IDEA – Informfully Dataset with Enhanced Attributes

In this repository, we share a list of different datasets that resulted from a 2-week field experiment with the Informfully app, where we exposed N = 593 respondents to daily scraped news feeds consisting of 26 articles across 13 different topics per day.

Note that week 1 had a clear 2x2 experimental design where only the placement of two of 26 articles as well as the content of one of the 26 articles varied across groups. Data from week two may be more complex to analyse as we introduced topic personalisation here, meaning that users may have seen different articles in different orders (even within groups).

Below you can see 1) an overview of the different datasets that we share including a brief description of what they entail and the number of unique data points per dataset, 2) a detailed description of the experimental setup and different recommendation logics, and 3) codebooks per dataset that describe all variables that were included (incl. item wordings for self-reported measurements).

Table of contents

Overview

Dataset	Description	Datapoints
Articles	Collection of all news articles that were shown in the app.	10954
Bookmarks	Comprises all instances where a user bookmarked an article.	2479
Favourites	Comprises all instances where a user favourited an article.	3115
Interactions	Comprises all interactions of valid users with individual articles between Nov 23rd and Dec 8th.	34890
PageViews	Comprises all instances where a user accessed any page on Informfully (including articles, homescreen, settings, etc.).	84747
Ratings	Comprises all instances where a user rated an article as either easy to understand or interesting.	28382
Recommendations	Stores every article that was recommended to any given user during the course of the experiment.	207220
ReadabilityScores	Stores readability scores for all nudged articles (both in their original and rewritten version)	32
Surveys	Comprises all responses to in-app surveys from week 1 and week 2.	43078
Users	Comprises all users that were active between Nov 23rd and Dec 8th. Includes a list of self-reported attitudes and preferences (see codebook), as well as two.	593

Recommendation logics

The data in this repository stems from a 2-week long field experiment on the effects of choice architecture design on engagement with and learning from environmental news. As we exposed users to different experimental conditions throughout the experimental period, the extent to which different users' news feeds resembled each other varied.

During the first experimental week, we had a simple 2 by 2 experimental design where we only varied the position (position 1 vs position 5) and the content (original vs rewritten) of the article. Accordingly, all users saw the same articles in the same order apart from this manipulation.

In the second experimental week we introduced explicit and implicit topic personalisation for a sub group of users (see also Figure 1). For this, we either asked users for their explicit top 3 preferred news topics, or inferred them implicitly from their reading behaviour during the first week. If topic personalisation was present, we then populated positions 2 through 4 and 6 through 8 with their preferences. Positions 9 through 26 consisted of 9 preference-based articles (3 articles per topic preference) and 9 filler articles (1 article per non-preferred topic to ensure sufficient diversity) that were randomly ordered. If no topic personalisation was present the news feeds looked just like they did in the first week.

Finally, while we had planned to keep the initial four experimental conditions constant throughout the experiment, due to a technical issue the group allocations of some users changed. This is reflected in the dataset (see variables experimental_group_w1 and experimental_group_w2 in the users dataset).

Position	Week 1 (random except position 1 and 5)	Week 2 (preference based except position 1 and 5)
1 (fixed)	**Environment**	**Environment**
2	Business	Sport
3	World News	Business
4	Sport	UK News
5 (fixed)	**Top news of the day**	**Top news of the day**
6	Football	Sport
7	Crime	Business
8	UK News	UK News
9 through 26	Random order of filler articles (two articles per topic)	Random order of 9 preference-based articles (3 articles per topic preference) and 9 filler articles (1 per topic)

Table 1. Example news feeds for experimental week 1 and 2.

Codebooks

In the following codebooks, we report and explain all variables for each dataset and report the exact items for self-reported measures. Scale reliability analyses can be found in the attached pdf of our RMarkdown output.

Articles

Articles	
article_id	*Unique article id*
flag	*Indicates the nature of an article in relation to the original experimental set-up. If no flag is added, the original article was scraped. Flags are:* - *Popular: This is a popular article from the Guardian in its original form* - *Nudge-wnv_OG: This is an environmental article from the Guardian in its original form* - *Nudge-env_RW: This is an environmental article from the Guardian that was automatically rewritten with ChatGPT for lower text complexity*
url	*Link to the original article*
primaryCategory	*Topic category as indicated by the Outlet's website*
subCategories	*If present, subCategories are tags assigned to articles on news outlets' websites*
title	*Title of the article*
lead	*Teaser of the article*
author	*Author(s) of the article*
datePublished	*Timestamp of when the article was published online by a given outlet*
dateScraped	*Timestamp of when the article was first scraped*
dateUpdated	*Timestamp of last update to the article*
outlet	*Outlet that published the news article*
image	*Link to the image of the article*
word_count	*Number of words in an article*
sentence_count	*Number of sentences in an article*
date	*Date on which the article was shown in news feeds*

Bookmarks

Bookmarks	
_id	*Unique id*
articleID	*Id of the article that was bookmarked*
informfully_id	*Unique user id*
createdAt	*Timestamp of when the bookmark was made*
articlePublishedDate	*Timestamp of when the article was published*
removedAt	*Timestamp of when the bookmark was removed*

Favourites

Favourites	
_id	*Unique id*
articleID	*Id of the article that was bookmarked*
informfully_id	*Unique user id*
createdAt	*Timestamp of when the article was favourited*
articlePub-lishedDate	*Timestamp of when the article was published*
removedAt	*Timestamp of when the article was unfavourited*

Interactions

Interactions	
_id	*Unique interaction id*
articleId	*Unique article id*
informfully_id	*Unique user id*
createdAt	*Timestamp of first interaction with the article*
updatedAt	*Timestamp of latest interaction with the article*
articlePublishedDate	*Timestamp of when article was scraped*
duration	*Overall time spent on article in milliseconds*

maxScrolledContent	*Maximum scroll percentage for any given article*
views	*Number of separate clicks on a given article*
date	*Date of the interaction*
unique_days_before	*Unique days before that a given user was active*
Experimental_week	*Indicates whether an interaction took place in the first or second week of the experiment (that is, before or after the first in-app survey). This matters for how the news feeds were assembled with different recommendation logics in different weeks.*

Page Views

Page Views	
_id	*Unique page view id*
Informfully_id	*Unique user id*
Page	*The page being logged*
previousPage	*The previous page that the user was on*
article_id	*Internal unique identifier for page that was viewed (e.g. homepage, settings, …)*
isBookmarked	*Indicates whether an article is in the bookmarks*
isFavourited	*Indicates whether an article has been favourited*
primaryCategory	*Primary Category of the article viewed*
fromArticleScreen	*Indicates how an article was opened (e.g. from favourites, bookmarks, homescreen, …)*
createdAt	*Timestamp of when the page view was created*

Ratings

Ratings	
_id	*Unique id*
informfully_id	*Unique user id*
articleQuestionId	*ID indicating the question that was asked (Note that ids vary across conditions and between users; see legend below).*

	QuestionId Legend: *Question 1 (I find this article interesting to read):* • *DZyvQCx4sTr6xhd67* • *yZ7Db3PsTNW2na9hZ* • *EiEAq5Ms4WfukLwG8* • *bQ925uvAhfuLJX9w8* • *8chxji6FtzC7c6eca* • *FKf2wJTAEMSDipPga* • *9kkHAP44YsWdv238y* • *HNTDMFkjt3Ynvhy7d* *Question 2 (I find this article easy to understand):* • *yGoJwWDeDdug6x3Dg* • *ZiZJw4TLtxzGkPpGY* • *XpAM2pajfeiP6MXrR* • *yx9tNErJn37qNWZZq* • *HozRh4v7WkCPfH6je* • *vn3xBCvWXt8AvfMyf* • *wCJoa9Jwpte34MB8i* • *pu5tmJQ2AtYthuMCv*
articleAnswer	*Response to a question; either 1 (thumbs up) or -1 (thumbs down)*
createdAt	*Timestamp of when the rating was given*
removedAt	*Timestamp of when the rating was removed*

Recommendations

Recommendations	
informfully_id	*Unique user id*
articleId	*ID indicating which article was recommended*
recommendationAlgorithm	*Indicates the underlying recommendation logic (either the original random recommendation logic or a more personalised version for a sub group of users in the second week)*
Prediction	*Indicates the position in the news feed. 100 is position 1, 99 is position 2, and so forth*
createdAt	*Timestamp of when the recommendation was created*
Experimental_week	*Indicates whether an interaction took place in the first or second week of the experiment (that is, before or after the first in-app survey). This matters for how the news feeds were*

	assembled with different recommendation logics in different weeks.
Flag	*Indicates whether an article was a filler article, a popular article, a nudged article in its original, or a nudged article in its rewritten form*

ReadabilityScores

Readability Scores Calculated with the Python readability package	
articleId	*ID indicating which article was recommended*
X_FRE	*Title/Teaser/Text Flesch reading ease score*
X_GFI	*Title/Teaser/Text Gunning Fog index*
X_K	*Title/Teaser/Text Kincaid*
X_CL	*Title/Teaser/Text Coleman_Liau*

Surveys

Surveys *Surveys are generated from a question pool, meaning that the same questions may appear in different surveys.*	
informfully_id	*Unique user id*
X_id	*Unique interaction id*
surveyId	*ID of the survey as fielded inside the app. Note that any survey is a collection of questions.*
createdAt	*Timestamp of first interaction with a (battery of) question(s)*
questionID	*ID of the individual survey question fielded within a particular in-app survey. Since surveys are generated from a question pool, the same question may be present in several surveys.*
questionText	*Survey item as presented to respondents in the in-app surveys*
selections_id	*ID for a given response option within a particular survey*
selections_text	*Response option*
selections_value	*Numerical response value (for knowledge items, values are 1 if the answer was correct and 0 if not)*

selectionsUserInput	*Responses to open questions if applicable*
question	*question code:* *rec_XX_01: Recall question for day XX (general gist question)* *rec_XX_02: Recall question for day XX (factual question)* *rec_XX_03: Recall question for day XX (which one is not true question)* *fillerX_Y: Filler knowledge question X item Y* *sat_X_wY: Satisfaction qustion ; item X in week Y* *talked_env: "How many days during the last week did you talk to friends or acquaintances about the environment?"* *read_env: "How many days during the last week did you read news about the environment outside of this app?"* *opn_pro: "Is there anything that you particularly like about this app?"* *opn_con: " Is there anything that you particularly dislike about this app?"* *opn_sug: "Do you have any suggestions for how we can improve the app moving forward?"* *subj_kno_wX: Subjective knowledge question in week X*
date	*Date where survey was completed*
first_survey	*Date where first survey was completed*
week	*Indicates whether the survey took place after week1 or week2*
knowledge_question	*Dummy variable indicating whether the response is an answer to a knowledge question*
open_question	*Dummy variable indicating whether the response is an answer to an open question*
payment_question	*Dummy variable indicating whether the response is an answer to payment questions that were posed in the second in-app survey*
doesn't_know	*Dummy variable indicating whether the response was don't know*
article_displayed	*If a response concerns a knowledge item, this variable indicates the date where it was displayed to respondents*
survey_taken	*Timestamp of when the survey was taken*
days_difference	*Days between seeing an article and answering a knowledge question about it (note that if respondents fell out of the standard schedule for filling in the in-app surveys, they may have been questioned about some of the articles twice)*

Users

Context variables	
informfully_id	*Unique user Id*
duration in seconds	*Amount of seconds that it took a given respondent to complete the Qualtrics intake survey*
informed_consent	*1 if respondent gave informed consent, 0 otherwise*
experimental_group_w1	*Group allocations for week 1* *1 = Original environmental news article nudged on position1* *2= Rewritten environmental news article nudged on position 1* *3 = Original environmental news article shown on position 5* *4 = Rewritten environmental news article shown on position 5*
experimental_group_w2	*Group allocations for week 2; same logic as above but different allocations. Also note that there were personalisation sub-conditions*
personalisation	*Indicates which type of news personalisation was present during the 2nd week.* *Wed_ / Thu_ / Fri indicates the day where personalisaiton was introduced* *_edi / _imp_ / _exp indicates whether recommendations were editorial (same random selection procedure as in week 1) or based on implicit or explicit user preferences from the prior week / in-app survey 1*
Internal political Efficacy *"Please indicate to what extent you agree with the following statements on a scale from 1 (completely disagree) to 7 (completely agree)"* ***[Taken from Niemi et al., 1991]***	

pol_eff_1	*Sometimes politics and government seem so complicated that a person like me can't really understand what's going on."*
pol_eff_2	*"I feel that I have a pretty good understanding of the important political issues facing our country."*
pol_eff_3	*"I consider myself well-qualified to participate in politics."*
pol_eff_4	*"I feel that I could do as good a job in public office as most other people."*
pol_eff_5	*"I think that I am as well-informed about politics and government as most people."*
pol_eff	*Overall internal political efficacy scale*
Political interest ***[Taken from Moeller et al., 2020]***	
pol_int	*"Generally speaking, how interested are you in politics? Please answer on a scale from 1 (extremely disinterested) to 7 (extremely interested)""*
Political position ***[Taken from Moeller et al., 2020]***	
pol_pos	*"People generally talk about 'left-wing' and 'right-wing' politics when it comes to political ideas. Where would you place your ideas on a 10 point scale where 1 represents left and 10 represents right?"*
Issue importance	
iss_imp	*"How important do you consider issues around climate change and the environment to be? Please answer on a scale from 1 (extremely unimportant) to 7 (extremely important)."*
News app usage	
nws_app	*"Except for Google News and Apple News, do you currently have a news app installed on your phone?"*
News interest	

nws_int	*"How many days during the last week did you read at least one news article either online or offline?"*
Environmental news interest	
env_int	*"Generally speaking, how interested are you in news about climate change and the environment? Please answer on a scale from 1 (extremely disinterested) to 7 (extremely interested)"*
Environmental news consumption	
env_new	*"How many days during the last week did you read news about the environment?"*
Environmental discussions	
env_diss	*"How many days during the last week did you talk to friends or acquaintances about the environment?"*
News selection preferences *"Please indicate to what extent you agree with the following statements on a scale from 1 (completely disagree) to 7 (completely agree)"* Two separate sub constructs will be measured here: 1. Algorithmic preferences (alg_pre) 2. Journalistic preferences (jou_pre) ***[Taken from Scheffauer et al., 2023]***	
alg_pre_1	*"Having stories automatically selected for me on the basis of what I have consumed in the past is a good way to get the news,"*
alg_pre_2	*"News based on an algorithm does not limit my exposure to important news,"*
alg_pre_3	*"Relying on a news diet based on algorithms ensures that I receive news that is most relevant to me,"*
alg_pre_4	*"Having stories automatically selected for me on the basis of what my friends have consumed is a good way to get news,"*
alg_pre_5	*"News based on an algorithm of what my friends consume does not limit my exposure to important news"*
alg_pre	*Overall algorithmic preferences scale*

jou_pre_1	*“Having stories selected for me by editors and journalists is a good way to get news,”*
jou_pre_2	*“Editors and journalists know best what is most relevant to me,”*
jou_pre_3	*“News selection by editors and journalists ensure that I get exposed to important news*
jou_pre	*Overall journalistic preferences scale*
Topic interest *“For each of the following topics, please indicate how interested you are in receiving news about this topic on a scale from 1 (very disinterested) to 7 (very interested)”*	
top_int_1	*Business*
top_int_2	*Crime*
top_int_3	*Entertainment & Arts*
top_int_4	*Football*
top_int_5	*Health*
top_int_6	*Lifeandstyle*
top_int_7	*Politics*
top_int_8	*Science*
top_int_9	*Sport*
top_int_10	*Technology*
top_int_11	*UK news*
top_int_12	*World news*
top_int_13	*Environment*
News Information overload *“Please indicate to what extent you agree with the following statements on a scale from 1 (completely disagree) to 7 (completely agree)”* ***[Taken from Joris et al., 2023]***	

inf_ovr_1	*I often feel overwhelmed about the large amount of daily news*
inf_ovr_2	*I give up following the news due to the large amount of news*
inf_ovr_3	*I often feel that there is more news than I could process*
inf_ovr_4	*I often doubt whether I do not miss out the most important news of the day due to the large amount of news*
inf_ovr_5	*I often do not know where to start due to the large amount of news*
inf_ovr_6	*I often feel stressed about the large speed of news coverage*
inf_ovr	*Overall information overload scale*
Attitudes towards algorithms *"Please indicate to what extent you agree with the following statements on a scale from 1 (completely disagree) to 7 (completely agree)"* ***[Taken from Schepman & Rodway, 2023]***	
ata_pos_1	*For routine transactions, I would rather interact with an artificially intelligent system than with a human.*
ata_pos2	*Artificial Intelligence can provide new economic opportunities for this country.*
ata_pos_3	*Artificially intelligent systems can help people feel happier.*
ata_pos_4	*I am impressed by what Artificial Intelligence can do.*
ata_pos_5	*I am interested in using artificially intelligent systems in my daily life.*
ata_pos_6	*Artificial Intelligence can have positive impacts on people's wellbeing.*
ata_pos_7	*Artificial Intelligence is exciting.*
ata_pos_8	*An artificially intelligent agent would be better than an employee in many routine jobs.*
ata_pos_9	*There are many beneficial applications of Artificial Intelligence.*
ata_pos_10	*Artificially intelligent systems can perform better than humans.*

ata_pos_11	*Much of society will benefit from a future full of Artificial Intelligence.*
ata_pos_12	*I would like to use Artificial Intelligence in my own job.*
ata_pos	*Overall positive attitudes towards algorithms scale*
ata_neg_1	*Organisations use Artificial Intelligence unethically.*
ata_neg_2	*I think artificially intelligent systems make many errors.*
ata_neg_3	*I find Artificial Intelligence sinister.*
ata_neg_4	*Artificial Intelligence might take control of people.*
ata_neg_5	*I think Artificial Intelligence is dangerous.*
ata_neg_6	*I shiver with discomfort when I think about future uses of Artificial Intelligence.*
ata_neg_7	*People like me will suffer if Artificial Intelligence is used more and more.*
ata_neg_8	*Artificial Intelligence is used to spy on people.*
ata_neg	*Overall negative attitudes towards algorithms scale*
Diversity values *"Please rate the statements below in terms of importance. On a scale from 1(extremely unimportant) to 7 (extremely important), how important do you think it is for people to get..."* ***[Taken from Kim & Pasek, 2022]***	
div_val_1	*News from multiple viewpoints*
div_val_2	*News from multiple sources*
div_val_3	*News that balances information from every possible point of view*
div_val_4	*News that reflects the diverse viewpoints within our society*
div_val_5	*News from both liberal and conservative viewpoints*

div_val_6	*News from sources that are owned by different owners*
div_val_7	*News that pits different viewpoints against one another*
div_val	*Overall diversity values scale*
Dummy variables	
took_survey_week1	*Indicates whether a given respondent successfully completed the 1st in-app survey during the experiment*
took_survey_week1	*Indicates whether a given respondent successfully completed the 2nd in-app survey during the experiment*