatope" locations, as it is the case with e.g. television, and the potential for the personalisation of the received content, render mobile news not only an attractive alternative, but also a platform introducing affordances not offered by traditional means of disseminating news.

Content consumption on-the-go and in diverse environments is probably the most striking change as we move towards mobile news apps. Such means of use is shown to lead to a change in behaviour, where news are "snacked" [15]. Consequently, mobile news apps tend to provide short bundles of engaging textual or video content, rather than full-fledged articles and news reports that are typical in traditional media. Furthermore, providers are often harnessing the knowledge of the user's context, such as the location sensed by mobile phone's sensors, to further tailor the news content [1]. On the other hand, when it comes to the actual *rendering* of the news content on a mobile device, surprisingly little is being done to adapt the presentation to the context in which the application is used as well as to the preferences of the user using the app.

In this paper we aim to fill this gap and to uncover, potentially individually-coloured, relationship between the context of usage and the preferred manner of displaying news information on a mobile device. We design and develop a mobile newsreader app that allows a number of different display parameters, such as the layout, the font size, theme colour, and others, to be adapted, and through a real-world study that involves mobile context sensing and experience sampling method (ESM) collect data pertaining to the context and the display preferences when reading news. Using hierarchical mixed-effect modelling we identify the (interplay of) contextual factors that affect the perception of different means of displaying news.

Our preliminary study finds that certain aspects of the layout, such as the presence of pictures in news articles, remain preferred across the contexts. We further identify physical activity and ambient brightness as main factors guiding successful contextual adaptation of the interface. Acknowledging individualism in the viewing preferences, we then develop a novel method for iterative personalisation of the displayed layout. Our method is based on a combination of active and reinforcement learning, thus minimises the amount of querying needed in order to deduce individual user's

UMAP '21, June 21–25, 2021, Utrecht, Netherlands

ACM ISBN 978-1-4503-8366-0/21/06…$15.00
https://doi.org/10.1145/3450613.3456808

preferences. In a separate user study the interfaces adapted through our personalisation method receive 26% higher user acceptance level than interfaces adapted in a user-oblivious manner.

In summary, the contributions of our work include:

- A hierarchical model explaining the impact of contextual descriptors on the general newsreader app display preferences;
- An active-reinforcement learning approach for the personalisation of the newsreader app that deduces individual user's preferences while keeping the amount of user querying at a minimum;
- An open-source codebase of our models and two anonymised datasets containing users' mobile news reading preferences and the information on the context of use: https://gitlab.fri.uni-lj.si/lrk/mobile-news-perception

Our work provides not only immediate guidelines for UI developers building relatively static "middle ground" solutions for mobile news reading, but also a method enabling personalised context-dependent interface adaptation that can potentially be used beyond news reading, for instance in mobile online social networks or advertising domains.

2 RELATED WORK

The majority of smartphone owners are using their devices to access news content, while for certain categories, such as those aged between 18 and 34 this represents the most likely means of obtaining news [16]. Furthermore, the diversity of smartphone usage patterns [8] is also reflected in the diversity of the ways and situations in which mobile news is consumed. As a result, while traditional media tend to assert their domination in certain locations and contexts (e.g. watching TV at home in the evening) termed "mediatopes" [22], mobile news has penetrated into virtually all segments of our everyday lives, including those that were previously out of the traditional media reach [25].

A pivoting change in how news are consumed stems from the mobile device's ability to engage a user during, what Dimmick et al. term "interstices" – opportunities that arrive at borders of other activities, for instance, while waiting for a friend, and outside of standard mediotopes, e.g. in a supermarket, at a restaurant, etc [5]. Consequently, mobile news exhibit substantially less context stability than standard media [23]. The length of the "interstices" and the opportunistic fashion in which they are exploited lead to news reading sessions that are more frequent and shorter than engagements with legacy media [17].

The affordances of the new medium have, thus, induced changes in how the content is tailored for mobile devices. The "snacking" behaviour with respect to mobile news reading leads to generally shorter and simpler articles [15]. Harnessing the fact that the mobile is a highly personal device, mobile news reading apps adapt the content according to a user's preferences, further refining "the Daily Me" concept enabled by the Web [19]. Finally, a range of sensors available on modern smartphones enable mobile news reading apps to adapt the content to the context of usage, usually to the sensed geographical location of the user [1].

While the content of the news articles has evolved to adapt to the new platform, the way in which the content is rendered on a mobile device has experienced little adaptation. This is especially surprising having in mind the diversity of the context in which mobile news is consumed and the importance of context-informed mobile user interface (UI) adaptation demonstrated in different areas of mobile computing [7]. For instance, previous research points out that context-based UI adaptation may reduce the cognitive load needed for finding information on a mobile's screen [9]. Further, a study of the impact of a user's physical activity on mobile reading demonstrates that the preferred font size may vary depending on whether a user is walking or standing still [27]. Finally, the need for outside brightness-related adaptation was recently recognised even at the mobile operating system level, thus starting from 2019 both Android and iOS phones support switch from light to dark theme at sunset.

The lack of context-driven adaptation is not the only opportunity that mobile news reading apps fail to capitalise on. The UI of these apps could also harness individual preferences for adaptive personalisation [14], and could deliver a tailored version of the UI based on the model constructed for a particular individual. According a study by Constantinides et al. [2], popular mobile news reading apps implement various means of manual interface customization, yet, automatic adaptation is not attempted. With Habito News app [3] the authors capture reading habits and identify three groups of users based on how often and where they access the content, and whether they read articles thoroughly, skim through them, or merely read a few words of the article. The authors then implement automatic app adaptation, yet, the adaptation is done on a single parameter, article layout, does not take the context into account, and harnesses only a coarse model that classifies a user in one of the three predefined groups.

In this paper we investigate both context- and person-based adaptation, and customise a range of parameters including the article layout, the presence of images, the font size, and the theme of the app. Furthermore, our approach goes beyond a three-class classifier, and instead through a combination of active and reinforcement learning for the first time crafts a truly individual model capturing context- and person-dependent mobile news reading preferences.

3 METHODOLOGY

To close the gap identified in the previous section and to deepen our understanding of the relationship between the context of usage and the preferred manner of displaying news information we conduct a real-world study of mobile news reading behaviour. For this purpose we develop a full-fledged mobile news reader app and first conduct a study where the information on the UI preferences is juxtaposed against the context. Based on the data collected in the first study we construct general predictive models that allow us to adjust parameters of the news display according to the context. However, general predictive models do not capture individual differences in the news app display preferences. Therefore, we augment our application with a powerful active & reinforcement learning engine that enables us to identify individual preferences with the least amount of user querying. We then conduct the second real-world study where the app is allowed to monitor the behaviour and construct an on-device personalised adaptation model, which we then evaluate in the remainder of the user study.

3.1 Data Collection App

To capture users' context-dependent preferences, we need to expose the users to various display options. Note that the existing popular news reading applications, such as Flipboard, Google News, BBC News and others, do not allow such data to be collected. From the UI perspective, these apps are adaptable, but not adaptive [2]: while they allow users to change the default theme or font size, the changes remain manual, static, and context-oblivious. Therefore, we construct an application that adapts various UI parameters, but also captures the sensor data pertaining to a user's context, and uses the experience sampling method (ESM) to query the user about the display preferences.

Informed by the existing news apps and the literature on context-adaptive UIs, we opted for allowing the following aspects of the UI to vary within our app:

- **Layout**: represents one of the major UI aspects that can be modified by commercial news reading apps. In our app we support four layout options which were chosen amongst the most frequent ways of arranging news feed items that existing popular news reading applications provide. We labeled them as follows: "largeCards", "gridView", "miniCards", and "xLargeCards" (Figure 1). The first category corresponds to swipe-based up-down navigation where news feed items are represented with a full text description and a medium-sized image. The second option "gridView" has news feed items arranged in a picture grid each having a relatively short text. "miniCards" is similar to the first option, but with smaller images and the length of the news feed description limited to 55 characters. The final option "xLargeCards" enables left-right navigation. In addition, the interface contains only one news feed item at a time with its full description and a large representative image.
- **Font size**: is a parameter whose adaptability is shown to be important in the mobile realm, especially as a user transitions through different physical activities (e.g. sitting, walking, etc.) [27]. In our app we support two values for the font size, "large" and "small".
- **Presence of images**: is included as a toggleable parameter in our app to cater to situations where the images are either unnecessary or impractical, for instance, due to high latency and low bandwidth connectivity. The users can decide to switch the image rendering on or off.
- **Application theme**: is nowadays a standard UI parameter. In our app the users can select between the light or the dark application theme.

Since "gridView" layout cannot be used when images are not present or when "large" font size is selected (as it covers too vast of an area of the image), and there are no differences between "miniCards" and "largeCards" when the images are not rendered, the above parameters result in a total of 22 different combinations of mobile news displays.

An array of sensors embedded in a modern smartphone enable numerous contextual attributes to be collected. Our goal is to cover as many different aspects of the context as possible, but at the same time we want to ensure that the dimensionality of the collected data is low enough to allow reliable modelling. We therefore cover the three broader groups of contextual parameters [13], and within each we sample the following:

- **Technical context:** internet speed, battery level, and screen brightness;
- **User context:** current user's physical activity;
- **Environmental context:** time of day while using the application, and the ambient brightness.

Relying on the best practices in app development and modern tools, such as Ionic, Capacitor in Angular, we build a hybrid (Android and iOS) mobile app that supports battery-efficient sensing of the above contextual parameters. Namely, we implement different sensing modalities as follows:

- **User's physical activity:** we use Google Activity Recognition API to detect whether the user was "still", "walking" or "in vehicle" while using our app;
- **Ambient brightness:** we query the built in sensor at times a user is using the app and obtain a positive numerical value representing outside brightness in lux.
- **Screen brightness:** we query a software sensor returning a numerical value from the interval $[0, 255]$.
- **Time of day:** current hour of application usage ranging from the interval $[0, 23]$.
- **Internet speed:** represented with an integer value from the interval $[0, 2]$ where the increasing value means higher internet speed. The value is calculated based on the signal strength and/or the connectivity type.
- **Battery level:** we query the current battery percentage represented as an integer value from the interval $[0, 100]$.

Finally, through dynamic UI presentation and the experience sampling method (ESM) [6] we assess how various interface renderings are perceived in different contexts. We expose a user to a randomly configured UI by varying the layout, font size, image rendering, and theme parameters each time an app is opened. At any point of time, however, a user can change the display parameters. After a 20-second period without modifications a questionnaire with three statements appears: 1) *Rate your overall view experience, based on personal preference;* 2) *Current view is clear and readable;* and 3) *Current view is informative* (see Figure 1). The first statement measures the overall view experience. The second one measures the perceived readability, whereas the last one assesses the informativeness of the way of displaying the news. Each statement is answered on a five-point Likert scale, providing values ranging from strongly disagree to strongly agree, later mapped to an integer value in the interval $[-2, 2]$. The answers, together with the contextual information sensed each time an ESM questionnaire is fired, are recorded on a device and later synchronized to a remote database.

3.2 Data Collection Campaigns

Our first study aiming to uncover the relationship between the context and the news app viewing preferences included ten participants and lasted for five weeks. The average participant age was 30 with the majority of users between 20 and 23 years old. We collected a total of 836 ESM questionnaire answers during this study. The second study geared towards the evaluation of our on-device UI personalisation method included six participants and lasted for four

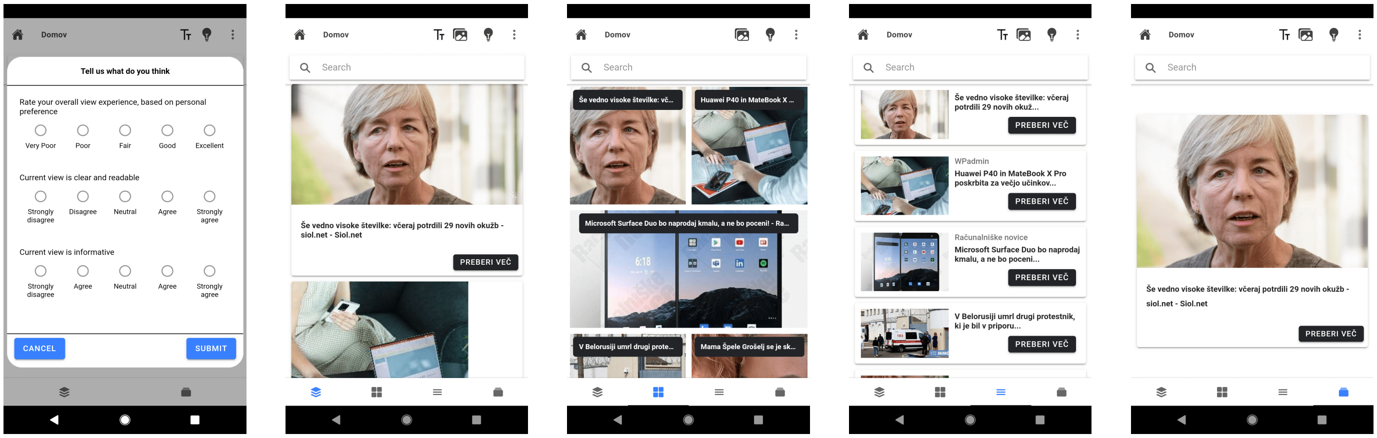

Figure 1: The leftmost image represents the ESM Questionnaire used to assess a user's satisfaction with the configured UI. The following images are snapshots of different application layouts which are labeled in the following sequence: "largeCards", "gridView", "miniCards", "xLargeCards".

weeks. The average participant age was 28 and a total of 606 data points were collected. In both studies participants were recruited from the students and colleagues at our institution and were not reimbursed, monetarily or otherwise, for their participation.

The studies were conducted during the COVID-19 pandemic, which impacted both the contextual data distribution as well as the app usage. Nevertheless, we believe that, collected in-the-wild, our dataset suffices for the identification of the relationship between the context and the news UI viewing preferences, as well as for the evaluation of our novel adaptation algorithm.

4 INVESTIGATING AND MODELLING MOBILE NEWS READING BEHAVIOUR

4.1 Impact of Context and Display Parameters

Our first study resulted in rich multidimensional data about the context (e.g. outside brightness, a user's physical activity, etc.), different UI parameters (layout, theme, font size, etc.) and users' assessments of the interfaces shown. We commence our analysis with the identification of potential links among the recorded variables and the users' perceptions. While the ESM questionnaire contains three questions answered on [-2,2] interval, we observe a very high level of correlation among the three categories, thus we add the scores and work with a single value in the range [-6,6] instead. More specifically, we assign an aggregate score for each UI and context combination that a user has evaluated, whilst dissecting the data along one of the dimensions. We then compare the means of the groups using Welch's ANOVA. We use a .05 p-value cutoff with the Bonferroni correction for multiple (in our case 18) hypothesis testing, thus report as significant only those p-values that are lower than 0.0028. Furthermore, to evaluate the size of the difference between two means we calculate multiple Cohen's d coefficients, taking the suggested .8, .5 and .2 values as the limits for large, medium, and small effect sizes, respectively [10].

User's physical activity is one of the most prominent contextual variables and the one previously shown to have impact on a user's perception of a mobile app interface [28]. Indeed, we find a strong statistically significant link between a user's physical activity and the final view score in the collected data. The Cohen's d coefficients for different pairwise activity comparisons are on the interval $[0.76, 2.18]$. The largest difference between two means of user's scores are observed when users were *still* and *in vehicle*. The smallest variability between the means of users' scores is found when participants were *walking* and *still*.

The battery level recorded in our dataset was overall relatively high, averaging 62%, yet, with significant variability (interquartile range 39.25%). Nevertheless, the value of Pearson's correlation coefficient between the battery level and the final user's score was relatively small (-0.36) as interpreted according to the guidelines from [18], leading to the conclusion that only a weak negative correlation between users' scores and the battery level exists.

Presence of images was noted at 80.4% of the data points, indicating that participants frequently choose to read the news where images were displayed. According to the Welch's ANOVA there is a statistically significant preference for interfaces with images and large difference between the average scores of views with images and those without them, with Cohen's d value of 1.13. Hence, we conclude that presence of images strongly relates with a user's perception of a news app interface.

Views where the **application theme** was set to light were almost always positively rated on the Likert scale, which is not the case with the dark theme. Welch's ANOVA confirms the existence of this difference with a medium effect size (Cohen's d coefficient equal to 0.27). Thus, we conclude that our users exhibit a medium statistically significant preference for the light theme.

Similarly to the application theme, we observe that views using the large **font size** rarely receive negative scores, unlike those using the small font size. Furthermore, the difference between the mean score of the small and large font size views is statistically significant and the effect size, according to the Cohen's d coefficient, is medium. Therefore, we conclude that our users have a medium preference for the large font.

Layout is a particularly interesting parameter that may affect the news app interface perception due to a vast design space of possible layouts and a potential role of mobile UI trends (e.g. Android Material design [21]) that may affect interface acceptance. Nevertheless, even within a modest space defined by the four layouts used in our application we find statistically significant differences (according to Welch's ANOVA) between the mean scores of different layouts. Values of Cohen's d coefficients range across the interval $[0.23, 1.1]$. Therefore, we conclude that there is a statistically significant medium to large preference for certain layouts, with "largeCards" being the most preferred and "gridView" being the least preferred.

Finally, we note that one of the parameters, Internet connectivity speed, was omitted from the analysis as all the data points indicated good connectivity during the app use. In addition, no statistically significant results were noted with the following parameters: screen brightness, time of day, and ambient brightness.

4.2 Hierarchical modelling

The statistical analysis elaborated in the previous section does not take into account the combined effect of different contextual and UI parameters on the user's perception of the mobile news app interface. To address this issue we used hierarchical linear models. They represent an extension of simple linear regression models that incorporates the dependence between parameters arising from a hierarchical structure. Building such models is usually an iterative process where we gradually increase the complexity of a previously built model by adding a new parameter as a part of a fixed or a random effect. To evaluate whether additional parameters are necessary we use AIC (Akaike information criterion) and BIC (Bayesian information criterion) to compare the goodness of a new model and the previously built model that does not contain a newly introduced variable. Both AIC and BIC evaluate the amount of information loss when a statistical model is used to represent the process that generated the data, with BIC incurring a larger penalty for models that include a larger set of parameters.

Before a multilevel model is constructed we need to decide whether each parameter will be used as a fixed or a random effect. We must also consider the number of different values of grouping variables. A small number of different values (less than five) causes lack of information to accurately estimate group-level variation [12]. Due to the fact that non of our categorical parameters consist of five different values we introduced a new parameter labeled as **"layout_images"**, which represents a combination of the layout parameter and the presence of images. We combined these two parameters since they represent the most informative view parameters, as the average users' scores differ the most when participants were using different layouts whilst having images displayed or not.

We construct the first **intercept-only model** to estimate the appropriateness of the grouping variable – "layout_images". The models allows merely the regression function intercept to vary across different values of random effect parameters. By calculating

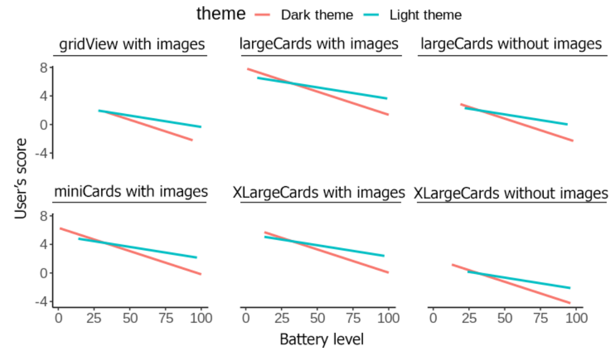

Figure 2: Regression functions of a multilevel model which includes an interaction term between the app theme and the battery level, grouped by a random effect parameter "layout_images". We see that the predicted scores generally vary among the "layout_images" groups, depend on the app theme (with the light theme generally being favored), yet also depend on the combination of the current battery level and the used theme.

the intraclass correlation coefficient (ICC) we estimate how much of the user score variation is explained by the clustering along our "layout_images" dimension. We find that the ICC for the intercept-only model is equal to 0.31 indicating relatively weak grouping.

In the next step we include a **user's physical activity** as a fixed effect parameter. We find that a user's physical activity as the first-level predictor increases the performance of our model, since it reduced both AIC and BIC compared to the intercept-only model. Increasing the complexity of the previous model by adding a term that accounts for **the interaction between the font size and user's physical activity** improves the model further. A deeper inspection of the new model uncovers that although participants generally prefer large font size, the preferences for large font size is particularly obvious when the users are walking. The contrary holds as well: a combination of walking and the smaller font size leads to approximately -2.47 lower scores according to our model.

Using the **application theme** as a part of the fixed effect parameters further improves the previously built model. Examining the coefficients of the new model we find that the participants' scores are higher when the theme is set to light. Accounting for the **interaction between the application theme and battery level** improves the previous model. We observe that users' scores are decreasing when the battery level is increasing. However, when battery level is low (approximately below 30%) the scores for the dark application theme are higher than the scores for the light theme. We visualise the relevant regression lines in Figure 2.

Intuitively, the ambient brightness affects the choice of the application theme. Indeed, adding a term that captures **the interaction between the ambient brightness and the application theme** improves the model (reduces both AIC and BIC). Further investigation of the new model reveals that participants rate light application theme higher when the environmental brightness increases, while the users prefer the dark theme in darker environments. Note that a simple statistical analysis we conducted in Section 4.1 fails to account for this subtle effect of the ambient brightness on the news app UI perception.

Adding the remaining contextual parameters, time of day and screen brightness, does not improve our model. To conclude our final best performing multilevel model includes "layout_images" parameter as a random effect and the following parameters with previously mentioned interactions as a part of fixed effects parameters: user's physical activity, font size, application theme, ambient brightness, and battery level.

4.3 Predictive modelling

The statistical analysis presented in Section 4.1 points towards the contextual and the UI-related variables that may impact the users' perception of the displayed mobile news screen. The hierarchical analysis from Section 4.2 further uncovers the structure among the predictors, for instance, uncovering the interplay between the physical activity and the font size. Nevertheless, these analyses are not suitable for predicting whether a user will favorably rate a particular view in a particular context. For this, we construct machine learning models that take the context and view-related parameters' values at the input and predict the user's score.

To make the prediction task more realistic, albeit more challenging, from our dataset we filter out the instances for which the images were not shown, since views with images tend to be preferred overall. Further, we filter out "miniCards", and "gridView" layouts, due to generally low scores received by these layouts. This is to mimic logical design decisions of not providing an image-free option nor layouts that are not favoured under any circumstances. Thus, the interface variables we have at our disposal are layout (with possible values "largeCards" and "xLargeCards"), theme ("dark" or "light"), and font ("large" or "small"), giving us a total of eight combinations of the **"layout_theme_font"** outcome (referred to as a "view" in the rest of the paper).

To reduce a chance of overfitting, we exclude time of day and screen brightness, as these parameters proved to be unimportant in our previous statistical analyses. Finally, we compared the effect size of the interactions between the battery level and the application theme, as well as the ambient brightness and the application theme in our final multilevel model. The absolute value of the interaction term between the former variables was lower than interaction between the latter variables. Therefore, we omit the battery level to simplify our predictive model.

Analysing the distribution of answers to the ESM questionnaire we notice that user's have generally (in 63% cases) answered all three statements from the questionnaire with the same Likert scale value. Furthermore, the majority of user's answers were positive (75% cases) and skewed towards the maximum value. Splitting our dataset so that the highest scored questionnaires are coded as "yes – the view is appropriate" and questionnaires scored lower than the maximum value coded as "no – the view is not appropriate" results in a roughly even class distribution and we use such a split for training and testing our machine learning models.

We train Random Forest, Naive Bayes , AdaBoost, and Decision Tree classifiers. We divide our dataset randomly into train (contains 80% dataset points) and test sets (contains 20% dataset points), and harness cross-validation over the training set for hyperparameter adaptation. We then evaluate the best classifiers on the test set and display the results in Table 1. The values indicate a clear superiority of the random forest model. Still, with 75% accuracy our best model is far from perfect. This could partly stem from a limited dataset we have collected. Yet, we suspect that the main culprit is the lack of adaptation to individual behaviour. Thus, in the next section we harness active and reinforcement learning to personalise our model in real time.

Algorithm	Prec.	Recall	F1-score	Acc.	AUC
Random Forest	0.75	0.77	0.76	0.75	0.78
Naive Bayes	0.64	0.31	0.42	0.56	0.65
AdaBoost	0.68	0.66	0.67	0.67	0.71
Decision tree	0.77	0.66	0.72	0.73	0.75
Majority classifier	0.00	0.00	0.00	0.50	0.50

Table 1: Performance of general predictive models.

5 MODELLING PERSONALISED MOBILE NEWS READING BEHAVIOUR

In the previous section we devised context-aware, yet general models for predicting an appropriate way of displaying mobile news content. On the other hand, news reading tends to be performed differently by different (groups of) people [2]. Therefore, in this section we build personalised predictive models, which allow iterative learning of a user's preferences and subsequent automatic adaptation of the mobile news app views.

We conduct a separate study divided into two parts, each lasting for two weeks. During the first half of the study, starting from the random forest-based general predictive model (described in Section 4.1), on each mobile device we develop a personalised model for that device's owner. The personalisation is guided by the feedback from the user, thus to reduce the querying burden, we specifically optimise our approach to miminise the number of questionnaires sent out to each participant. During the second half of the study, we interchangeably adapt the views shown by the application using either the personalised or the general model. A user's opinion towards the displayed view is captured via experience sampling, enabling us to assess whether, and to what extent, our personalised approach manages to deliver more favourable views than an approach based on the contextual adaptation only.

5.1 Active & Reinforcement Learning for UI Adaptation

Active learning aims to maximise the performance of a machine learning model, while minimising the amount of labelled training data needed for the model construction [24]. It does so by requesting the labels for only those training data instances that are the most useful for the classifier construction. In our setting, the labels are provided by a user – she decides whether the displayed view is appropriate or not. Thus, in our mobile system active learning could be used to minimise the number of questionnaires sent out to a user, focusing on interrupting a user only when it is expected that the acquired information (i.e. the label) is going to substantially improve the quality of the model (i.e. classifier).

Deciding whether labelling an instance is going to "pay off" or not is not trivial and several strategies for determining the informativeness of an unlabeled instance exist. However, the utility of a particular strategy may differ depending on the underlying data distribution, which, in our case implies, that the strategy might be user-dependent. To adapt the model to individual users with minimum querying, we harness four different active learning strategies and embrace a *reinforcement learning* framework in order to converge to the best fitting strategy for a particular user. More specifically, we start with an existing model that predicts to what extent each of the eight **layout_theme_font** combinations will be favoured by the user and shows the highest ranked combination (i.e. the one for which the random forest confidence is the highest). At each opportunity to get feedback from the user, we rely on the following active learning strategies (or none) to label an instance:

- **Margin of confidence:** decides whether the user needs to label an instance or not by calculating the difference between the confidence of top two most confident predictions. In our case, the questionnaire is shown if the distance between the top two confidences falls below 20%.
- **Random sampling:** even though it is the simplest strategy it explores the most diverse feature space since the decision whether to query a user or not is random.
- **Random sampling based on the distribution of user's physical activity values:** a user's physical activity is one of the contextual parameters that has the large impact on the final user's score. However, most of the data points (approximately 66%) from our first study were collected when participants were still whilst reading the news. Therefore, we decided to prioritise surveying when participants are in a vehicle or walking by lowering the probability of showing a questionnaire whilst users are being still.
- **Least confidence sampling:** asks a user to label instances where the most confident prediction drops below a threshold. We set this value to 0.5, meaning that we query a user when the prediction confidence is 0.5 or lower.

To select which strategy to use, we must balance between the exploitation of an apparently well-working strategy and the exploration of potentially better strategies. This problem is often termed the multi-armed bandit and is described with a gambler presented with the opportunity to pull any of the n one-armed bandit machines. Wishing to maximize the overall reward, in each step the bandit has to decide whether to pull the lever that was the most profitable up until the current point or to explore other levers [26]. There are numerous ways in which a particular lever, or in our case, a particular active learning strategy, can be selected. We use a lightweight reinforcement learning algorithm termed the *upper confidence bound (UCB)* [4]. The fundamental equation of the UCB algorithm is the following:

$$A_t = argmax_a(Q_t(a) + c\sqrt{\frac{ln(t)}{N_t(a)}}) \qquad (1)$$

The first factor in the equation ($Q_t(a)$) represents the current reward estimate for action a calculated by averaging the previously obtained rewards for that particular active learning strategy. The second factor represents the uncertainty in the estimates that optimistically assumes that the least frequently selected action (i.e. the one for which the number of times the action has been tried – $N_t(a)$ – is the lowest) could be the most appropriate action. In the initial phase the UCB algorithm selects each action once, since it has no previous knowledge about the most appropriate action. In the later iterations the UCB algorithm proceeds to converge towards the best action. In the equation, c represents a parameter balancing between the impact of the exploitation and the exploration.

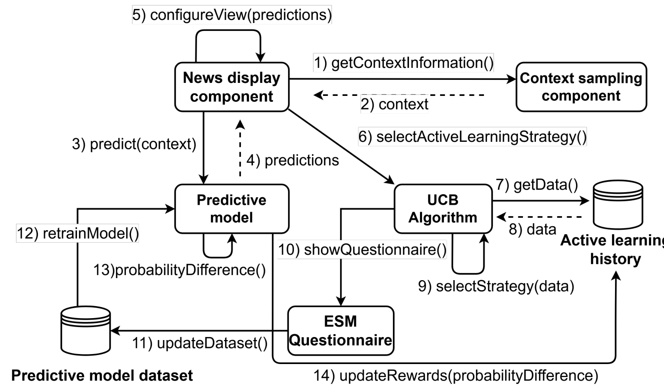

Figure 3: Active and reinforcement learning protocol.

The award function is defined so that the selected active learning strategy is rewarded in case the newly-provided label leads to an *increased classification confidence* of the true class. Intuitively – we reward an active learning strategy that imporoves our understanding of the user. The label for each decision is provided with the ESM questionnaire shown to a user in two different scenarios. The first case being when the selected active learning strategy decides that querying is required, whereas the second one ensures that user's quick (under 20s) modifications of the configured view are intentional.

To conclude, our model personalisation protocol (Figure 3) proceeds as follows. When a user opens an app, the physical activity and ambient brightness are sampled (step 1 and 2), the predictive model is used to assess the suitability of each **layout_theme_font** combination (steps 3 and 4), with the highest ranked combination shown to the user (step 5). The UCB algorithm is then instructed to decide whether to query the user or not. The decision is made from the history of each active learning method, i.e. the award and the number of trials as per equation 1 (steps 7, 8, and 9). In case the ESM questionnaire is shown to the user (step 10), the feedback is noted (step 11) and used to retrain the predictive model (step 12). The difference between the correct label prediction confidence of the new and the previous model is then used as a reward or penalty for the selected active learning strategy (step 13).

5.2 Personalised Predictive Modelling – User Study and Results

We deploy our updated mobile news reader app fitted with the above reinforcement & active learning personalisation framework to six phones and allow for two weeks of in-the-wild use for the models to converge. In the subsequent two weeks we compare the

ID	Rewards	GM accuracy	PM accuracy	No. inst. 1^{st} part	No. inst. 2^{nd} part
0	$[31.72, 5.47, -3.40, -1.98]$	47%	71%	80	77
1	$[3.79, 4.08, -1.90, -1.93]$	44%	69%	16	22
2	$[15.34, 2.07, -1.95, -1.75]$	52%	63%	25	80
3	$[26.04, 12.60, -1.41, -1.85]$	47%	73%	70	41
4	$[-0.02, 2.50, -0.50, -2.50]$	68%	98%	78	79
5	$[1.38, 0.39, -1.71, -2.25]$	33%	71%	25	13

Table 2: Performance comparison of the general model (GM) and the personalised model (PM) based on active & reinforcement learning. For each user (represented by a separate line in the table), the PM performs better than GM.

performance of personalised predictive models with the general one (described in Section 4.3) by alternating between the adaptation based on the general model and the personalised model every 24 hours, and querying the users via ESM to rate the appropriateness of the displayed view.

In Table 2 we summarize the results for both general and the personalized predictive model, and for each active learning strategy, which was selected by the UCB algorithm. The first column contains unique user identifiers, while the second one contains vectors representing rewards of the UCB algorithm for each of the active learning strategies: margin of confidence sampling, random sampling, random sampling based on the distribution of user's physical activity, and the least confidence sampling, respectively. The third column represents prediction accuracy for the general predictive model, whereas the fourth one contains prediction accuracy for the personalised model. The final two columns represent the number of instances each user has contributed in the first and the second part of the study.

5.2.1 Predictive performance and algorithmic decisions. The results demonstrate that the personalised predictive model achieves greater accuracy, i.e. a higher ratio of views labelled as "appropriate", with all six users. The mean accuracy increase over the general model is 26%, ranging from 11% and 38% across all users. Analysing the per-strategy rewards from Table 2 we conclude that for the majority of the users, the algorithm converged towards repeatedly selecting *the margin of confidence* and *random sampling* strategies (the first two fields in the "Rewards" column). The result is interesting, as it demonstrates that a combination of a method that explores the most extreme values (the margin of confidence) and the one that explores a more diverse area (random sampling) is a suitable means for driving the personalisation.

5.2.2 Personalisation factors. Finally, we aim to infer which aspects of the view require context-based personalisation. Was it the layout, the font size, or the app theme that the personalised model adapted the most? We compare the distributions of the values for the three aspects of the view between the data obtained with the personalised and the data obtained with the general model in different contexts. Since the last parts of the study were done during a severe COVID lockdown we limit ourselves to the analysis of adaptation to the environmental brightness contextual parameter, as the physical activity parameter values were highly skewed towards "still". We use Wasserstein distance (also known as earth mover's distance [11]) to compare the distributions (over a single user) of font size, app theme, and layout values selected by the general and the personalised model. We then average the results over all users. We observe that the font size parameter is subject to the largest level of personalised adaptation, however, the layout and the app theme trail close behind, indicating that all three parameters need to be considered when personalising a mobile news app interface according to the context.

6 DISCUSSION

The overarching research question we aim to address in this paper is whether a user's perception of the interface of a mobile newsreader app is dependent on the context and personal preferences, and how can automatic sensing of relevant factors be harnessed to intelligently adapt the app's display according to the situational and personal preferences. The actual definition of the notion of "context", "interface", and "preferences" evolved throughout the two studies we have conducted. First, we noted that certain contextual parameters do not vary significantly in our dataset (e.g. "Internet speed"), thus cannot have any effect on the outcome, while some parameters tend to be redundant in presence of other, more relevant parameters (e.g. "time of day" and "screen brightness" in presence of "ambient brighness"). We developed four different layouts (Figure 1) adaptable to show images or not, modify font size, and change the theme brightness. Nevertheless, users highly preferred that images are shown and favoured two of the layouts "largeCards" and "xLargeCards" over the others. Insisting on preserving all of the originally designed layouts throughout the analysis would boost the accuracy scores, but would also prevent us from exploring the nuanced relationship among personal preferences, context, and different aspects of the display, and have us focus on answering a relatively trivial question of whether a display is well designed or not. Finally, while we collected the preference data on a Likert scale, we have soon realised that the answers are skewed, and have therefore decided to identify whether the users gave the maximum rating to a displayed view or not. Conveniently, such a decision resulted in a roughly evenly balanced dataset.

The machine learning methodology, too, evolved over the course of our research. We have first conducted the investigation of individual contextual factors and later their combination on the score that a particular view has received. Faced with an uneven number of datapoints per user we used hierarchical modelling. We believe that this is the only statistically sound method for addressing such datasets, and avoids the issues that arise with alternatives such as averaging per-user data (which reduces the explanatory power of the data) or bundling all the data together (which is plain incorrect). The predictive modelling we performed with the data collected in the first study (Section 4.3) relied on the analysis of an already collected trace. Thus, the particular task our machine learning algorithm solved was whether a particular combination of layout, theme, and font size shown in a given context is "appropriate". Note that this is a pure binary decision problem, as we could not change the shown combination during the analysis. The second study, instead focuses on a much more practical issue of which layout, font size, and theme combination to show to a user. While the feedback is again binary – the display is appropriate or not – we now have eight combinations to chose from in real time. Thus, the personalised predictive modeling results from Section 5.2 should not be compared with the results from Section 4.3.

The major limitation we faced during our analysis stem from the restrictions imposed during the COVID-19 pandemics, the start of which coincided with the start of our first data collection campaign. Not only did the restrictions impact our ability to recruit volunteers, but it also impacted the breadth of the collected data – some contextual factors, in particular those related to a user's physical activity, exhibited less variation that expected. Furthermore, the restrictions were fluctuating throughout the course of our data collection, with various movement limitations, e.g. closed schools, public transport shutdowns, movement limited within a municipality, etc., active at various times. Nevertheless, we believe that the final result of our study, the active & reinforcement learning framework for context-based personalised adaptation, is well engineered, and we are prepared for recruiting a larger group of users as soon as the situation allows us to. Furthermore, we have prepared the codebase and the two anonymised datasets (available at https://gitlab.fri.uni-lj.si/lrk/mobile-news-perception) for other researchers to investigate and use in their applications.

7 CONCLUSION

In this paper we investigated the impact of personal preferences and contextual factors on the perception of a mobile newsreading app. We conducted two user studies with the total of 16 users that together lasted 9 weeks, and identified the key contextual factors, a user's physical activity and the ambient brightness, that affect the display perception. We then devised an active & reinforcement learning-based method that runs on a mobile device and builds a personalised context-aware model guiding the UI adaptation. The method led to an improved acceptance of the displayed view among all of the users in our in-the-wild study. We believe that our contribution may complement the existing content personalisation efforts and lead towards truly personalised mobile news reading experience.

REFERENCES

[1] Chih-Ming Chen. 2010. Intelligent location-based mobile news service system with automatic news summarization. *Expert Systems with Applications* 37, 9 (2010), 6651–6662.

[2] Marios Constantinides and John Dowell. 2016. User Interface Personalization in news apps. In *CEUR Workshop Proceedings*, Vol. 1618. CEUR workshop Proceedings.

[3] Marios Constantinides, John Dowell, David Johnson, and Sylvain Malacria. 2015. Exploring mobile news reading interactions for news app personalisation. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services.* 457–462.

[4] Wojciech Czarnecki and Igor Podolak. 2014. Adaptive Active Learning as a Multi-armed Bandit Problem. *Frontiers in Artificial Intelligence and Applications* 263 (01 2014), 989–990.

[5] John Dimmick, John Christian Feaster, and Gregory J Hoplamazian. 2011. News in the interstices: The niches of mobile media in space and time. *New media & society* 13, 1 (2011), 23–39.

[6] Nikolaos Dimotakis, Remus Ilies, and Timothy Judge. 2013. *Experience Sampling Methodology.* https://doi.org/10.4324/9780203585146

[7] Jacob Eisenstein, Jean Vanderdonckt, and Angel Puerta. 2000. Adapting to mobile contexts with user-interface modeling. In *Proceedings Third IEEE Workshop on Mobile Computing Systems and Applications.* IEEE, 83–92.

[8] Hossein Falaki, Ratul Mahajan, Srikanth Kandula, Dimitrios Lymberopoulos, Ramesh Govindan, and Deborah Estrin. 2010. Diversity in smartphone usage. In *Proceedings of the 8th international conference on Mobile systems, applications, and services.* 179–194.

[9] Leah Findlater and Krzysztof Z Gajos. 2009. Design space and evaluation challenges of adaptive graphical user interfaces. *AI Magazine* 30, 4 (2009), 68–68.

[10] Catherine Fritz, Peter Morris, and Jennifer Richler. 2011. Effect Size Estimates: Current Use, Calculations, and Interpretation. *Journal of experimental psychology. General* 141 (08 2011), 2–18. https://doi.org/10.1037/a0024338

[11] A. Gardner, C. A. Duncan, J. Kanno, and R. R. Selmic. 2018. On the Definiteness of Earth Mover's Distance and Its Relation to Set Intersection. *IEEE Transactions on Cybernetics* 48, 11 (2018), 3184–3196.

[12] Andrew Gelman and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge University Press. https://doi.org/10.1017/CBO9780511790942

[13] Claude Ghaoui. 2005. Encyclopedia of human computer interaction. *Perspectives* 2021 (2005), 299.

[14] Anthony Jameson. 2007. Adaptive interfaces and agents. In *The human-computer interaction handbook.* CRC Press, 459–484.

[15] Nancy Miller. 2007. Manifesto for a new age. *Wired Magazine, March* (2007).

[16] Amy Mitchell and Tom Rosenstiel. 2012. The state of the news media 2012: An annual report on American journalism. *Pew Research Center Project for Excellence in Journalism. http://stateofthemedia. org* (2012).

[17] Logan Molyneux. 2018. Mobile news consumption: A habit of snacking. *Digital Journalism* 6, 5 (2018), 634–650.

[18] Mavuto Mukaka. 2012. Statistics Corner: A guide to appropriate use of Correlation coefficient in medical research. *Malawi medical journal : the journal of Medical Association of Malawi* 24 (09 2012), 69–71.

[19] N. Negroponte. 1996. *Being Digital.* Vintage Books. https://books.google.si/books?id=5A-u4op92gEC

[20] Nic Newman, David AL Levy, and Rasmus Kleis Nielsen. 2015. *Reuters Institute digital news report 2015: Tracking the future of news.* Reuters Institute for the Study of Journalism.

[21] Lumpapun Punchoojit and Nuttanont Hongwarittorrn. 2017. Usability Studies on Mobile User Interface Design Patterns: A Systematic Literature Review. *Advances in Human-Computer Interaction* 2017 (09 Nov 2017), 6787504. https://doi.org/10.1155/2017/6787504

[22] Thorsten Quandt and Thilo von Pape. 2010. Living in the mediatope: A multimethod study on the evolution of media technologies in the domestic environment. *The Information Society* 26, 5 (2010), 330–345.

[23] Kim Christian Schrøder. 2015. News media old and new: Fluctuating audiences, news repertoires and locations of consumption. *Journalism studies* 16, 1 (2015), 60–78.

[24] Burr Settles. 2009. *Active Learning Literature Survey.* Computer Sciences Technical Report 1648. University of Wisconsin–Madison.

[25] Samson Struckmann and Veronika Karnowski. 2016. News consumption in a changing media ecology: An MESM-study on mobile news. *Telematics and Informatics* 33, 2 (2016), 309–319.

[26] Richard Weber et al. 1992. On the Gittins index for multiarmed bandits. *The Annals of Applied Probability* 2, 4 (1992), 1024–1033.

[27] Tetsuo Yamabe and Kiyotaka Takahashi. 2007. Experiments in mobile user interface adaptation for walking users. In *The 2007 International Conference on Intelligent Pervasive Computing (IPC 2007).* IEEE, 280–284.

[28] T. Yamabe and K. Takahashi. 2007. Experiments in Mobile User Interface Adaptation for Walking Users. In *The 2007 International Conference on Intelligent Pervasive Computing (IPC 2007).* 280–284. https://doi.org/10.1109/IPC.2007.94