MANCHESTER 1824
The University of Manchester

Using Data to Understand How Audiences Engage with Interactive Media

A thesis submitted to the University of Manchester for the degree of
Doctor of Philosophy
in the Faculty of Science and Engineering

2022

Jonathan Carlton
Department of Computer Science

Contents

Word Count: 40,988

List of figures

List of tables

Abstract

Media is evolving from traditional linear narratives to personalised experiences, where control over information (or how it is presented) is given to individual audience members. Measuring and understanding audience engagement with this media is important: a post-hoc understanding of how engaged audiences are with the content will help production teams learn from experiences and improve future productions. Engagement is typically measured by asking samples of users to self-report, which is time consuming and expensive. In some domains, however, interaction data have been used to infer engagement. The nature of interactive media facilitates a much richer set of interaction data than traditional media; this thesis aims to understand if these data can be used to understand and infer audience engagement and, by extension, the abandonment of content.

This thesis reports studies, run in collaboration with the BBC, of engagement and abandonment using data captured from audience interactions with an interactive TV show and an adaptive tutorial. It was found that engagement can be modelled and predicted in the interactive TV show, and that users appear to behave differently based on their level of engagement. For example, high engagement is associated with consumption-type behaviours, while low engagement is associated with skipping-type behaviours. When investigating the data collected from the adaptive tutorial, the results revealed that user context, rather than user interactions, affects the engagement of users. Abandonment was investigated using a wider dataset collected from the national release of the interactive TV show; it was demonstrated that abandonment could be accurately predicted from the interactions of users. An increase in moving backwards and forwards in the show were indicative of an increase in abandonment, suggesting an exploratory-type behaviour. When exploring the link between abandonment and engagement, it was found that low engagement users were predicted to drop out further from the end, suggesting a relationship between the two. The results demonstrate that interaction data is a viable method for the evaluation of media is this evolving domain.

To move towards consistency in the interaction data analysis field, the thesis proposes a framework to provide methodological support for researchers. Through an analysis of the literature, meta-issues were identified in the communication of research which create barriers in reproducibility and reduces transparency. The framework provides structure for those undertaking research on understanding users through their interactions and a terminology that can be applied consistently across the different disciplines in this area. It is conjectured that using such a framework should improve both the quality of science and science communication in the area, with more reproducible and transparent research being enabled.

Declaration of originality

I hereby confirm that no portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright statement

Acknowledgements

To my parents and brother, your patience and support throughout this journey has been inestimable – I could not have done it without you.

To my supervisors, Andy Brown, Caroline Jay, and John Keane, thank you for believing in me. Without your guidance, support, and friendship, this would not have been possible.

To Lukas, Miguel, and the two Tom's, you have made this journey unforgettable and I will always be indebted to you for your friendship, advice, and continuous support. To the IAM lab, thank you for the academic discussions which have helped shape the work and for the socials.

Finally, I would like to thank BBC R&D (and OBM team) and the EPSRC for supporting and sponsoring this project.

Chapter 1

Introduction

Interactive media is becoming more mainstream, driven by the shift to consuming content over the Internet. This delivery mechanism allows for individual audience members to receive different – personalised – content, where the information can be tailored to their individual needs, context, or knowledge, or to their input. In this context, a form of media that is neither linear video nor traditional computer game is re-emerging (Roth and Koenitz, 2019; Ursu et al., 2008). Differing from traditional media, where a single piece of content is produced that aims to engage and entertain the whole audience, these experiences aim to engage the individual through personalisation. With personalisation at the forefront of the experience, the audience may be able to get different information, levels of detail, or stories; information could be presented in different audiovisual forms, or a scene might be viewable from a different perspective. This is where the future of TV is heading. For example, Netflix are investing heavily in this area with the creation of several "interactive specials" (a dedicated section on their platform) such as "Black Mirror: Bandersnatch", 2018 and "You vs. Wild", 2019 which allow the viewer to make decisions that change the path through the narrative and affects the story presented to them.

The emergence of interactive media is born out of wanting to produce more engaging and enjoyable media content. While a one-size-fits-all piece of content (traditional media) can engage an audience, the individual characteristics and nuances that an audience bring are not considered, which both limits the engagement and may exclude a portion of the audience. With personalisation at the forefront, a larger proportion of the audience could be engaged as the content is tailored to the individual as per the methods previously mentioned. Further, excluded audience members, whether down to taste or content preference, may be included through tailoring to their context, needs, knowledge, or based on their input.

The BBC's approach to the creation of these experiences is to compose the programme from smaller objects (for example, "Object-Based Media", 2017). Each object contains some media that conveys a certain meaning (e.g., an audio or video snippet), and metadata describes how these can be arranged, opening a wide range of content creation possibilities that are tailored to the individual. This approach has been used to create a variety of experiences, including: an adaptive tutorial where the audience have control over the presentation and can move backwards and forwards between the steps (for example, "BBC: Make-along: Origami Jumping Frog", 2017); an interactive cookery show that adapts to the viewer's progress (for example, "BBC: Cook-Along Kitchen Experience", 2016); and a

documentary where preference questions are asked up-front and the content automatically alters to suit the individual (for example, "BBC: Instagramification", 2019). The BBC want to understand what benefits these types of experiences bring to audiences. They have therefore built a research platform that supports building, delivering, and testing experiences. One part of this research is understanding how success can be measured, and whether this can be done by analytical methods.

The types of experiences described above, which are a subset of personalised experiences, are the focus of this thesis. Behind the scenes, interactive media experiences are formed of narrative elements (pieces of content) and are linked together (forming paths through the experience). The links between narrative elements can have conditions associated with them, and may or may not be exposed to the audience as a choice. Narrative elements can have multiple representations (video or audio, for example), each of which gives broadly the same information, but in a different form. The personalisation in these experiences can be implicit (e.g., questions asked up-front to reorder the content) or interactive (e.g., opportunities for the audience to interact with the content).

The personalised nature of the media makes them more complex than traditional linear media to evaluate, as different audience members can have differently tailored experiences. To further conflate the challenge, and as the area is new, there is a lack of an established understanding and techniques that media producers can draw from to assess the success (or non-success) of experiences. Being able to measure the success of these experiences means that findings can feed forward into future productions, to understand what worked and what did not.

In this thesis, two aspects of success are investigated – engagement and abandonment, with higher engagement and lower abandonment rates being indicators for success for media creators. The interactive nature of these experiences presents an opportunity to leverage user interactions as a method to infer engagement and abandonment. Interaction data describes the actions a user has performed over time with an online system and takes a range of different forms, from sequences of actions (L. Guo et al., 2019; Zhuang et al., 2018) to statistical summaries of interaction sessions (Z. Liu et al., 2020; Youngmann and Yom-Tov, 2018), and has shown to be effective across a range of domains (Constantinides and Dowell, 2018; Kim et al., 2014; Lu et al., 2018; Zhuang et al., 2017). As it is an unobtrusive and scalable form of data, it allows for large scale experiments to be performed outside of the laboratory environment where the entire population can be sampled. When collected and analysed, the data be used to understand and model user behaviours (Grinberg, 2018; Lehmann et al., 2012), along with providing proxies for higher level user characteristics, namely engagement (Arapakis and Leiva, 2016; Barbieri et al., 2016) and abandonment (Diriye et al., 2012; Williams and Zitouni, 2017). The focus in this thesis is on the interaction data and whether it can be used to understand and model both engagement and abandonment.

The BBC provide the motivation, access to relevant systems, and data to explore research goals that suit their, as well as the university's, research objectives. The ethical collection

of data is considered throughout the work. At the start of each interactive experience, a privacy notice is presented to the audience explaining what data is collected and that it is being collected for research purposes; the audience are free to reject this. For University of Manchester systems, full ethical approval is sought from the ethics committee at the university.

1.1 Aims & Objectives

As this is an emerging domain of media creation, there is a desire by media producers to understand whether experiences were successful or not and whether audiences are interacting in a positive or negative way. As such, the following aim is addressed in this thesis:

To explore whether user engagement and abandonment can be inferred from interaction data

This thesis reports the findings on two aspects of success: whether audiences were engaged and whether they reached the end or not. To address the overall aim of the thesis, the following objectives are addressed:

1. To investigate whether there are signals in interaction data that are predictive of engagement

One measure of success is how engaged the audience were with an experience. Engagement is a quality of the user experience, characterised by the user's depth of investment when interacting with a digital system (H. O'Brien, 2016). Capturing a reliable measure of user engagement typically involves directly questioning the audience (O'Brien et al., 2018; H. L. O'Brien and Toms, 2010; Schoenau-Fog, 2011), which introduces limitations in its measurement – usually, a smaller and more controlled sample is collected in lab-based studies. Due to this, recent work has focused on deriving measures of engagement from scalable and unobtrusive approaches, notably user interactions, which are collected from the entire audience; this objective is directed towards understanding if these approaches can be used in this domain. We therefore propose the following research questions:

E-RQ1: Can user engagement with interactive media experiences be inferred from interaction data?

E-RQ2: What interaction-derived metrics are important when inferring engagement?

E-RQ3: Are there commonalities between interactive media experiences?

2. To investigate whether there are signals in interaction data that are predictive of abandonment

The abandonment of media content – where audience members dropout – can be measured by counting the number of users that reach the end as a proportion of the total audience.

While this type of metric can provide a broad overview of success – the higher the value, the better – it does not provide any detail, e.g., what led to someone leaving? Are the reasons content- or user-based? Answering these types of questions could provide fine-grained insight for media producers. Additionally, as the narratives in interactive media experiences can be more complex, with multiple sub-stories making up a broader narrative, some audience members may only be interested in one or two components, watch them in their entirety, and then leave. With the basic metric described previously, these types of users would be considered to have had a negative experience as they left early, but for media producers, this could be considered a success – the user has left after consuming a piece of media in its entirety, even if there are questions around engagement which could be answered by the previous objective. As such, we aim to explore modelling abandonment in interactive media experiences and investigate the following research questions:

A-RQ1: Can abandonment in interactive media experiences be modelled using interaction data?

A-RQ2: What interaction-derived metrics are important when inferring abandonment?

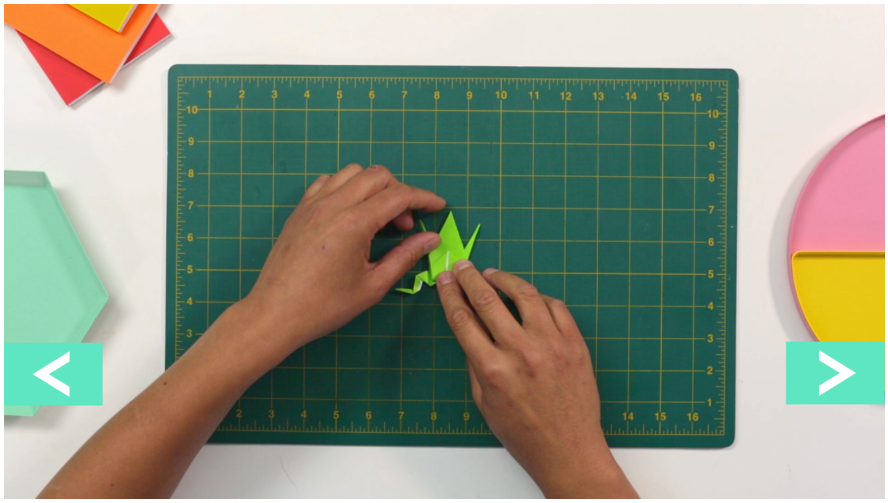
A-RQ3: Is there a relationship between abandonment and user engagement?

3. Based on the above, we will investigate whether a generic framework can be established to guide and assist researchers when working with interaction data

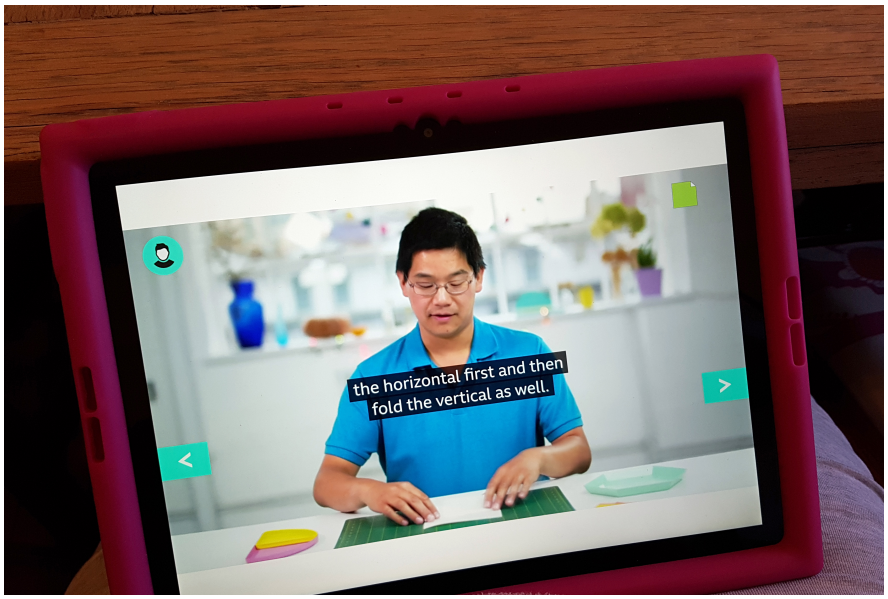
The work presented in this thesis to investigate the objectives and research questions described above borrows from and sits between two domains: machine learning and human-computer interaction. Using the approaches and techniques from these two domains, we are looking to understand the mapping between interaction data, user behaviours, and even high level user goals. In the carrying out the research reported in this thesis, it was identified that there is a need to be more rigorous in how ideas are discussed and tested. There are communication barriers that hinder reproducibility and reduce transparency. Therefore, a fresh analysis of the literature is performed to establish a generic framework that can provide methodological support when working with interaction data, enabling clear communication and facilitating the production of more reproducible and transparent research.

1.2 Interactive Media Experiences

The studies reported in this thesis focus on two interactive media experiences: an adaptive tutorial and an interactive branching narrative. The first is called *"BBC: Make-along: Origami Jumping Frog", 2017* (shown in Figure 1.1), which is a step-by-step guide showing the audience how to create an origami frog from scratch. The tutorial is formed of 27 steps (each a narrative element) organised linearly with each step building on the previous one. The audience have greater control over the presentation of the content than they would in a standard linear tutorial, along with the ability to progress through at their own



(a)



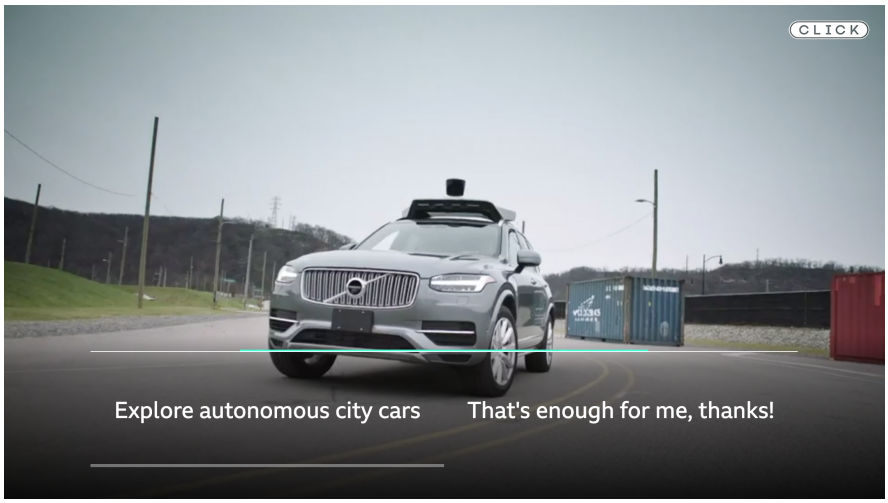
(b)

Figure 1.1. The Make-Along: Origami Frog Experience: (a) shows one of the steps in the tutorial with the next and back buttons displayed, while (b) shows an alternative camera angle during one of the steps in the tutorial, with both the next, back, and change camera angle buttons displayed.

speed and easily repeat steps. The default experience progresses as a standard tutorial, of the type found on YouTube or similar video platforms. However, after each step, the experience pauses showing a diagrammatic view of that fold; when the user has completed the step, they press the next button to move on. The audience can also move between narrative elements at any time using next and back buttons, while chapter icons allow the audience to jump to any step. Within the narrative elements, the user can manipulate and personalise the presentation of the content through controls on the interface, including changing camera angles for more detailed views and replying or rewinding the current step.

The second is called *"BBC: Click 1000", 2019* and was a one-thousandth special edition of a long-running BBC technology show (shown in Figure 1.2). This edition of Click was an interactive branching narrative where the audience determine their path through the story based on their interests. The show is made up of four sub-stories which each cover different topics; the two main sub-stories are about technology use in Malawi and autonomous vehi-

(a)



(b)

Figure 1.2. Examples of choices given to the audience in the interactive episode of Click: in (a) viewers can choose between a long or a short version, whilst in (b) they can either explore another aspect of a topic or move on.

cles. The audience can control what content they see, e.g., they can choose to view one or both use-cases for the autonomous vehicle technology (industrial and/or consumer) along with having the option to go into more or less detail about the technology. The on-screen host prompts the audience to make decisions (shown in Figure 1.2a), but if the audience decides not to interact, then a default path is automatically followed. The audience were also given next and back buttons to enable them to move to the next or previous segment of content (narrative element). Within a segment, the usual video controls could be used to move the video playhead or to replay the narrative element. In contrast to the Origami experience, viewers of Click had no ability to change the presentation.

1.3 Contributions

The main contributions of this work are:

It has established the usefulness of collecting and analysing user interactions from inter-

active media experiences

The usefulness of interaction data in providing a measure of success and an understanding of users is untested in interactive media, having found success in other domains. Two studies were carried out, one on an adaptive tutorial and another on an interactive TV show, where interaction data and self-reported engagement were collected. By training a model to predict levels of engagement for the interactive TV show and analysing how the model placed importance on interaction metrics derived from user interactions, it was found that engagement levels could be accurately predicted and there were signals in the data that pointed towards behaviours. For example, that high engagement is associated with consumption-type behaviours, while low engagement is associated with skipping-type behaviours. When investigating the data collected from the adaptive tutorial, engagement (and the four factors that make up engagement) could be accurately predicted but the models relied on single metrics to make the predictions, limiting their usefulness. However, by performing this analysis, the results revealed that user context is important when considering engagement in this setting.

Following the studies of engagement, an exploration of the wider dataset collected from the interactive TV show was performed to discover if there were predictive signals for abandonment in the interactions of users. It was demonstrated that abandonment could be predicted, with temporal metrics being important in its prediction, that users exhibit an exploratory-type behaviour, and that there was a link to engagement. More specifically, the proportion of events triggered when moving backwards and forwards in the show were indicative of an increased likelihood of abandonment. In exploring the link between abandonment and engagement, there was a significant difference in the distance prediction between low and high engagement users, with higher values (dropping out further from the end) being predicted for low engagement users.

The monitoring of the metrics that make up these behavioural proxies could provide media producers with the ability to understand two aspects of success from the data alone, feeding into the creative process or to react to a change in the user behaviour in real-time. All of which establishes that interaction data is a viable method for understanding users and assessing the success of content in this re-emerging domain.

A framework to provide methodological support for researchers in the analysis of interaction data to understand users

Through a fresh review of the literature to further explore issues identified in the original literature review, issues in the communication of research were found that create barriers in the reproducibility of the work and reduce transparency. A framework was proposed to standardise reporting of the different types of data and processes encountered in these domains. It provides researchers with methodological support and can help them structure their research. The framework consists of the following four layers (each building on the last; presented in full in Chapter 5):

Interaction Data: The data collected from systems and/or generated by the system itself.

Interaction Metrics: An abstraction from the interaction data and the analytics applied.

Behavioural Proxies: Semantically meaningful and observable patterns of action or groupings of interaction metrics.

Outcome of Interest: The specific characteristic that is being measure, investigated, or attempted to be understood.

The framework provides structure for those undertaking research on understanding users through their interactions and, as it is a multi-disciplinary area with researchers from a range of domains, a consistent terminology for stakeholders. In addition, the framework facilitates open and reproducible research in the interaction data analysis field, improving the quality of science and science communication.

1.4 Publications & Open-Sourcing

Publications

[MM'21] J. Carlton, A. Brown, C. Jay, and J. Keane. "Using Interaction Data to Predict Engagement with Interactive Media." In *Proceedings of the 29th ACM International Conference on Multimedia (MM'21)*, 2021.

[CHI'19] J. Carlton, A. Brown, C. Jay, and J. Keane. "Inferring User Engagement from Interaction Data." In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI'19 Extended Abstracts)*, pp. 1-6. 2019.

[DSJM'18] J. Carlton, J. Woodcock, A. Brown, C. Jay, and J. Keane. "Identifying Latent Indicators of Technical Difficulties from Interaction Data." In *ACM KDD Workshop on Data Science, Journalism and Media (DSJM)*, 2018.

[MB'18] J. Carlton, A. Brown, J. Keane, and C. Jay. "Using Low-Level Interaction Data to Explore User Behaviour in Interactive-Media Experiences." *Measuring Behavior 2018*, 2018.

Open-Sourcing

To ensure the work presented in this thesis is reproducible, and to contribute to the open-research space, there has been an additional objective to open-source as much of the data and materials as possible. Working alongside the BBC, the following open-source contributions have been made:

- J. Carlton, A. Brown, C. Jay, and J. Keane. (2022). Code and Data – User Abandonment in Interactive Media (v1.0-thesis). Zenodo. https://doi.org/10.5281/zenodo.5879067

- J. Carlton, A. Brown, C. Jay, and J. Keane. (2022). Interaction-lib: A library for the processing of interaction data collected from interactive media (v1.0-thesis). Zenodo. https://doi.org/10.5281/zenodo.5879129
- J. Carlton, A. Brown, C. Jay, and J. Keane. (2022). Chaos in the Clicks: Towards a Framework for the Analysis of Interaction Data (v1.0-thesis). Zenodo. https://doi.org/10.5281/zenodo.5883794
- J. Carlton, A. Brown, C. Jay, and J. Keane. (2021). Using Interaction Data to Predict Engagement with Interactive Media (Multimedia'21 Paper Analysis Code) (v1.0.0). 29th ACM International Conference on Multimedia (MM). Zenodo. https://doi.org/10.5281/zenodo.5137806
- J. Carlton, A. Brown, C. Jay, and J. Keane. (2021). Code and Data – User Engagement with Make-Along: Origami Frog (v1.0). Zenodo. https://doi.org/10.5281/zenodo.5783522

1.5 Structure

The thesis is structured as follows: Chapter 2 provides an overview of relevant related work. Chapter 3 presents two studies investigating engagement with two different interactive media experiences, and Chapter 4 presents the study of abandonment in interactive media experiences. The framework to provide methodological support for those working with interaction data to understand users is in Chapter 5. Chapter 6 discusses the findings in the thesis and Chapter 7 draws conclusions and considers future work.

Chapter 2

Literature Review

Interactive media experiences are an emerging form of media ("Object-Based Media", 2017), where the content is tailored to the individual. As the domain is developing, there is a desire from media producers to understand the audience's experiences, assess the success of the content, and then feedback into the creative process; for example, what presentation style does or does not work. At the moment, there is a lack of existing evaluation approaches and techniques that media producers can draw on to understand the audience's experience beyond relatively basic methods, such as how many people viewed the content. The increase in interactivity of the media presents a chance to collect the interactions of users to facilitate an evaluation of their experience. Here, the literature around the use of interaction data to understand user experience is explored, focusing on two aspects of success of value to media producers: user engagement and abandonment.

The chapter is structured as follows: an overview of engagement and how it can be measured is presented in Section 2.1; Section 2.1.3 provides an overview of interaction data; approaches that make the link between phenomena – engagement and abandonment – and interaction data through analytical means are discussed in Section 2.2; and, a summary of how the literature informs the thesis investigation and a discussion of issues discovered while reviewing the literature is presented in Section 2.3.

2.1 User Engagement

User engagement is one aspect of success that media producers wish to understand, with higher engagement signalling a successful experience for the user. User Engagement is a complex, multifaceted phenomenon (Attfield et al., 2011), with H. O'Brien broadly defining it as a quality of user experience characterised by the depth of an user's investment when interacting with a digital system.

As user engagement is often a phenomenon that is of interest to the research community, several methods to measure it have been proposed. These methods are composed of three main groups (Attfield et al., 2011; Hong and Lalmas, 2020; Lalmas et al., 2014; Lehmann et al., 2012; Webster and Ho, 1997):

(1) Self-reported methods.

(2) Physiological measurements.

(3) Online analytics.

2.1.1 Self-reported Methods

In group (1), a measure of engagement is captured through subjective and mostly qualitative methods, such as questionnaires and interviews. A popular and frequently used survey is the User Engagement Scale (UES) (O'Brien et al., 2018; H. L. O'Brien and Toms, 2010) which measures engagement with digital technology and consists of either 12 (short-form) or 30 (long-form) questions depending on the intended application. In the most recent edition of the survey, engagement is captured through a combination of four factors:

Focused Attention: Feeling absorbed in the interaction and losing track of time.

Perceived Usability: A negative affect experienced as a result of the interaction and the degree of control and effort expended.

Aesthetic Appeal: The attractiveness and visual appeal of the interface.

Reward: The value gained from using the system.

The following are sample questions from long-form UES and are adapted to the system being evaluated by replacing *Application X* with the name of the system: "I lost myself in this experience" is one of the questions (seven in total) for the focused attention sub-scale; "I found this *Application X* confusing to use" is one of eight perceived usability questions; "This *Application X* was attractive" is one of five for the aesthetic appeal sub-scale; and, "My experience was rewarding" is one of ten reward sub-scale questions. An alternative is the Engagement Sampling Questionnaire (Schoenau-Fog, 2011) which provides the means of regularly sampling the participant throughout a study to capture how engagement changes over time. The approach is broken down into four smaller surveys and applied at particular points during the study: demographics, before the experience, during the experience, and after the experience, with the *during the experience* survey repeatedly applied at various points during the study.

2.1.2 Physiological Measurements

In group (2), more objective and mostly quantitative measures of engagement are captured through physiological measurements (Doherty and Doherty, 2018), such as electrocardiography (Arapakis et al., 2019) and eye-tracking (Nakano and Ishii, 2010). For example, in (Belle et al., 2011) ECG and other physiological data were collected from participants in a study to measure and assess the attention and engagement levels while watching interesting and non-interesting videos. The interesting (highly engaging) videos consisted of documentary clips, popular movie scenes, and high-speed car chases, while the non-interesting

videos were repetitive and monotonous animations (a clock ticking and still images). During the study, participant feedback – whether they found the videos engaging or not – was collected with each viewing two 20-minute sets of three- to five-minute long clips. The authors found that they could accurately predict high and low engagement using features derived from the physiological data. These types of approaches (both (1) and (2)), however, are impractical to use at scale and are typically deployed in smaller scale user studies within a laboratory environment.

2.1.3 Online Analytics

While providing a reliable measure of engagement, self-reporting methods such as those described previously have their limitations. Notably, they can become impractical to use at scale and in-the-wild, which limits their application in large scale media productions. Additionally, even with physiological measures demonstrating their ability as effective proxies for engagement, they are hampered with similar, but more acute, impracticalities as self-reporting measures.

Group (3) is associated with online analytics, where more objective, quantitative, and large-scale measures are used – addressing the impracticalities of (1) and (2). Interaction data is the data source in online analytics; it is straightforward to capture at scale, without disruption, and the burden of audience members to retrospectively self-report engagement is removed (**paper_54**); Attfield et al., 2011; Lehmann et al., 2012. The data can take a range of different forms, depending on the collection purpose, for example, application-specific actions (Ge et al., 2020; Z. Liu et al., 2020; H. Zhang et al., 2020), mouse cursor events (Arapakis et al., 2019; Lagun and Lalmas, 2016; Zhuang et al., 2017), or mobile interactions (Constantinides and Dowell, 2018; Lu et al., 2019).

However, as engagement and abandonment is being measured indirectly through the interactions, there is greater potential for error in their inference; further, when the user is aware their interactions are being collected, their behaviours may change (Doherty and Doherty, 2018). The interaction data captured from the interactive media experiences reported in this thesis take the form of application-specific actions, for example, clicking the play/pause button or changing window orientation.

2.2 From Data to Phenmena

As measuring engagement and abandonment with interactive media experiences is a new area, learning from approaches used in domains where understanding user behaviour through their interactions are more established is key. In particular, examining the types of data collected, the behaviours inferred, and what analysis techniques were applied.

Extracting behaviours and high-level user characteristics from interaction data has been performed in a range of contexts from security and medical systems (Chien et al., 2020;

Shen et al., 2020) to optimising search engine results (Q. Guo et al., 2011; Huang et al., 2011). For example, interaction data has been captured and analysed to identify possible cognitive impairment (Gledson et al., 2016), to help in identifying difficulties in navigating online media (Thomas, 2014), and pointers towards personality traits of users completing a visual search task through the synthesis of interaction (Brown et al., 2014).

However, interaction data in its raw form is typically too fine-grained to make the link between data and behaviour, so authors often extract interaction metrics from their data. These typically take the form of descriptive statistics, such as the number of page visits, dwell time, or the number of clicks (Hong and Lalmas, 2020; Lehmann et al., 2012). From these metrics, users' behaviours can be inferred and used as proxies for the phenomenon of interest (Barbieri et al., 2016; Kim et al., 2014; Miroglio et al., 2018; Su et al., 2018). For example, C. Liu et al., 2010, captured two-weeks of interaction data using a web browser plugin which recorded the searches and browsed pages for users that opted-in. Using the visit time to consecutive pages in a browsing session, the authors calculated the dwell time for each page visit. The authors find through fitting a Weibull distribution that users adopt a "screen-and-glean" behaviour where they first check a web page prior to a more detailed examination.

To determine how familiar a user is with a website, Apaolaza et al., 2015, collected cursor movements from users and extracted metrics which described different types of scrolling behaviours. Doherty and Doherty, 2018, find that a document recommendation panel on an online word processing platform helps users find their documents faster but that some miss the recommendations and resort to complicated alternative methods to find the document. The study demonstrates that an evaluation of a system can be performed at scale and without interruption using user interactions, a direction in which media producers want to move towards using interaction data and one that this thesis aims to lay the foundation for.

The possibility of further personalisation of content through the monitoring of interactions for behavioural proxies is demonstrated in (Constantinides and Dowell, 2018). The authors find that user news reading behaviours, such as reading frequency, browsing strategy to select articles, and reading style (detailed, skimming, or scanning), could be extracted and inferred from interaction data. Further, that the behaviours could be matched with user preferences for different interface elements or interactions, for example, additional features for users that track and follow the news throughout the day.

A trend that is noticeable in the literature is the inconsistent use of the term 'behaviour'. For example, as previously discussed, C. Liu et al., 2010, found through their analysis and subsequent inference a "screen-and-glean" behaviour – which is more akin to an actual behaviour – whilst works such as (Belletti et al., 2019; Chen et al., 2018) use the term to refer directly to their interaction data – that is, that the metrics, whether statistics or sequences, extracted are described as behaviours.

2.2.1 User Engagement

A measure of success is how engaged the audience was with an experience. As discussed previously, capturing a reliable measure of engagement typically involves directly questioning the audience, introducing limitations in its measurement. An aim of this thesis is to investigate whether there are predictive signals for engagement in interaction data, with three research questions on whether engagement can be inferred (**E-RQ1**), what interaction metrics are important in its prediction (**E-RQ2**), and whether there are commonalities between interactive experiences (**E-RQ3**). Here, focusing on **E-RQ1** and **E-RQ2**, the existing literature around deriving measures of engagement from interaction data is discussed.

Lehmann et al., 2012, logged user interactions through a web browser toolbar from approximately 2 million consenting users visiting a wide range of websites such as news, weather, and movies. From the interaction data, the authors extracted metrics such as those described previously: the number of page visits, dwell time, and number of clicks. Using the metrics, simple models of engagement were trained that reflected different aspects of engagement: popularity (represented by the number of users), activity (dwell time), and loyalty (the return rate of users), where the higher and more frequent, the more engaged. Dwell time is an often-used simple proxy measure for engagement (Agichtein et al., 2019; Yom-Tov et al., 2013), for example, dwell time after clicking on an advertisement is modelled in (Barbieri et al., 2016) to improve the ranking of advertisements in a search context. The authors find that engagement – measured by dwell time – increases when the model is integrated into the ranking function used to serve the advertisements.

However, there is a limited amount of depth that can be gleaned from dwell time alone to measure engagement and other traits (Dupret and Lalmas, 2013; Q. Guo and Agichtein, 2012). Using this as a motivator, Lagun and Lalmas, 2016, reconstructed reading patterns from scrolling events and temporal statistics to develop a more nuanced understanding of user attention beyond dwell time in the context of online news reading. Articles were split into areas-of-interest (top, middle, bottom, comment, start, and leave) and scrolling actions were modelled as transitions between these states in a Markov chain model. By applying a Markov mixture model and comparing statistics, clusters of reading behaviours were identified which the authors characterised as: *bounce, shallow, deep*, and *complete*. Between the behaviours, significant differences in the interaction metrics were found; for example, users determined to be in the *completed* group spent significant time in the comment section and recorded an increase in the number of clicks on an article compared with the other three detected behaviours. In a similar approach, Grinberg, 2018, extracted dwell time, maximal reading depth, active engagement (the amount of interaction), the proportion of an article visible on screen, scrolling speed, and normalised active engagement from users interacting with news articles. The data was collected from an online web analytics company and consisted of 66,821 news articles which were viewed a total of 7.7 million times by 4 million unique users. Using the metrics, a multivariate normal mixture model was trained to identify clusters in the data. Through a combination of interpretation and statistical comparison

between the identified clusters, six reading behaviours were found (also termed engagement measures): bounce back, shallow, scan, idle, read, and read long. The authors then investigated how information gain in the article text relates to each behaviour. To understand engagement with online news content, Arapakis et al., 2014, extracted a large set of metrics from cursor movement data, including time, distance, speed, and acceleration metrics. They found certain types of mouse gestures were negatively correlated with engagement, specifically the UES focused attention sub-scale (previously defined in 2.1). The authors, however, do not reveal what combinations of interaction metrics determined the clusters.

To make the link between interaction data and the phenomena in question, authors often capture a notion of ground truth through the application of surveys, for example, by using the UES survey previously described. In the search and information retrieval domains the focus is often user satisfaction (analogous to user engagement in a search context (Hong and Lalmas, 2020)). Using the novelty sub-scale of the original UES survey (see H. L. O'Brien and Toms, 2010), Zhuang et al., 2018, investigated the relationship between user behaviour sequences and perceived novelty in a web browsing context. The authors carried out a study where participants performed tasks on a web browser, with their interaction data (mouse clicks, scrolling, and keystrokes) logged and a post-study questionnaire about their experience was administered. The interactions were distilled into descriptive actions, for example, clicking a result on the search engine result page and clicking next for the next set of search results. The authors then identified frequent sub-sequences in the data and performed a chi-square test between the frequent sub-sequences and levels of novelty (high and low). They found that several sub-sequences could discriminate between novelty levels, for example, the most discriminatory was: clicking next for the next set of search results followed by clicking on a result. The authors hypothesised that this demonstrated a switch from exploration to immersion. Subsequently, the authors trained a model using the non-/presence of the top 15 most discriminatory sub-sequences, finding that the model could accurately predict high and low novelty.

UES sub-scales were used as separate prediction targets to investigate behaviour during a search task in (Zhuang et al., 2017). Data was collected from participants in a controlled study, where each participant was asked to carry out a series of search-based tasks with their interactions collected throughout – the specific interactions logged were not described. A range of behavioural features were extracted based on four categories: click (described clicking behaviour, e.g., number of clicks), query (related to the queries issued, e.g., number of queries), result (features related to the results viewed by the user, e.g., number of pages viewed), and time (e.g., the total time on the task). Using the behavioural features, models for each sub-scale were trained and the mean decrease in accuracy was used to identify important features. The authors find the time spent on the search engine result page is predictive of usability and they hypothesise that it indicates that the longer spent on the result page (searching) is indicative of a lack of usability (struggling to find information). In a similar approach, attention, usefulness, and perceived task duration were predicted from mouse cursor data by (Arapakis and Leiva, 2016). Data was collected from a crowd-

sourcing study where participants were asked to carry out two search tasks. Their mouse cursor data was logged throughout and self-reported measures for attention, usefulness, and perceived task duration were captured in a 3-question survey (one question per measure; sourced from UES). A large number of features were derived from the mouse cursor data and used to test whether each self-reported measure could be predicted. The authors find that each could be accurately predicted, but links between the metrics and measures were not discussed or uncovered.

User preference for news content and topics were collected from participants during a study investigating how the quality of news affects the probability of clicking on news articles and their reading behaviours (Lu et al., 2019). Interaction metrics were derived from interaction data to describe the participants session, specifically: viewport time (time spent reading), dwell time, reading ratio, reading speed (combination of dwell time and reading length), scroll direction change times, and number of intervals (careful examinations of information). Through statistical analysis of the metrics, the authors found that when users read low quality (determined by expert annotation) news, they spend less time reading, leave articles earlier, read slower, have fewer revisits, and fewer careful examinations. In a similar study, Lu et al., 2018, extracted temporal, reading, and scrolling metrics from user interactions with news articles. By applying correlation analysis between user groups, which were determined by answers to a news reading preference survey, they found users who liked an article tended to read more, at a slower pace, and revisited content. Both works demonstrate findings that can be used by online news producers to evaluate their stories to better understand whether they are serving the needs of their readerships.

In a different context, specifically users searching for music, A. Li et al., 2019, distinguishes between focused (where the user is looking for one thing in particular) and non-focused (where the user is open to different results) search mindsets to improve design choices to boost user experience. The authors used a mixed methods approach where responses to a survey were collected to determine user mindset and interaction data was collected and analysed to predict mindset. They found that focused users expend more effort through issuing longer search queries and spend longer both searching and until their first click, scroll down further, and click on lower-ranked entities. In contrast, non-focused users tend to put in less effort and instead rely on the music platform for search results and suggestions. Continuing in the same context, evaluating music recommendations based on user satisfaction and intent is performed in (Mehrotra et al., 2019). Following a mixed methods approach of capturing interviews, an in-app survey on intent and satisfaction, and the collection of user interaction metrics, the authors identify eight user intents and demonstrate the importance in explicitly considering the intents when predicting user satisfaction. In each of the intents, different interaction metrics were important. For example, one intent is to explore artists or albums more deeply; for this the most important metrics were the total number of clicks on the homepage, whether the user saved or downloaded any track or album, and the average value of a click (a metric to describe the relationship between the number of clicks and the number of music streams).

2.2.2 Abandonment

Alongside user engagement, another measure of success of interest to media creators is abandonment. One aim of this thesis is to investigate whether there are signals in interaction data that are predictive of abandonment. Similar to the investigation of user engagement, three research questions are proposed on whether abandonment can be modelled using interaction data (**A-RQ1**), what interaction metrics are important when inferring abandonment (**A-RQ2**), and whether there is a relationship between abandonment and engagement (**A-RQ3**). Answering these research questions could provide detail to media creators to aid in answering questions such as: what led to someone leaving? Are the reasons content- or user-based? Here, the existing literature on abandonment in a range of different contexts is examined.

In the context of interactive media experiences, abandonment is where audience members dropout and stop watching the content before it has finished. As shown previously, the information that can be gleaned from the interactions of users can provide rich insights in various domains and settings when investigating user engagement and satisfaction. As an extension to understanding satisfaction, work has been carried out on understanding and predicting abandonment within the context of search. Abandonment in that domain is where the user does not click any search result and leaves the search engine result page, resulting in a negative signal and possible user dissatisfaction (J. Li et al., 2009). To provide more detail than can be revealed by the click-through rate, (Das Sarma et al., 2008). defined a bypass rate, which quantified the number of search results ignored by the user before clicking on a search result, providing a signal for results that were not of use to the user. Extending this, J. Li et al., 2009, showed that a large proportion of queries performed by users result in good abandonment, where the user finds what they are looking for, rather than bad abandonment (they do not).

The relationship between satisfaction and abandonment was further explored by Diriye et al., 2012, where the rationales behind users' abandoning were explored. Through an in-situ study, they detect when a user was about to abandon (either by users not clicking on results or by trigger events, for example, closing the tab) and ask the users to provide reasons for abandoning. The users were able to say that they either found what they were looking for (they were satisfied with the result), that they were dissatisfied with the results, or there was some other reason for abandoning (a distraction, for example). They found that metrics derived from cursor data can accurately predict dissatisfaction, specifically the speed in which the user clicked on a search result. Following a similar approach but focusing more on developing a better model for abandonment, Song et al., 2014, explores both query-level and session-level user behaviours and their relationship to abandoned queries. For example, users that abandon (but were satisfied; *good abandonment*) tend to spend less effort in sessions compared to those who abandoned (but were not satisfied; *bad abandonment*), in addition to longer session length being associated with bad abandonment.

Williams et al., 2016, investigate signals in mobile gestures to detect user satisfaction and

use it to differentiate between good and bad abandonment in a mobile search context. A range of interaction metrics were extracted, such as the total number of swipe actions and the number of up and down swipes, along with more descriptive statistics including dwell time and session duration. They found, through correlation analysis, that the time spent interacting with answers on a search result page were positively correlated with satisfaction and good abandonment, while swipe interactions were negatively correlated with satisfaction. Expanding on the prior work, Williams and Zitouni, 2017, investigated how sequences of user interactions differ between good and bad abandonment. By training a recurrent-neural network and by collecting labelled data through crowd-sourced annotation, they found that queries which exhibit bad abandonment lead to more interactions as users spent more time searching for information, while good abandonment had shorter sequences - dwell and scrolling actions were more common in bad abandonment. Proposed as an alternative but supplementary method to predict good and bad abandonment, mouse cursor data is used and modelled also using a recurrent-neural network in (Brückner et al., 2020). The authors explore various architecture and data sampling methods, finding that they can predict abandonment with reasonable accuracy but do not explore what features of the data contribute to the prediction.

The abandonment of online content has been of interest in domains outside of the search community, for example, in online tutorials and courses (Halawa et al., 2014; Ramesh et al., 2013; Taylor et al., 2014). Yan et al., 2017, explore the feasibility of predicting which learners – in an online programming tutorial – are likely to complete the next lesson in the tutorial. They examine whether metrics derived from the interactions are either positively or negatively correlated with abandonment and find that users who saved their progress and seek further instruction from the in-tutorial help were less likely to leave. From modelling abandonment, the authors found that metrics relating to commitment and effort were relied on for prediction.

In some works, the focus is on predicting how long a user will stay on an online service. For example, Gupta and Maji, 2020, model session length in e-commerce search and Tian et al., 2021, predicted the next application that someone will use and how long they will use it for. Vasiloudis et al., 2017, used survival analysis techniques to model session length. They found that almost half of the users in their dataset (consisting of more than four million streaming sessions) demonstrated negative ageing which is where their session becomes less likely to end as it increases in time. The authors derive a set of metrics from their interaction data but the data – what the events are and their composition – are not described, and only a handful of metrics are presented (with the remaining omitted). Through fitting a model on their metrics, the authors find that they could accurately predict the length of a session using metrics available at the start of a session. Similarly, in (Dedieu et al., 2018) the authors focus on proposing a new model to predict session length from the moment a user logs into the platform using a Bayesian inspired approach. By extracting feature importance directly from their model, they find that device and time-related metrics, specifically absence time and the average time the user spends on the service, to be the most

important ones.

2.3 Summary

This review describes the context in which the work presented in this thesis sits. The review looked at the different methods to measuring engagement, finding that they fit into one of three groups: self-reported, physiological, and online analytics. In addition, a review of the literature on abandonment was also performed.

Qualitative methods provide a more subjective measure of engagement and produce metrics which are closer-to-ground-truth, for example, the User Engagement Scale (O'Brien et al., 2018; H. L. O'Brien and Toms, 2010). There are difficulties, however, with applying these methods at scale and without disrupting the user experience. This has led to a range of approaches that make the link between the interactions of users – a type of data that can be unobtrusively captured at scale – and user engagement.

One aim of the thesis is to investigate whether there are predictive signals for engagement in interaction data, which is broken down into three research questions. For **E-RQ1**, which focuses on the extent to which engagement can be predicted in interactive media experiences, the literature demonstrates that engagement can be predicted from interaction in other domains. In some cases, proxies for engagement are predicted such as dwell time, but in others, a measure of engagement gathered from self-reported approaches has been used as a prediction target. As part of the modelling or analysis process, relationships between the interaction data and measure of engagement are investigated which allow for inference to be made about the behaviours of users – the focus of **E-RQ2**. The literature from other domains suggests this is a worthwhile approach to explore. In a number of works, the interactions of users are collected, for example, application-specific events or mouse cursor data. These data are then processed into another form such as statistics or sequences, with statistics consisting of contextualised and summative descriptive metrics. From there, the metrics are used as input into models or statistical tests to predict or understand engagement, with authors finding behavioural differences between engagement levels.

The abandonment of content is another measure of success that media producers are interested in understanding, with lower rates of abandonment being preferential. In reviewing the literature, it was found that most progress in the modelling of abandonment has been achieved in the search domain. The measure of abandonment differs from that of engagement as it can typically be represented by a binary – for example whether or not a user left the search result without clicking. There has been work on gathering a more nuanced measure of abandonment where it is broken down into good and bad abandonment. This review of the literature has found that there is evidence that abandonment – in different forms – can be accurately predicted using the interactions of users (**A-RQ1**). Similar to engagement, approaches for doing so focus on extracting descriptive metrics or sequences from the data and then using that abstraction as input into models or statistical tests. Inference is made

from models or the results of statistical tests to understand what behaviours are important (**A-RQ2**). Key findings are that as the length of an interaction session increases, the less likely it is to end (Vasiloudis et al., 2017) or that users are unlikely to abandon if making use of the features available to them on the platform (Yan et al., 2017). The relationship between engagement and abandonment (**A-RQ3**) has been explored in a search context. Often in the literature, and much like in the measurement of engagement, a form of ground truth for satisfaction is used to provide a level of reliability in the investigation of abandonment. As such, making use of the engagement measure captured in investigating the prior aim will be crucial in discovering links between the two.

The literature review shows that modelling the engagement of users, or their rates of abandonment, through their interactions with digital content works in other domains – the gap this thesis aims to address is whether it applies to interactive media. In the synthesis of the literature, there were noticeable inconsistencies in the use of the term 'behaviour', which introduced challenges in forming a coherent picture. Highlighted earlier in the text, some authors refer to meaningful groupings of interactions as behaviours while in others, the data itself is referred to as a behaviour. These discrepancies present an opportunity, and the final aim of this thesis investigates this in more depth. In the next chapter, two studies of user engagement in two different interactive media experiences are presented, addressing the main aim of the thesis.

Chapter 3

User Engagement in Interactive Media Experiences

A key metric for success is how engaged the audience members were with the experience. However, user engagement is a complex, multifaceted phenomenon (Attfield et al., 2011) and as media is typically consumed by a geographically distributed audience, in different contexts and on different devices, an accurate measurement has both practical and methodological challenges. As shown in the review of the literature, interaction data can be used as an unobtrusive and scalable way of monitoring user behaviours and engagement, and given that interactive media experiences offer additional interaction opportunities to the audience than traditional media, is it possible to infer engagement from the monitoring of interactions?

The retrospective analysis of interactions to understand engagement can not only be fed into the creative process to improve future media, but user interactions could allow near real-time monitoring of engagement and enable the content to adapt on-the-fly. For example, audience members who appear highly engaged may be offered supplementary material, while those who appear to have low levels of engagement might be given a shorter or simpler conclusion to the content. These types of interventions could significantly enhance the user experience, but are demanding and are only really possible if engagement can be monitored in real-time. Firstly, however, a retrospective understanding of engagement needs to be established.

This chapter aims *to investigate whether there are signals in interaction data that are predictive of engagement*. To direct the investigation, the following research questions are explored:

E-RQ1: Can user engagement with interactive media experiences be inferred from interaction data?

E-RQ2: What interaction-derived metrics are important when inferring engagement?

E-RQ3: Are there commonalities between interactive media experiences?

To answer the questions and address the aim, the chapter presents two studies of two separate interactive media experiences: one an adaptive tutorial and the other an interactive

branching narrative. Both studies capture a ground truth notion of engagement from audience members, with the former in a more controlled setting and the latter an in-the-wild production environment.

3.1 Make-Along: Origami Frog

The first study into engagement with interactive media experiences is focused on an adaptive tutorial experience called *Make-Along: Origami Frog*, which is a step-by-step guide showing the audience how to create an origami frog from scratch – shown in Figure 1.1. The tutorial is broken down into 27 steps organised linearly with each step building on the previous one. The audience have greater control over the presentation of the content than they would in a standard linear tutorial, along with the ability to progress through at their own speed and easily repeat steps. The default experience progresses as a standard tutorial, of the type found on YouTube. However, the audience can move between narrative elements using next and back buttons, along with chapter icons that allow the audience to jump to any step. Within the narrative elements, the user can manipulate and personalise the presentation of the content through controls on the interface, including changing camera angles for more detailed views and replying or rewinding the current step. The standard video controls are also available: play/pause, fullscreen, volume control, and video scrubbing.

3.1.1 Methodology

Study Design

In the first step towards understanding engagement with interactive media experiences, a lab-based study was carried out where interaction data and engagement metrics were collected from Make-Along: Origami Frog. Building on the initial results presented in **[CHI'19]**, in which the same study was carried out but with a smaller set of participants where preliminary results suggested differences in user interactions based on their engagement. The study presented here was designed to be as ecologically valid as possible, with participants taking part remotely and free to complete it at their leisure with any device. Participants were first instructed to read an information sheet, detailing both the study and data being collected, then to complete a consent form; the data for those that did not complete the consent form were discarded. To ensure that all study components functioned correctly and that instructions were clear, two separate pilot studies were run. In each, two participants performed all study tasks, and their feedback allowed for refinement of the process. To recruit participants, advertisements were posted on internal and external message boards within the university, public advertising space, and through word-of-mouth; an incentive for participation was entrance into a random prize draw to win a gift voucher. For the components of the study that were hosted by University of Manchester systems, i.e., the surveys, ethical approval granted by the University of Manchester Research Ethics Committee. A privacy

notice was presented to the participants when they first started the adaptive tutorial (hosted by the BBC), which detailed the types of data collected during the experience, with acceptance being required to progress.

Data Collection

To capture information about the sample of participants taking part, the participants were asked prior to starting to complete a demographic survey. The survey contained seven questions on: the participant's age, gender identity, level of education, current employment status, competence with technology, previous origami experience, and their preferred method of consuming video-based content. To quantify engagement, the User Engagement Scale (UES; long-form; 30 questions) was administered after the experience. The survey was chosen as it has been validated in a range of other contexts and is widely used for measuring engagement (Doherty and Doherty, 2018). Following the guidance of the authors of UES (O'Brien et al., 2018), engagement scores were calculated as the mean of the factor means. Interaction data was captured in the study by built-in analytics and in the form of application-level events, these were logged whenever a user performed an action on the interface. The data takes the following form: *user_id* - an anonymous identification string, *timestamp* - millisecond granularity timestamp, *action_type* - the type of event that occurred, *action_name* - the button clicked/context change, and *data* - additional metadata about the event. The analytics captured the following events: switching camera angles, play/pause, next and back, subtitles, volume changes, video scrubs, repeating the step, fullscreen, changing chapter, and narrative element changes.

Interaction Metrics

In its raw form, interaction data alone conveys little about the behaviours of users. To create a description of the participants activity during the study, descriptive statistics were calculated. A total of 57 metrics were extracted per participant, which included the total number of events, individual event counts (for example, the number of times the pause button was clicked), and relative frequency of events. Temporal statistics were also calculated: the time to completion (how long it took the user to reach a defined endpoint in minutes) and session length (the amount of time the user was actively using the media). Four types of pauses between interaction events were also captured: short (between one and five seconds), medium (between six and 15), long (16 and 30), and very long (more than 30), using the values defined in (Williams and Zitouni, 2017) which have demonstrated their usefulness in other domains (Mehrotra et al., 2017; Williams and Zitouni, 2017). The defined endpoint was step 25 in the tutorial, which is the point where the Origami Frog is complete and the remaining two steps demonstrate how to play with the creation.

Statistical Analysis

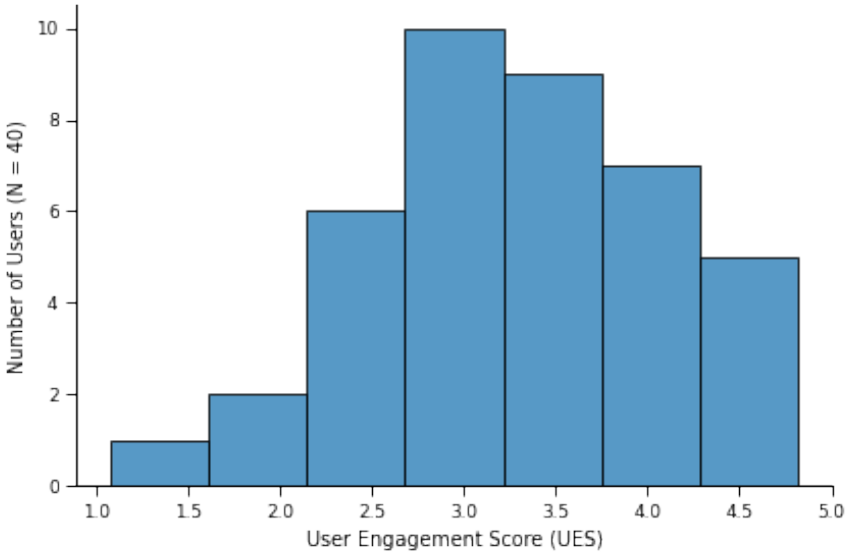
To understand how engaged the participants were with the experience, an analysis of the engagement metrics – derived from the responses to the survey – was performed. Descriptive statistics are reported for the four UES sub-scales, with Spearman's rank correlation coefficient calculated to test the distinctiveness between each of the sub-scales. To better understand whether demographics influence the engagement of participants, statistical tests were carried out to test for differences between the four engagement factors (perceived usability, aesthetic appeal, focused attention, and reward), as well as engagement (an average of factor averages), and demographic information provided by participants, specifically their familiarity with technology and previous experience in origami making. Before testing for differences, two statistical tests were performed to evaluate normality and variance equality: a Shapiro-Wilk test and a Levene test (F), respectively. In testing for statistical differences, the following procedure was carried out: if the distributions were found to be normal and to have equal variance, then a one-way ANOVA (F) was performed; if normal and a non-equal variance, then a Welch ANOVA was computed; if non-parametric, then a Mann-Whitney U test with common language effect size (f) was carried out.

Testing for statistical differences and correlations between metrics can only reveal so much in terms of understanding the relationships between metrics and engagement. To explore whether interaction metrics are predictive of engagement and the individual factors, interpretable models were trained and evaluated. As the data sample is relatively small, Decision Tree regression and classification models, where the regression learns to predict scores and the classification factor group membership (high or low), were trained. Optimised models are trained in both cases, to ensure the models were best fitted to the problem – mean squared error (MSE) and accuracy are used as scoring metrics. Evaluation of the model performance is reported through MSE, Mean Absolute Error (MAE), Root-MSE (RMSE), and R^2 score for regression, and accuracy, recall, precision, F1, and area under the curve (AUC) for classification. To evaluate whether demographics are predictive of engagement alongside interaction metrics, binary representations of previous origami making experience and technology experience are included as features in the models.

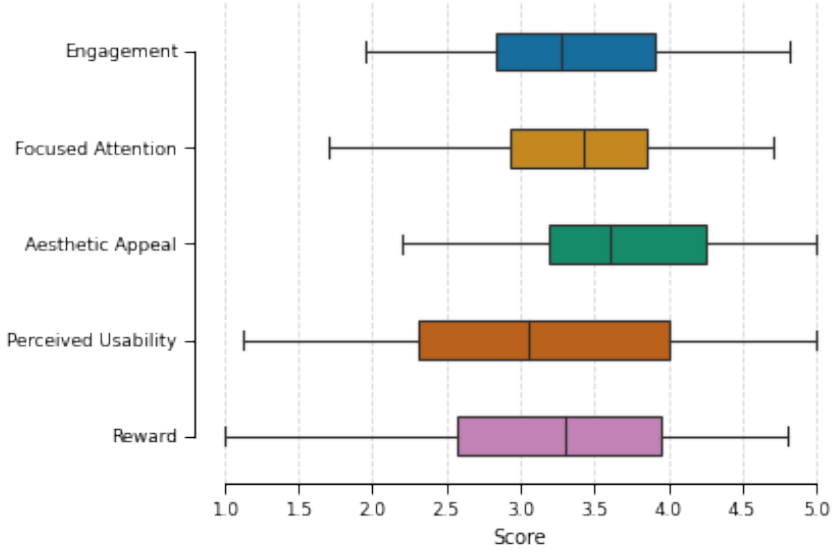
To answer **E-RQ2**, which focuses on understanding what interaction metrics are important, and as the models used are transparent in their decision-making process, the internal importance of metrics were extracted and permutation importance – where a metric is randomly permuted while the other remain static – was calculated.

3.1.2 Results

In total, 40 participants chose to take part in the study. As shown in Figure 3.1, the participants self-reported mixed engagement experiences: $M = 3.32$, $STD = 0.86$. The individual UES factors show (in Table 3.1 and Figure 3.1b) that most participants found the aesthetics of the experience appealing and that it captured their attention. There was a wider



(a) Distribution of engagement scores reported by audience members



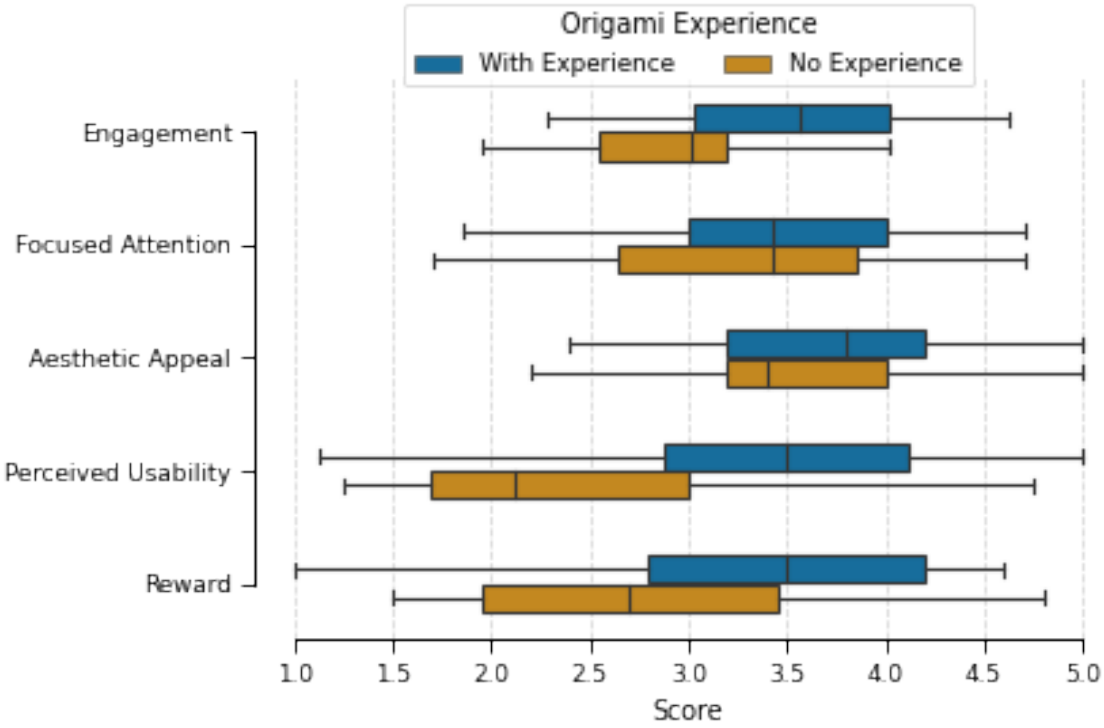
(b) Distribution of the individual factors and engagement

Figure 3.1. Distributions of engagement scores & UES Factors for Make-Along: Origami Frog

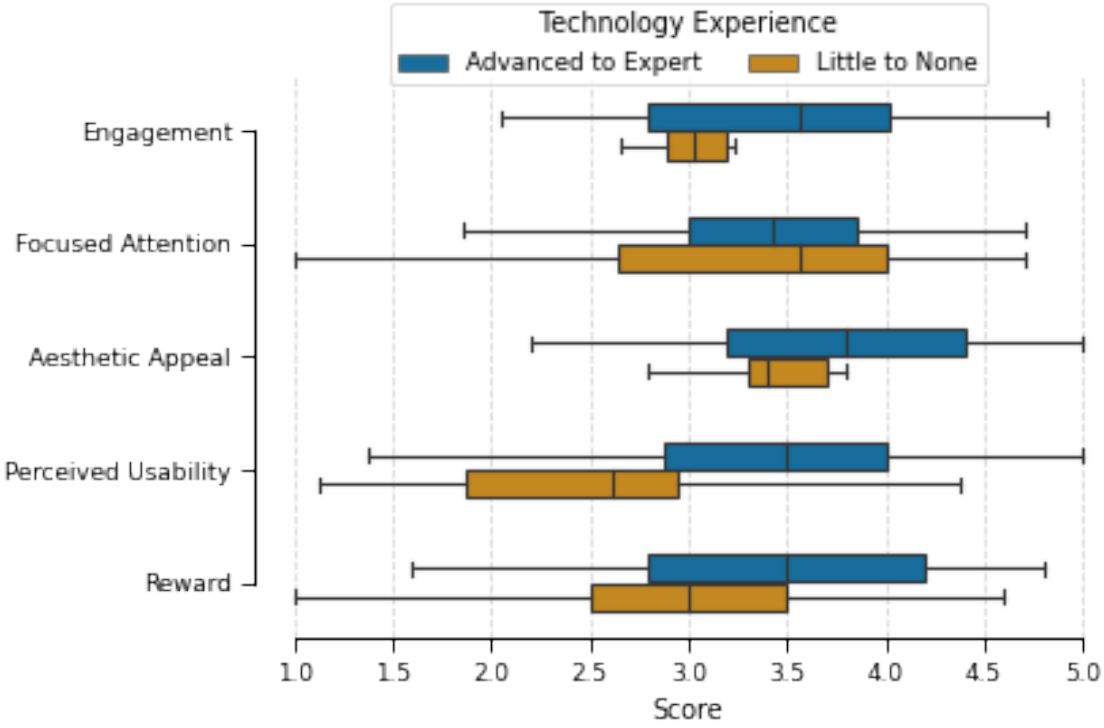
Table 3.1. The descriptive statistics of the UES factors self-reported by participants in the study. Spearman's rank correlation coefficient is reported (significance level: ***= $p < .001$, *= $p < .05$)

UES Factor	Mean	Std. Dev	Median	AE	PU	RW
Focused Attention (FA)	3.32	0.86	3.34	0.58***	0.45*	0.51*
Aesthetic Appeal (AE)	3.63	0.86	3.60	1	0.68**	0.83***
Percieved Usability (PU)	3.06	1.05	3.06		1	0.82***
Reward (RW)	3.19	0.95	3.30			1

range of opinions on the perceived usability of the experience with the factor recording the lowest mean score and the strongest disagreement between participants. Similarly, participants varied their opinion on the reward felt from using the interactive tutorial. Based on the reported correlations in Table 3.1, and as there are moderate correlations between the factors, it demonstrates that each cover different dimensions of engagement and that there is a distinctiveness between the factors.



(a)



(b)

Figure 3.2. The distribution of engagement and individual factors between the previous origami making experience and technology experience demographics

Figure 3.2 reports the distributions for each factor and engagement between two demographic factors: previous origami making experience (Figure 3.2a) and technology experience (Figure 3.2b). There is a disparity in the perceived usability in both cases – participants with no experience of making origami prior to the tutorial and those with little to no experience with digital technology do not find it usable in comparison with those that have experience (these are not the same groups of participants). In the tests for significant differences, it was found that there is a significant effect of previous origami making experience on the perceived usability: $F(1, 38) = 10.33, p < .05, \eta_p^2 = 0.21$, along with a significant effect of technology experience on the usability: $F(1, 38) = 6.46, p < .05, \eta_p^2 = 0.14$. There was also a significant effect of origami making experience on the reward: $F(1, 38) = 4.91, p < .05, \eta_p^2 = 0.11$. These effects are shown in Figure 3.3.

There were minor correlations between interaction metrics and engagement, as shown in Figure 3.4. The number of switch views – how many times the user changes the camera angle – is positively correlated with perceived usability, suggesting that the more a user changes their camera angle, the more usable they find the experience. Similarly, there were positive correlations between temporal metrics and focused attention; those that felt that the tutorial captured their attention take more time to go through the tutorial and spend more time on narrative elements.

When testing for statistical differences between interaction metrics and engagement, as well as the individual factors, there were no differences found between metrics and focused attention, aesthetic appeal, and reward. There were significant differences found between the number of switch view events and perceived usability ($U(N_{low} = 19, N_{high} = 18) = 265.0, p < .05, f = .77$), as well as the proportion of switch view events and perceived usability ($U(N_{low} = 19, N_{high} = 18) = 270.00, p < .05, f = .78$). These differences are again found between the same two metrics and engagement (number of switch view events: $U(N_{low} = 19, N_{high} = 18) = 238.00, p < .05, f = .69$; proportion of switch views: $U(N_{low} = 19, N_{high} = 18) = 240.00, p < .05, f = .70$) but with a weaker effect (as noted by f) and were likely carried over from being significant between the perceived usability groups. Figure 3.5 shows the differences between the switch view-based metrics with perceived usability and engagement. The results suggest that those who found the interactive tutorial more usable make greater use of the additional interactive features. This may be an artefact of the demographic differences found and presented earlier. However, it is worth noting the scale of the proportion metric, limiting the application of the proportion metric in real terms.

Figure 3.6 shows the correlation between two demographics and interaction metrics. Positive correlations were found between temporal metrics and previous origami-making experience, suggesting that the more experience a user has with origami, the more time they spend on the experience; hence taking longer to complete and spending more time on narrative elements. In contrast, there were negative correlations between the same temporal metrics and previous technology experience. There were also positive correlations between interaction-based metrics and previous technology experience which suggests that those

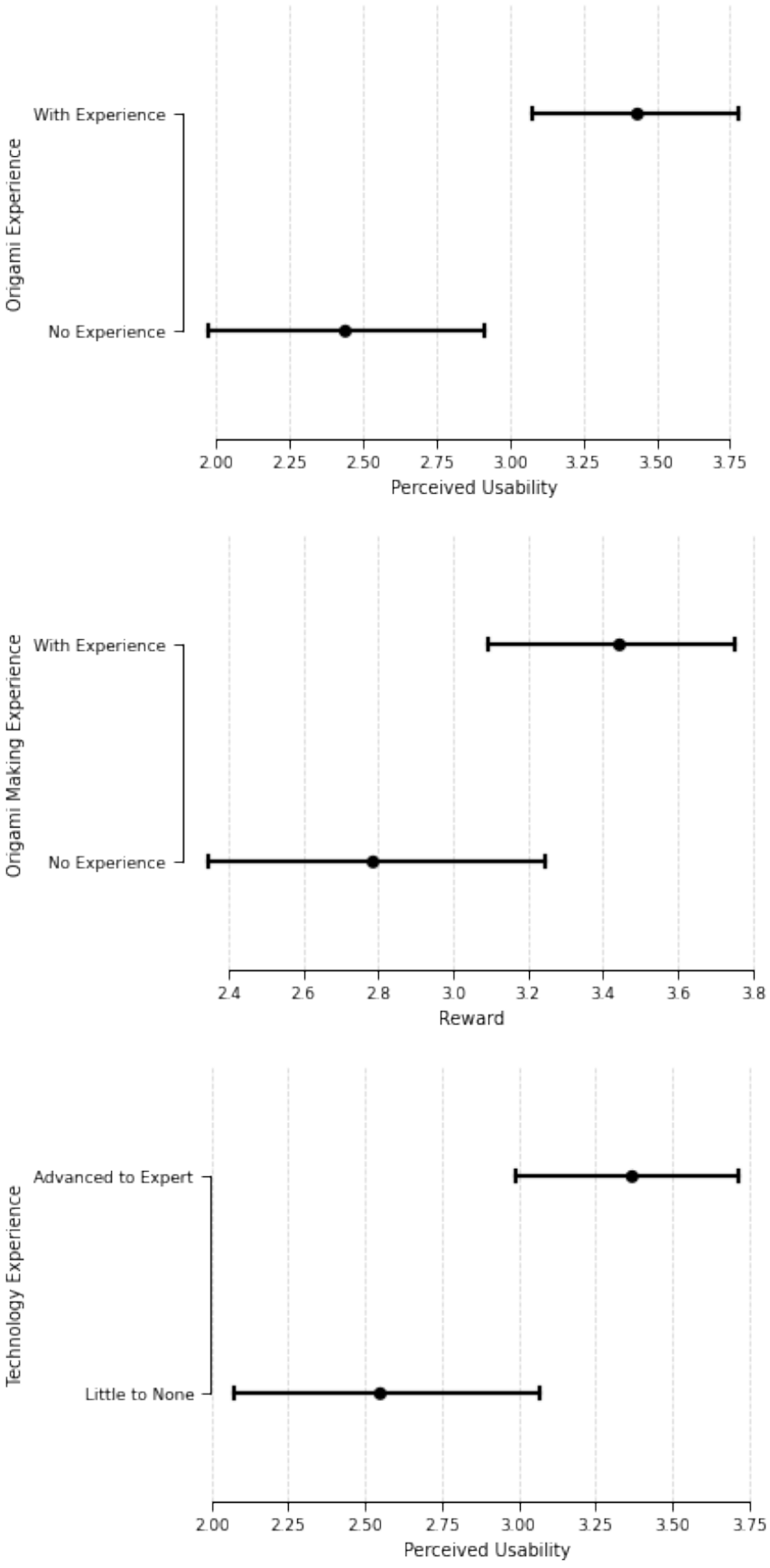


Figure 3.3. The significant effects between demographics (origami making and technology experience) and engagement factors (perceived usability and reward). The figures show that those with no experience (origami making experience) or little to none (technology experience) record lower perceived usability and reward scores.

Figure 3.4. The correlations between the interaction metrics and engagement factors. Correlations are calculated using Spearman's rank correlation.

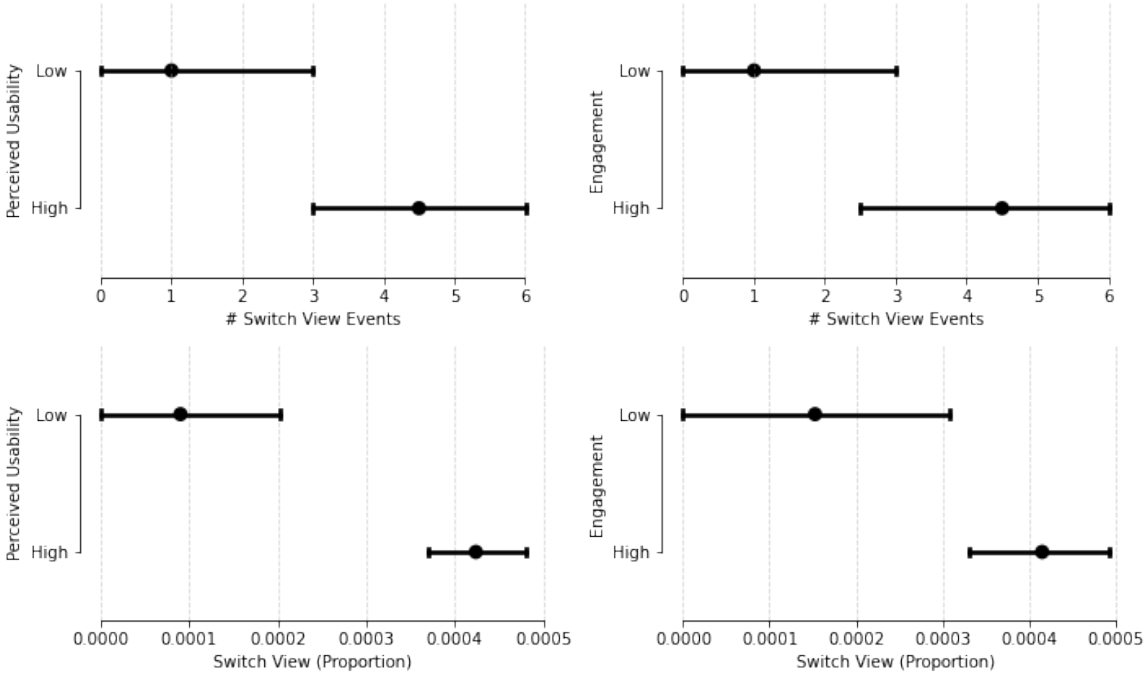


Figure 3.5. The effects between the switch view-based interaction metrics with self-reported perceived usability and engagement. Central tendency was estimated using the median.

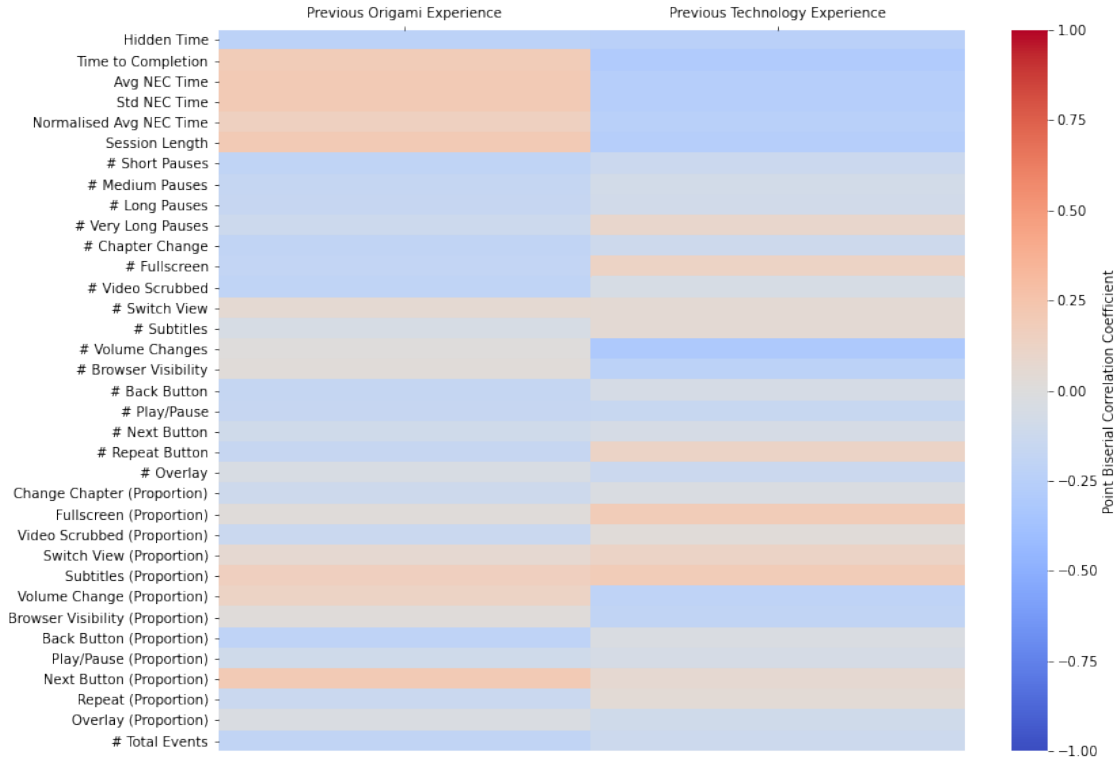


Figure 3.6. The correlations between the interaction metrics and demographics of participants. Correlations are calculated using Point-Biserial correlation

Table 3.2. The classification results from modelling engagement and individual factors. The results reported are split based on whether the two demographical features of interested are included as features, with the percentage change reported to measure the effect of their inclusion. UES = Engagement, FA = Focused Attention, AE = Aesthetic Appeal, PU = Perceived Usability, and RW = Reward

Factor	Accuracy		F-1		Precision		Recall		AUC	
	No Demo	Demo (±%)	No Demo	Demo(±%)	No Demo	Demo (±%)	No Demo	Demo (±%)	No Demo	Demo (±%)
UES	0.75	0.75 (0%)	0.75	0.75 (0%)	0.73	0.73 (0%)	0.77	0.77 (0%)	0.75	0.75 (0%)
FA	0.67	0.64 (-4.00%)	0.53	0.62 (16.72%)	0.77	0.61 (-21.42%)	0.41	0.64 (57.14%)	0.72	0.64 (-10.18%)
AE	0.75	0.78 (3.57%)	0.72	0.69 (-4.80%)	0.66	0.81 (22.72%)	0.80	0.60 (-25.00%)	0.78	0.83 (6.56%)
PU	0.81	0.81 (0%)	0.81	0.81 (0%)	0.78	0.78 (0%)	0.83	0.83 (0%)	0.81	0.81 (0%)
RW	0.78	0.67 (-13.79%)	0.78	0.57 (-27.61%)	0.71	0.72 (1.81%)	0.88	0.47 (46.66%)	0.82	0.71 (-12.54%)

with a higher degree of technology familiarity make more use of the features on offer in the tutorial, for example, switching views and fullscreen and subtitle usage. There were no significant differences found between the interaction metrics and demographics, where the metrics are compared between the two groupings created in the demographics.

Modelling

Beside statistical differences between self-reported usability and engagement with camera angle changing-based interaction metrics, there were no significant differences in the interaction metrics and engagement factors. The modelling of interaction metrics to predict engagement and individual factors that make up engagement could leverage latent relationships and reveal important predictive metrics.

The results of training Decision Tree classifiers are shown in Table 3.2. For each of the factors and engagement an optimised model was found, and the analysis reported is based on insights derived from each of the optimised models. Reasonable results were achieved in

Table 3.3. The regression results from modelling engagement and individual factors. The results reported are split based on whether the two demographical features of interested are included as features, with the percentage change reported to measure the effect of their inclusion. UES = Engagement, FA = Focused Attention, AE = Aesthetic Appeal, PU = Perceived Usability, and RW = Reward.

Factor	MSE		RMSE		MAE		R^2	
	No Demo	Demo (±%)	No Demo	Demo(±%)	No Demo	Demo (±%)	No Demo	Demo (±%)
UES	0.46	0.46 (0%)	0.68	0.68 (0%)	0.54	0.54 (0%)	0.24	0.24 (0%)
FA	0.49	0.37 (-24.94%)	0.70	0.60 (-13.36%)	0.58	0.46 (-20.66%)	0.31	0.48 (54%)
AE	0.62	0.46 (-25.845)	0.79	0.68 (-13.88%)	0.59	0.53 (-9.27)	0.06	0.30 (382.79%)
PU	0.71	0.74 (3.63%)	0.84	0.86 (1.79%)	0.61	0.72 (16.76%)	0.29	0.27 (-8.59%)
RW	0.71	0.63 (-10.72%)	0.84	0.79 (-5.51%)	0.70	0.55 (-21.23%)	0.15	0.24 (56.79%)

the prediction of all four factors, as well as engagement, which suggest that each can be predicted by the metric derived from the participants interactions collected during the study. Perhaps contradicting the earlier findings, introducing the prior origami making experience (with experience and without experience) and technology familiarity (little to none and advanced to expert) demographics as binary features tended to negatively impact the performance of the models. For example, the accuracy in predicting reward (RW) decreased by 13.39% and a similar effect was found in predicting focused attention (FA) with a percentage change of 4%. In some cases, the introduction of the demographics had a positive effect, for example, a 57% increase in the recall of the focused attention model. In the case of perceived usability – the most predictable factor – and engagement, the introduction of the demographics had no effect on the models suggesting that they rely solely on other metrics to predict at a constant rate.

An alternative approach is to model the self-reported scores for the factors and engagement; the results of training Decision Tree regression models are shown in Table 3.3. For MSE, RMSE, and MAE, these metrics describe the errors the models make when predicting the self-reported score and a decrease in the scores when demographics are added means the performance of the model has increased. Comparatively, R^2 is a measure of the goodness of fit and the closer to one the better is the model fit. The results show that the models can predict the scores with relatively small error rates but the performance of the models is poor. The introduction of demographics show an increase in performance across almost all factors, with the performance of the models increasing with respect to R^2, which suggests that including demographics adds some useful information.

Feature Importance

When extracting feature importance from the models and performing permutation importance, it was found that there is a reliance on a small number of metrics to predict either the high/low groupings (Table 3.4) or self-reported scores (Table 3.5), where only single metrics are used in the latter. The maximum number of interaction metrics that were important in both cases was three, with all other non-reported interaction metrics having no impact on the predictions of either model types. In the cases where more than a single metric is relied on in the prediction, the mean decrease in accuracy demonstrates that the models still rely heavily on the most important metrics. The relationship between perceived usability and

Table 3.4. The feature importance and permutation importance for the classification models with demographics included as features. Only the metrics that have recorded an importance score are reported - in both approaches of feature importance, the ranking of features was the same - all other metrics had no influence on the performance of the models. Importance is the models internal measure of importance and MDA is the mean decrease in accuracy with a higher number indicating a larger impact on the model when the metric is randomly shuffled.

Factor	Metric	Importance	MDA (Std. Dev)
UES	Switch View (Proportion)	1.00	0.33 (0.09)
	-	-	-
	-	-	-
PU	Prior Origami Experience	0.84	0.64 (0.10)
	# Fullscreen	0.15	0.09 (0.04)
	-	-	-
FA	Std. Dev NEC Time	0.64	0.59 (0.14)
	Fullscreen (Proportion)	0.22	0.13 (0.04)
	# Total Events	0.12	0.08 (0.02)
AE	Switch View (Proportion)	1.00	0.16 (0.04)
	-	-	-
	-	-	-
RW	Prior Origami Experience	0.45	0.66 (0.23)
	# Back Button	0.35	0.22 (0.15)
	# Switch View	0.19	0.12 (0.05)

Table 3.5. The feature and permutation importance for the regression models with demographics included as features. Only the metrics that have recorded an importance score are reported - in both approaches of feature importance, the ranking of features was the same - all other metrics had no influence on the performance of the models. Single interaction metrics were importance in the regression case. Importance is the models internal measure of importance and MD-MSE is the mean decrease in the mean square error.

Factor	Metric	Importance	MD-MSE (Std. Dev)
UES	Switch View (Proportion)	1.00	0.27 (0.07)
PU	Switch View (Proportion)	1.00	0.32 (0.07)
FA	Next Button (Proportion)	1.00	0.14 (0.07)
AE	# Switch View	1.00	0.16 (0.08)
RW	# Switch View	1.00	0.17 (0.09)

prior task experience is again shown, with prior origami making experience being the most important metric with the greatest effect on the performance when predicting high/low usability (along with reward). The variation in time spent on the segments of the tutorial were the most important in predicting focused attention and could suggest that a deviation in the average time spent on segments might be related to how the tutorial grabbed the attention of the participant.

In the case of predicting the self-reported scores, single interaction metrics were used by the models to make the prediction (Table 3.5). The importance of the switch view-based metrics is demonstrated here, with four of the modes relying solely on either the number of events or their proportion. The results however show that there is a limit to the generalisability and predictive power of the models, ultimately limiting the insights that can be drawn.

3.1.3 Conclusion

The findings from the analysis performed on the interactive tutorial, particularly the modelling of engagement and its factors, yielded little actionable insight. Significant differences were found between the demographics of the participants and their engagement, specifically their perceived usability of the tutorial. While there were statistical differences in the usage of the switch view functionality, these differences did not manifest themselves in a meaningful way when modelling and predicting engagement using the interactions of participants. However, the findings point towards needing to collect a different type of data to understand the audience in this context, perhaps through taking a qualitative approach. The controlled nature of the study is likely a causal factor: participants that have less experience with the task and technology are unlikely to actively choose the experience 'in-the-wild', but the results do demonstrate that those types of users should still be considered when designing and producing these types of experiences. Anecdotally, participants reported not interacting with the experience and not understanding the additional controls on the interface. In the next section, a study is carried out to investigate engagement in-the-wild.

3.2 Click

In this section, whether there are predictive signals of engagement in interaction data collected from a different example of an interactive media experience is investigated. The study presented in this section differs from the previous as it is less controlled and performed 'in-the-wild', with engagement and interaction data being captured from a large audience attracted to the experience organically, as opposed to being recruited.

The experience is a special interactive edition of the BBC TV technology show *Click*, which was created to celebrate its thousandth edition. It used a form of storytelling, known as an interactive branching narrative, where the user determines their path through the story based on their interests. The show is made up of four sub-stories which each covering different topics; the two main sub-stories are about technology use in Malawi and autonomous vehicles. The audience can control what they see both between and within these sub-stories, e.g., they can choose to view one or both use-cases for the autonomous vehicle technology (industrial and/or consumer) along with having the option to go into more or less detail about the technology. The decisions that the audience are asked to make are prompted by the on-screen host (shown in Figure 1.2). If the audience decides not to interact, then a default path is automatically followed.

As with the previous experience, the show is divided into narrative elements and the audience has control over the show and video content, such that they can navigate between the narrative elements using the back and next buttons and replay or rewind the current narrative element. In this experience, however, the audience do not have control over the presentation of the content, though standard video controls are available: play/pause, full-screen, volume control, and video scrubbing.

3.2.1 Methodology

Study Design

The study was designed to be opt-in and in-the-wild, set in a live production environment attached to the official national release of Click. Both interaction data and ground truth engagement metrics were collected in a privacy-first manner, with no identifying information collected. To differentiate between users, at the start of their session they were each randomly assigned a generated identification string which did not persist across multiple visits. Explicit consent was requested, using a privacy notice detailing the data collected during the experience, from each audience member for data collection for research purposes and the study presented in this section focuses on those that completed the engagement survey.

Data Collection

Much like the previous experience, interaction data were collected using built-in analytics, which log when a user performed an action on the interface or when a contextual change occurred (a window orientation or browser visibility change). The data takes the following form: *user_id* - an anonymous identification string, *timestamp* - millisecond granularity timestamp, *action_type* - the type of event that occurred, *action_name* - the button clicked/context change, and *data* - additional metadata about the event, for instance hidden/visible for browser visibility changes. As the Click experience has a different set of interaction opportunities for the audience – for example, the audience cannot change their camera angle – the set of events that are logged by the built-in analytics are different: play/pause, back, next, fullscreen, subtitles, volume, video scrub, seek backwards, seek forwards, browser visibility change, window orientation change, narrative element change, and link choice.

User engagement metrics were also captured using the UES survey. As the survey was attached to a live production system, a short-form version of UES containing 12 questions was used, instead of the long-form 30 question version used previously. To avoid disrupting the experience, the survey was administered post-credits, which does introduce a sampling bias to the data as only a self-selected sub-sample complete it. Similarly, audience members that did not reach the end were not questioned as it would require the tracking of the audience and predicting when they were about to leave, which is a research direction explored in the next chapter (see Chapter 4) and was not possible at the time of the study. Much like the previous study, the questions were altered to fit with the Click experience, for example “Using Application X was worthwhile” to “Using this interactive episode of Click was worthwhile”. The responses to the survey cannot be used in their raw form, so following the guidance of the UES authors to create scores (O’Brien et al., 2018), a score for each factor was calculated as a mean and then a mean of means for engagement to obtain a final score.

Interaction Metrics

As with the previous study (see Section 3.1.1), the interaction data collected from Click tells us little about the users during their time on the experience. To create a descriptive picture of the user session, metrics were extracted from the interaction data which included individual event counts, relative proportion of events, and the total number of events. Similar to the previous study, temporal statistics were also derived: the time to completion (the amount of time to a defined endpoint), session length (the overall time spent on Click in minutes), and hidden time (the time spent with the browser window hidden). To create a notion of the lack of interaction, the pauses between events were also calculated using the same splits as previously: short (between one and five seconds), medium (between six and 15), long (between 16 and 30), and very long (more than 30). In the calculation of the session length and time to completion, the hidden time was subtracted as interest is in the amount of active time that the user spent on the experience and to reduce the number of outliers caused by large session lengths. For the time to completion metric, a node was chosen in the Click experience story graph where all sub-stories are brought back together, situated just before the ending scenes of the show. The justification for choosing this endpoint is that it is where all paths through the story graph meet and the main portion of the narrative finishes, with only the credits remaining.

Analysis

To discover differences and relationships between interaction metrics and engagement levels, correlation analysis was performed, statistical differences between engagement levels were tested for, and investigated metrics that discriminate between high and low engagement. Prior to performing correlation analysis, Shapiro-Wilk's test for normality was performed to determine which correlation test to perform. For non-parametric metrics, Spearman's rank-order correlation coefficient (r_s) was used, whilst for normally distribution metrics, Pearson's r correlation coefficient was calculated. Statistical differences were tested for using the Mann-Whitney U test and common language effect size (f), corrected for multiple tests using False-Discovery Rate. The *high* and *low* engagement levels were determined by splitting users based on the median UES score as in (H. L. O'Brien and Lebow, 2013). To investigate the discriminatory power of interaction metrics between high and low engagement, and following a similar approach to (Mehrotra et al., 2017; Zhuang et al., 2018), a chi-square test (χ^2) was performed. This test evaluates the probability of the observed result given the null hypothesis being true, as such the following null (**H0**) and alternative (**HA**) hypothesises were proposed:

H0 There are no detectable differences in the interaction metrics between the two engagement classes (low and high).

HA There exist detectable differences in the interaction metrics between the two engagement classes (low and high).

To investigate if interaction metrics are predictive of engagement, interpretable models to predict high or low engagement were trained, feature importance was extracted, and a model-agnostic interpretability method called Shapley Additive Explanations (SHAP) (Lundberg and Lee, 2017) was applied. Each sample was scaled to a range between zero and one to account for differences in ranges between metrics, e.g., session length (minutes) and proportions (zero to one). Metrics containing more than 50% of zeros were converted into binary.

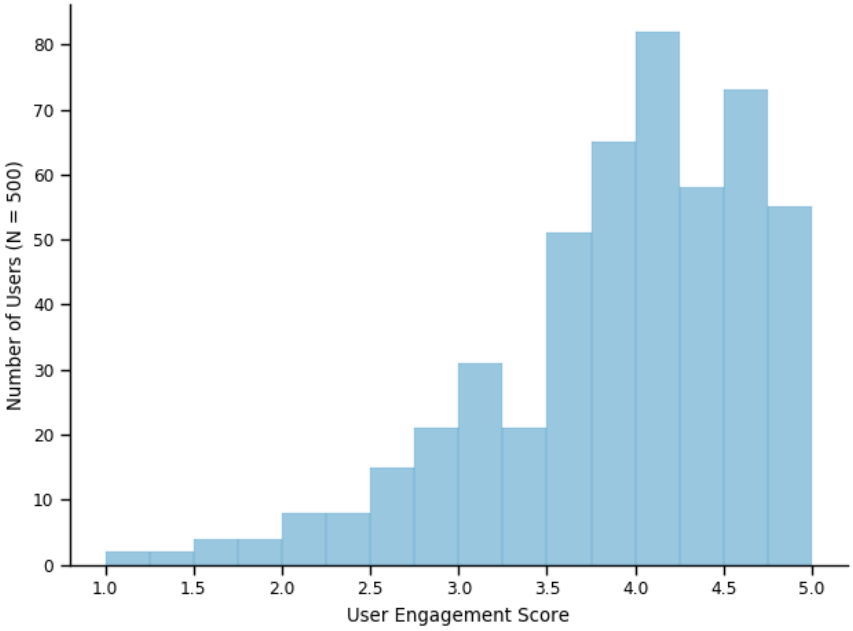
A range of interpretable models were trained and evaluated, with the processed interaction metrics as input, using 10-fold stratified cross-validation with Area Under the Curve (AUC) as the performance metric – this metric was chosen as to evaluate the model's ability to distinguish between the two target classes. The choice of models was primarily driven by interpretability, but also models that make different assumptions about the data and have been proven to work well on a range of prediction tasks. More complex models, such as Neural Networks, were not used due to added complexity in terms of interpretability and structure. The models evaluated were: Logistic Regression, Gaussian Naïve Bayes, Decision Tree, Linear Discriminant Analysis, K-Nearest Neighbours, Support Vector Machine, and SVM with stochastic gradient descent. Once the best performing model was identified, 10-fold stratified cross-validated grid search was performed to find the optimal hyper-parameters, an optimised model was trained using a $80 : 20$ train-test split - this model is used in the investigation and is the focus of the rest of the section.

To evaluate the importance of interaction metrics when predicting high and low engagement, feature weights were extracted from the model and SHAP values were calculated which provide a model-agnostic measure of how a feature value impacts the model prediction.

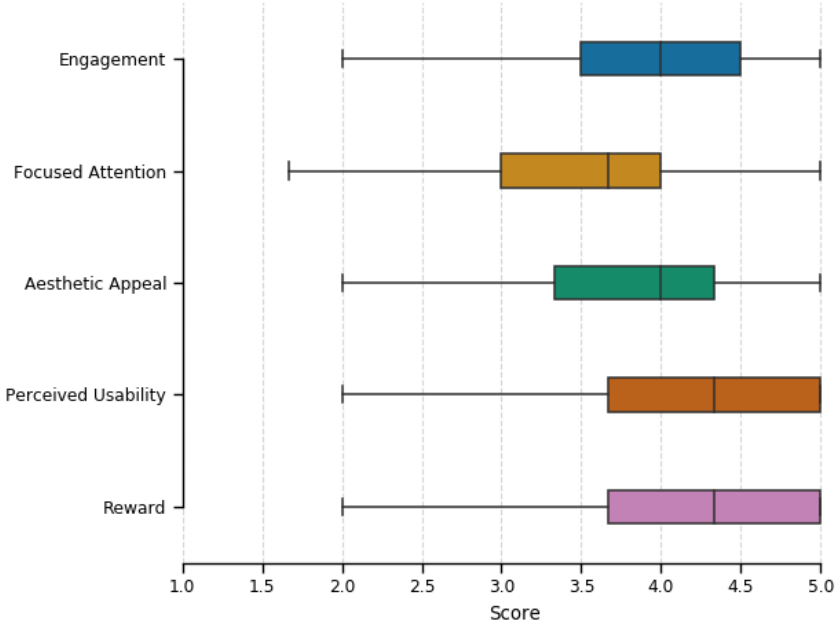
3.2.2 Results

Sample

In total, 500 members of the audience chose to take part in the study by submitting a response to the UES survey. The distribution of their engagement scores are shown in Figure 3.7a, and the scores of the four engagement factors are shown in Figure 3.7b. Overall, the audience were highly engaged with the experience ($M = 3.87, STD = 0.77$). However, due to the survey placement - post-credits - the sample captured may be skewed, and less engaged users are less likely to reach this part of the show. Nevertheless, there is still a proportion of users that were not engaged by the experience (recording scores below three, 14.77%, $M = 2.48, STD = 0.49$), suggesting that even if a user reaches the end of the experience, it does not necessarily mean they were engaged - a hypothesis for this is presented later. On the whole, as shown in Figure 3.7b, the audience found the experience to be usable ($M = 4.10, STD = 0.85$) and they felt reward from using it ($M = 4.12, STD = 0.93$). However, the audience felt that the experience did not capture their attention to a



(a) Distribution of engagement scores reported by audience members



(b) Distribution of the individual factors and engagement

Figure 3.7. Distributions of engagement scores & UES Factors for BBC Click

high level ($M = 3.46, STD = 0.97$), a crucial aspect of media creation. Similarly, the audience varied their opinion on the aesthetics of the experience ($M = 3.79, SD = 0.91$). From the 500 participants, 310,800 ($M = 621.60, STD = 379.83$) interaction events were recorded.

Engagement Relationships

Between the majority of metrics and engagement scores, minor positive and negative correlations were observed. Session length had the strongest correlation with the overall engagement scores ($r_s = .32, p < .001$), and time to completion was second strongest

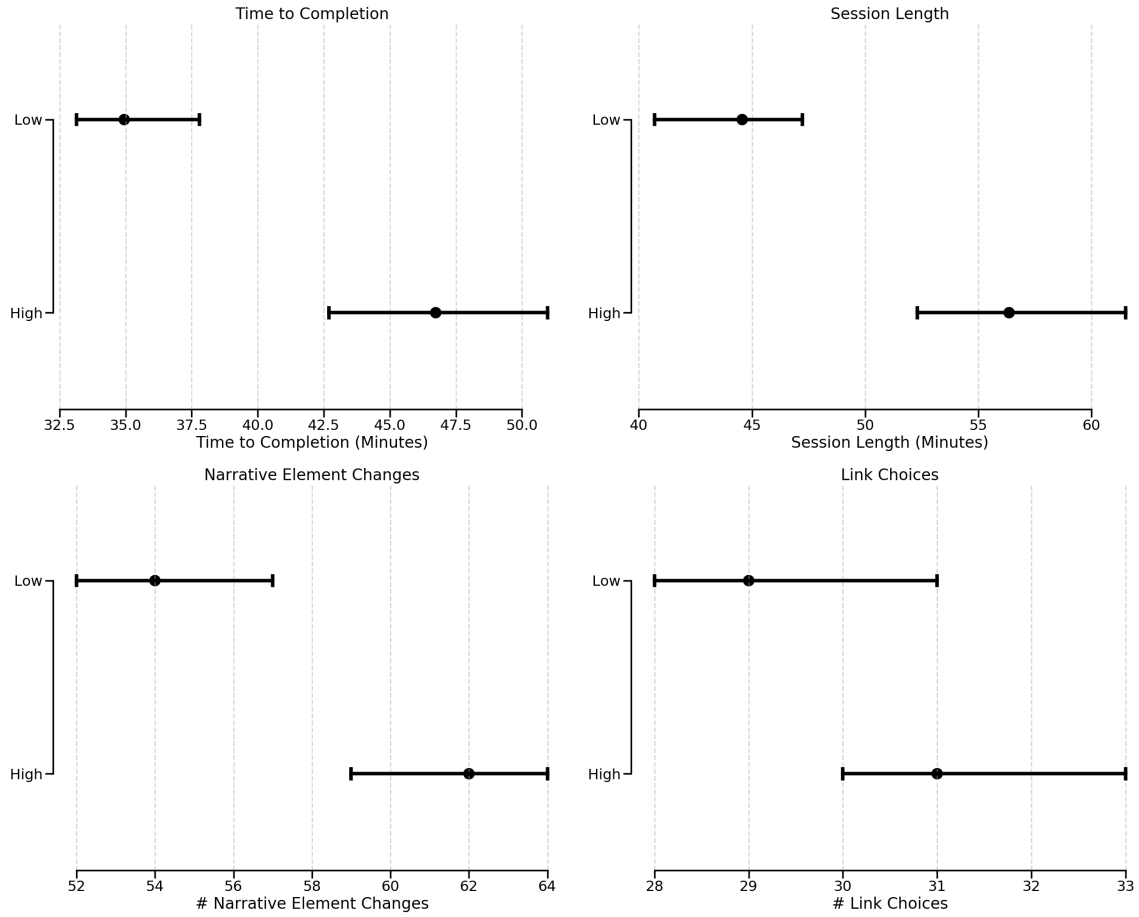


Figure 3.8. Estimations of the central tendency for the four metrics that are significantly different between the two engagement groups; the median was used as the estimator. The plots demonstrate that low engagement participants record lower numbers of the metrics.

($r_s = .30, p < .001$), both indicating a relationship between the time spent on the experience and engagement. These two metrics are related but differ slightly as the session length is how long the user's session lasts, from start to finish, whilst the time to completion metric is the time it takes for the user to reach a defined endpoint.

When applying Mann-Whitney's U test, with correction for multiple tests, to each metric between the two engagement groups ($N_{low} = 259, N_{high} = 241, low \leq 4.0 < high$), there were significant difference between low and high engagement with respect to time to completion ($U(N_{low} = 259, N_{high} = 241) = 22443.00, p < .001, f = .64$) and session length ($U(N_{low} = 259, N_{high} = 241) = 22111.00, p < .001, f = .65$)). For metrics relating to interaction events, the number of narrative element changes, recorded when the user moves from one story element to another, are statistically different: $U(N_{low} = 259, N_{high} = 241) = 23527.00, p < .001, f = .61$. The distribution of link choices, recorded when a user actively chooses a presented option, were also significantly different between engagement levels: $U(N_{low} = 259, N_{high} = 241) = 27724.50, p < .05, f = .54$. These differences between the two groups are shown in Figure 3.8, which demonstrates that users in the high engagement group typically have longer sessions, take more time to complete, record more narrative element changes, and make more active choices in their viewing path.

To test the ability of interaction metrics to discriminate between low and high engagement,

Table 3.6. Discriminatory statistical features (significance level: ***= $p < .001$, *= $p < .05$). pr_l and pr_h denote the Pearson residuals for the feature for low and high engagement. $total_l$ and $total_h$ represent the observed frequency for each feature in the low and high engagement classes. Direction represents how the relationship is weighted, a negative direction means the low engagement group recorded more of the feature.

Feature	χ^2	pr_l	pr_h	$total_l$	$total_h$	**Direction**
Next Button	175.09***	9.23	-9.48	2388	1464	-924
Short Pauses	84.06***	6.39	-6.56	1230	767	-463
Narrative Element Changes	39.67***	-4.39	4.51	14626	14918	292
Medium Pauses	36.33***	4.20	-4.31	731	488	-243
Long Pauses	24.38***	3.44	-3.53	566	388	-178
Play/Pause	11.59***	-2.37	2.43	742	835	93
Very Long Pauses	5.57*	-1.64	1.69	1245	1296	51
Fullscreen	4.35*	-1.45	1.49	314	350	36

the chi-square test was performed to evaluate the probability of the observed result given the null hypothesis being true (**H0**). Table 3.6 presents the statistically significant results, ordered by discriminatory power (χ^2). The number of next buttons (the function to skip through narrative elements) is the most discriminatory metric, suggesting that users in the low engagement group skip through content much more frequently than those in the high engagement group (indicated by the direction of the relationship). Further, the second-ranked metric, short pauses, demonstrates a significant difference in the way that users consume the media. Users in the low engagement group were more likely to record shorter periods between interactions; in contrast, users in the high engagement group were more likely to record very long pauses - indicating they likely spend more time watching content.

Predicting Engagement

When evaluating which model performed best in predicting engagement levels, Logistic Regression (LR) performed the best across all evaluations (LR $\mu AUC = .60$, all other models: $\mu AUC = 0.55 - 0.60$), and as it is a comparatively simpler and more interpretable model, it was used for the rest of the analysis. To tune the model to the data, the normalised penalty (L_1 and L_2), the strength of regularisation, and the stopping criteria tolerance were optimised; the most optimal ($AUC = .61$) hyperparameter configuration was: L_2 normalisation penalty, a regularisation strength of 1, and a stopping tolerance of 1×10^{-5}. Fitting an optimised model on a training set consisting of 400 samples and testing on 100 samples, the model could accurately separate the two engagement classes ($AUC = .66$, *precision* $= .61$, *recall* $= .61$, *f1-score* $= .61$), as shown in Figure 3.9. A paired t-test was computed to compare the performance of the model with a baseline that generates predictions uniformly randomly, finding that the Logistic Regression model performs significantly better ($t(32) = 4.47, p < .01$). These results suggest that engagement can be modelled and accurately predicted using interaction data collected from an interactive media experience.

To assess metric importance, regression coefficients were ranked, which describe size/direction of the relationship between a metric and target (Figure 3.10). Coefficients with a large positive value increase the probability of predicting high engagement; large negative values increase the probability of predicting low engagement.

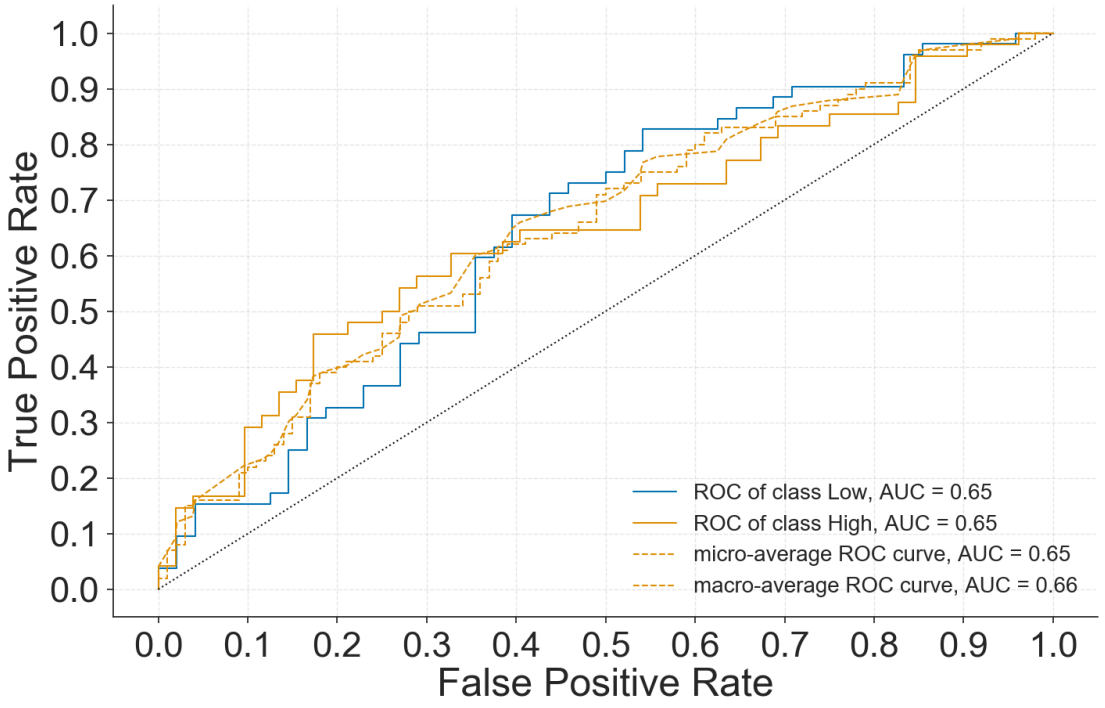


Figure 3.9. Receiver Operating Characteristic (ROC) curve demonstrating the model's ability to distinguish between low and high engagement

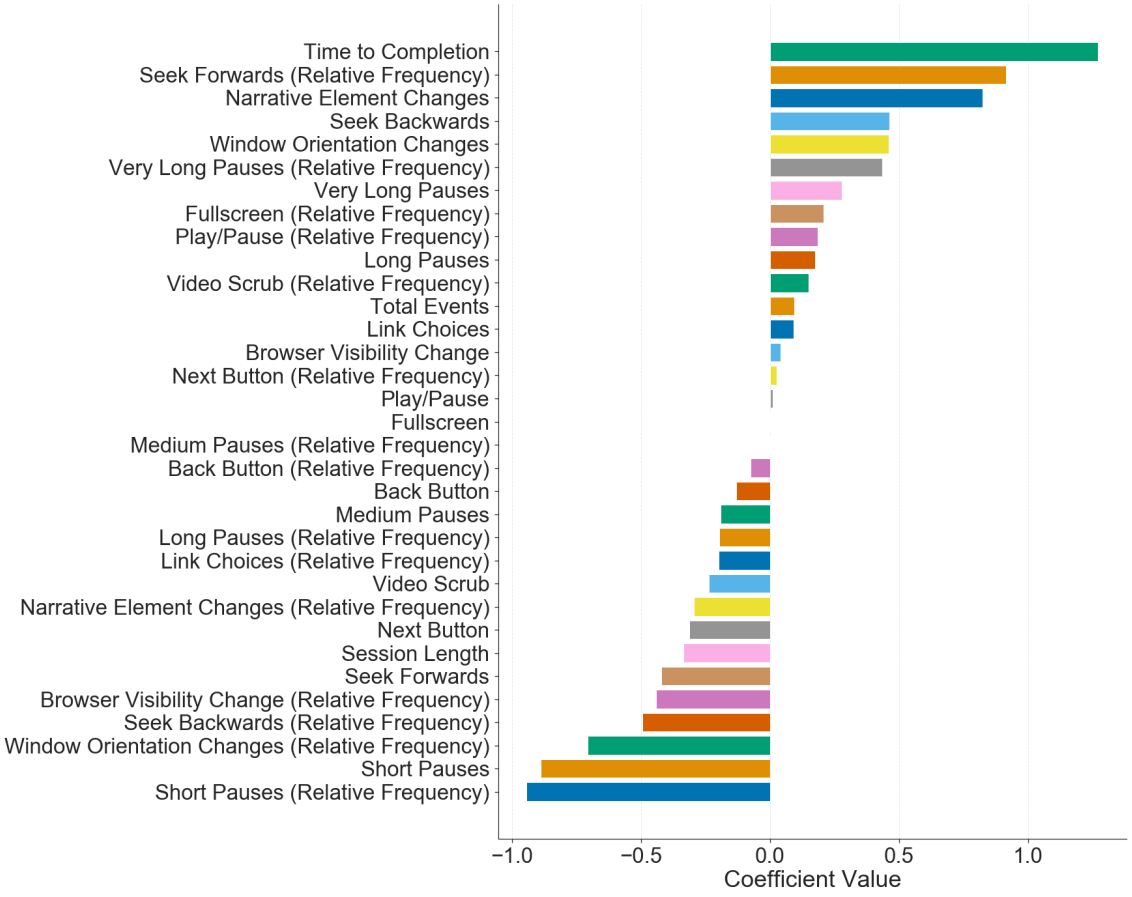


Figure 3.10. The coefficients for each metric that is used in the prediction of engagement. A positive coefficient value indicates that the metric weights the prediction towards high engagement and a higher coefficient value signals a heavier weighting on the prediction.

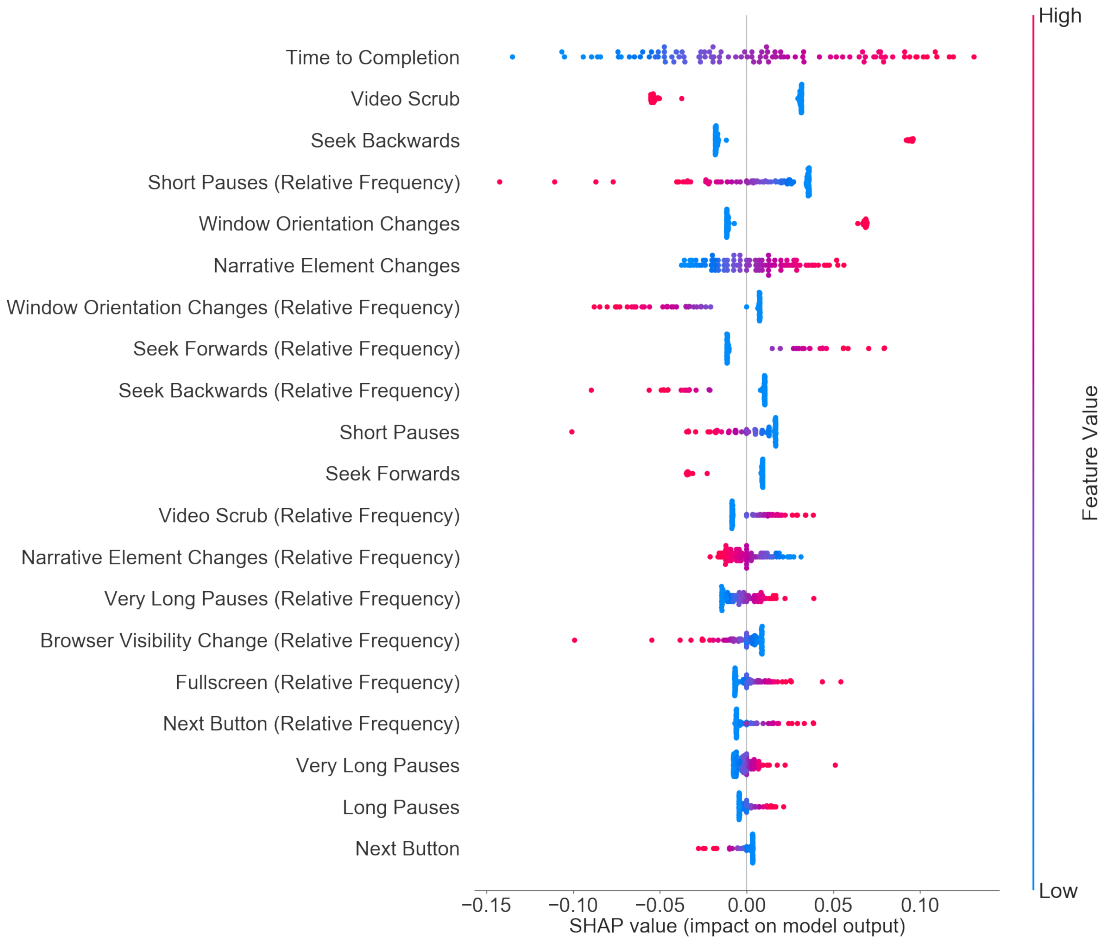


Figure 3.11. Feature importance based on contribution to model output (SHAP values). The position of the metrics on the y-axis is ordered by the sum of SHAP values across all samples.

Figure 3.10 shows that an increase in the time to completion metric correlates to higher engagement, supporting the significant difference between high and low engagement seen in Figure 3.8. Similarly, an increase in narrative element changes is positively associated with higher engagement; however, as the relative frequency of narrative element changes increases, the prediction output weights towards low engagement. The importance of very long pauses and short pauses (both counts and relative frequencies) is demonstrated by high engagement being positively associated with a large number of very long pauses, and negatively associated with a large number of short pauses. In contrast, a longer time to completion, higher count of narrative element changes, and more very long pauses are associated with high engagement prediction. The importance of temporal metrics is demonstrated; the time between events, rather than the events themselves, is important. It is also observed that the occurrence of a window orientation change event is associated with higher engagement, but a high frequency of such events is associated with low engagement.

To further explore the interaction metrics contributions to the prediction of high and low engagement, SHAP values were calculated – results are shown in Figure 3.11. The results reveal that a high value (indicated by the red colour) for the time to completion metric increases the probability of predicting high engagement, while a lower value increases the probability of predicting low engagement, further corroborating the findings shown in Fig-

ure 3.8. The second most important feature, video scrubs, shows that there are two groups of users: the first record higher values which result in pushing the prediction towards low engagement, whilst the second record smaller values which push the prediction towards higher engagement. Similarly, an increase in seeking backwards (rewinding the video by ten seconds) events results in higher engagement, while smaller values have little contribution to the model output.

3.3 Summary

This chapter aimed to investigate whether there are signals in the interaction data that are predictive of engagement within the context of interactive media experiences. Techniques and approaches that have had prior success in other domains are untested in this re-emerging area of media and the unobtrusive monitoring of interaction data could provide media creators with an ability to measure the success of experiences. The chapter presented two studies of engagement: 1) a semi-controlled study of an adaptive tutorial; and 2) an in-the-wild study of an interactive branching narrative – in both, interaction data and self-reported measures of engagement were captured.

To investigate **E-RQ1**, models of engagement were trained and evaluated in both studies. In the study of the adaptive tutorial, it was found that engagement, and the individual factors, can be predicted with a reasonable amount of accuracy. With the introduction of demographics that had a statistically significant effect on the engagement and perceived usability of the tutorial tended to negatively impact the performance of the models. These results are counterintuitive, but a potential explanation is a lack of interaction by users that found the experience less usable and therefore less engaging. Whereas, those with higher engagement (along with more origami making and technology experience) made more use of the interaction opportunities. When focusing on modelling high and low engagement with the interactive TV show, it was found that membership of the groups could be accurately predicted from interaction-derived metrics.

For **E-RQ2**, the internal representations of importance and model agnostic-importance measures were extracted and calculated. The individual engagement factors and engagement itself could be predicted at a reasonable rate in the adaptive tutorial. However, most of the models relied on a small subset of interaction metrics, usually only one, which limits the inference that could be carried out. Notably, however, through statistical analysis, it was found that there was a relationship between the context of the user and their felt engagement with the tutorial: participants with no prior experience in making origami found the tutorial less usable than those who had experience, and participants with less experience with technology also found it less usable. These differences manifested themselves in the interaction metrics, with a significant difference found in the number of switch view events (where the user could change the camera angle) between high and low usability and engagement. Differences may exist in modelling of more fine-grained interaction data (e.g., cursor movements) but these findings point towards the need to collect a different type of

data to understand the audience of an adaptive tutorial, perhaps through taking a more direct approach and collecting qualitative data.

The interactions that the model showed were most important in predicting low engagement were more next button clicks, more narrative element changes, and shorter pauses between interactions. In combination, these can be seen as the user jumping through the content and interacting more regularly (a skipping-type behaviour). This is in contrast to users that were more highly engaged, who typically interacted less (longer time to completion, higher count of narrative element changes, and more very long pauses). Given the nature of the content (similar to a traditional TV show), we postulate that engagement is associated with focus on the content and only interacting when necessary (a consumption-type behaviour). On the other hand, skipping through the experience indicates lower engagement with the content (analogous to fast-forwarding a video). In contrast, the Origami tutorial was designed to support a task, and users who are engaged with that task are likely to want to interact more with the experience. Those who sit back and watch are, in this case, less likely to feel engaged. It is worth noting, and a point that requires further investigation, that it is unclear whether audience members are skipping because they are not engaged with the content or whether they are unable to engage with the content because they are skipping.

Monitoring for the metrics that make up these potential behavioural proxies could provide content creators with a detailed picture of the success of their experience and the means to retrospectively understand what types of story experiences work better. For example, identifying a loss of engagement during the experience and detecting skipping could provide the opportunity to modify an experience on-the-fly to provide a shorter, summarised version of the narrative.

Interactive media experiences are a new type of media and are continually in development so understanding whether there is a commonality in user behaviour that is indicative of engagement across multiple experiences could allow media creators to deploy and perform the same retrospective analysis. The discovery of differences rather than similarities would enable media creators to focus on particular aspects of the user behaviours for specific types of interactive media experiences.

E-RQ3 focuses on whether there was a shared importance in user behaviours across the two interactive experiences or differences between the two interactive media experiences that are the subject of this chapter. While in each experience there were factors important in the understanding and prediction of engagement, there is a lack of commonality between the two. In the interactive tutorial, factors outside of the experience – the context of the user, their demographics – affect how usable and engaged they will be with the experience. While in the interactive TV show, a combination of interaction metrics can differentiate between engagement levels and point towards behavioural proxies being present in the data. These differences could be due to the nature of the experiences. The task of following a tutorial requires active engagement with the content, so we find more engaged users interact more. A TV show is designed for the audience to sit and watch the content, which results in

more engaged users interacting less.

The work presented in this chapter focused on gathering an understanding of engagement and whether it is predictable within the context of interactive media experiences. It was found that the context of the user affects their engagement and perceived usability with an interactive tutorial, whereas in an interactive TV show, audience members that experience a low degree of engagement appear to exhibit a skipping-type behaviour and those with a higher degree of engagement show a consumption-type behaviour – both of which could be monitored in real-time and used in retrospective analysis of these types of experiences by media creators. However, both studies focus on a small sub-sample of users, which may limit the generalisability of the results to the wider audiences attracted to these experiences. As such, the next chapter shifts focus to model, predict, and understand user abandonment with the interactive TV show in the wider dataset collected when the show was released to the public. In addition, the possible relationship between abandonment and engagement is explored.

To summarise, the following are the key takeaways from this chapter. Engagement can be predicted from audience interactions with an interactive branching narrative, but not with an adaptive tutorial. This could be owing to the small sample size or the nature of the more involved experience or both. Instead, for the adaptive tutorial, the context that an audience member brings to an experience affects their engagement. For the interactive branching narrative, audience members with a reported lower engagement appear to skip through the content, while those with higher engagement appear to sit-back and consume. Metrics that were important in each interactive experience were not shared across both, suggesting that experiences may need to be considered individually.

Chapter 4

Predicting Abandonment in Interactive Media Experiences

Having considered user engagement in interactive media experiences, the focus now shifts onto investigating abandonment in interactive media experiences; specifically, whether abandonment can be predicted, what interaction metrics are important for its prediction, and whether there is a relationship between abandonment and engagement. Abandonment is where audience members dropout of the content before reaching its end and is typically a signal of a negative experience for the user under the premise that if they were engaged or enjoyed the content then they would have stayed until the end. Audience abandonment itself is not difficult to measure: a measure of success currently deployed by media producers is counting the number of audience members that completed the show, which is a rudimentary measure but provides high-level insights into the performance of the content – the higher its number with respect to the total audience size, the better.

Prior work has shown that going beyond these types of basic measurements and using the interactions of users can provide a more nuanced understanding of abandonment. For example, through modelling *good* (the user finds what they are looking for) and *bad* (they do not) abandonment (J. Li et al., 2009) and understanding the potential underlying causes (Diriye et al., 2012; Williams and Zitouni, 2017). Additionally, the previous chapter demonstrated that interaction data is useful to differentiate between engagement levels, here we investigate whether it is also useful in predicting abandonment. Exploring whether abandonment is predictable in this context could have interesting downstream benefits for media creators, such as the real-time monitoring of a user's likelihood to abandon through an experience. But also, if there is an association between the interactions of users and abandonment, then it may facilitate a deeper understanding in the future as to whether it was good or bad abandonment.

Much like modelling engagement in the previous chapter, approaches that have been successfully deployed in other domains are largely untested with interactive media experiences and as this form of media can be more complex compared to traditional media, it is unclear how well the approaches translate. Additionally, how abandonment is represented poses a challenge. Grouping users into two groups - abandoned and completed - may not provide sufficient granularity in the context of a multi-story interactive media experience. Associ-

ating behavioural proxies to abandonment should provide a more refined and subtle measurement for media creators to deploy. Using this as motivation, this chapter presents an exploration of the wider dataset collected from the Click TV show to address the following objective: *to investigate whether there are signals in interaction data that are predictive of abandonment*. To direct the investigation, the following research questions are explored:

A-RQ1: Can abandonment in interactive media experiences be modelled using interaction data?

A-RQ2: What interaction-derived metrics are important when inferring abandonment?

A-RQ3: Is there a relationship between abandonment and user engagement?

In answering the questions, the analysis focuses on the two main sub-stories of the Click TV show rather than the experience as a whole. The reason for this is to simplify the analysis; one story chosen by the audience member, do they reach the end? Following the motivating example described previously, identifying *why* someone has left part way through a sub-story should provide a more detailed picture of abandonment than simply whether they reached the end of the whole experience.

4.1 Method

To investigate whether there are signals in interaction data that are predictive of abandonment, the dataset collected from the national release of the Click TV show is used - introduced in the previous chapter.

The data is split according to which story people chose to view at the first choice and contains the interactions of audiences in both sub-stories. Then, models of abandonment are explored to predict the abandonment likelihood of audience members. Each part of the investigation is presented in more detail in this section, first focusing on the dataset, how the interaction metrics were created, and how abandonment was defined as a target for modelling. Following from there, the modelling approach and measuring importance is presented, all of which aid in characterising the behaviour of audience members who abandon.

4.1.1 Data

The dataset used is from the wider dataset collected from the national release of the Click TV show, which includes both those that took part in the engagement study (presented in the previous chapter, see Section 3.2) and those that did not. The specific subset of data used is from the two main sub-stories in the show, which form a substantial part of the narrative: technology use in Malawi (Malawi) and autonomous vehicles (Cars), shown in Figure 1.2. The show allowed people to view one or both stories, and in either order. All users who started a sub-story are included and there is no differentiation according to what they

had done earlier in the show. Audience members that chose to take part in the engagement survey are separated from the dataset for later analysis. Approval for the use of data was sought from the audience using a privacy notice which is presented at the start of the experience.

Interaction Metrics

The same process used in the previous chapter to create interaction metrics is followed – making use of the custom interaction metrics library developed to work with data collected from interactive BBC experiences. Table 4.1 presents all the metrics extracted from the interaction data. To make the analysis and results more applicable to other experiences where different actions are collected but can be grouped into the same categories, these categories are also shown in Table 4.1. The temporal and pre-sub-story metrics are not included in these groupings.

Creating an Abandonment Target

Behind the scenes, the Click TV show is a graph-based structure where nodes are narrative elements and edges are choices/paths through the narrative (see Figure 4.1). To create an abandonment target, the graphical structure is leveraged to focus on predicting how far the audience made it through the story rather than a binary complete/non-complete target. For each audience member in the dataset, the number of narrative elements away from each of the endpoints in the sub-story is counted and the smallest value is taken, i.e., the shortest path to the end of the sub-story. There can be multiple endpoints in both sub-stories, which the user visits depends on their choices throughout the narrative, this is shown in 4.1 where the green coloured nodes are endpoints in both sub-stories. Creating a target metric that is a notion of how far the audience were from the end could provide deeper insight into the causes behind the audience leaving; did they leave straight away or half-way through? Additionally, a numerical target instead of a categorical target allows for an assessment of the relative impact of each feature on the prediction; does an increase in an interaction metric push the prediction in a negative (closer to completion) or positive (abandoning earlier) direction?

4.1.2 Analysis

Statistical Analysis

To discover relationships between interaction metrics and abandonment, as well as between interaction metrics themselves, an exploration of the data was performed. Correlation analysis was carried out to discover relationships and statistical difference testing was performed to find differences between abandonment distances. As the abandonment metric is ordinal,

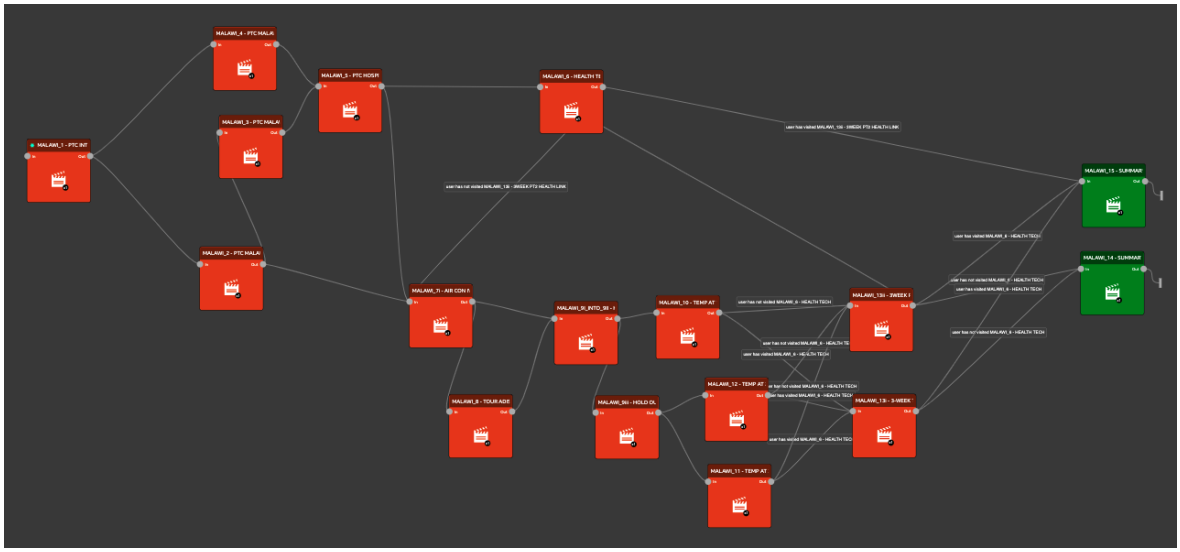
Table 4.1. The interaction metrics extracted from the interaction data collected in the Click TV Show. Includes the how the different metrics are grouped together.

Grouping	Interaction Metric
Navigation - Backwards	# Back Button # Seek Backwards
Navigation - Backwards (Proportion)	Back Button (Proportion) Seek Backwards (Proportion)
Navigation - Forwards	# Next Button # Seek Forwards
Navigation - Forwards (Proportion)	Next Button (Proportion) Seek Forwards (Proportion)
Contextual Changes	# Window Orientation Changes # Fullscreen
Contextual Changes (Proportion)	Window Orientation Changes (Proportion) Fullscreen (Proportion)
General Video Controls	# Play/Pause # Volume Changes # Video Scrubs
General Video Controls (Proportion)	Play/Pause (Proportion) Volume Changes (Proportion) Video Scrubs (Proportion)
Visibility Changes	# Browser Visibility Changes
Visibility Changes (Proportion)	Browser Visibility Change (Proportion)
	Average NEC Time (Pre) Norm Average NEC Time (Pre) # Link Choices (Pre) # Narrative Element Changes (Pre) Link Choice Ratio (Pre) Vertical Orientation (Pre) Average NEC Time Normalised Average NEC Time # Short Pauses # Medium Pauses # Long Pauses # Very Long Pauses Mobile Device (Inferred)

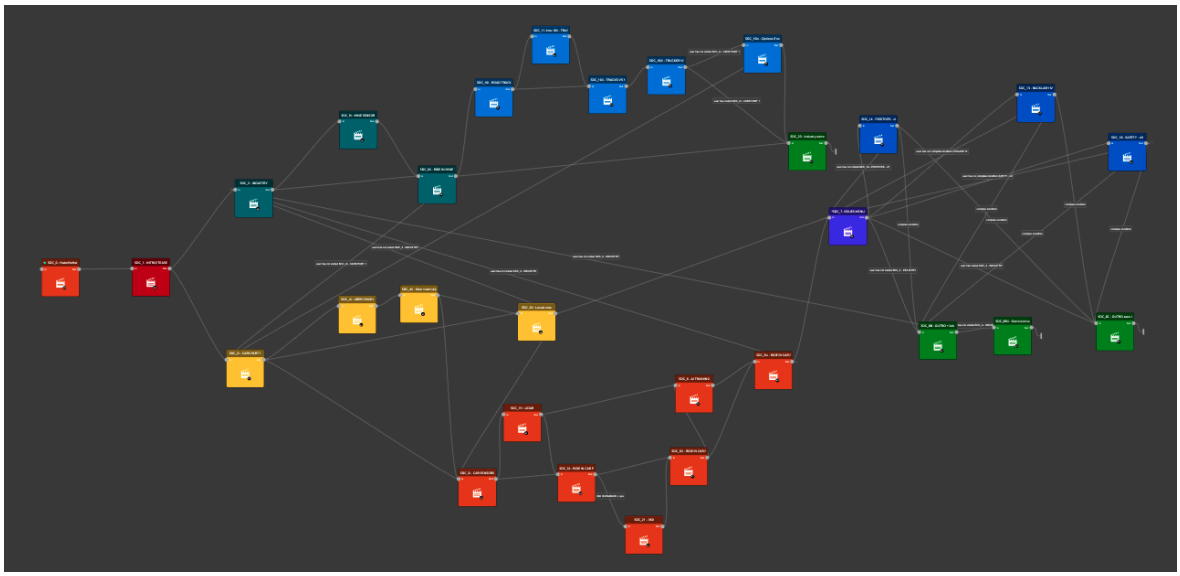
Spearman's rank correlation coefficient was used to measure the correlation between interaction metrics and abandonment and between the interaction metrics. In the case where the interaction metric was a boolean, point-biserial correlation coefficient was calculated. To test for significant differences between metrics and abandonment, the Kruskal Wallis test – used when there is a single independent variable (the metric) with two or more levels and the ordinal dependent variable (abandonment) – was applied.

Modelling Abandonment

The abandonment target is distance-based and produces a non-negative count-based metric where there is a maximum shortest path for each sub-story (Malawi = 4 and Cars = 6). As such, the modelling techniques applied to the data to predict abandonment are Poisson regression models. A Poisson model works under the assumption that the target variable



(a) Malawi



(b) Cars

Figure 4.1. The graphical structure of both the Malawi and Cars sub-stories

has a Poisson distribution and are used in modelling count data. Models are trained for each sub-story and are as follows: Poisson Generalised Linear Model, Decision Tree, and Gradient Boosted Regression Tree (with histogram-based learning). For the Decision Tree (DT) regressor, Poisson deviance criterion was used to evaluate the quality of split. A Gradient Boosted Regression Tree with histogram-based learning (HGBR) is a tree-based model that when building trees, rather than finding splitting points on the sorted input data (memory intensive), places continuous features into bins and uses the bins to create feature histograms which provides a coarse approximation of the input data (Ke et al., 2017). For this model, the loss function used is Poisson deviance.

To evaluate the models and find which to take forward in the analysis, 10-fold cross-validation is performed using the mean Poisson deviance as the performance metric to minimise (the mean squared error (MSE) and mean absolute error (MAE) are also reported in the results). Each model is compared to each of the other models, and a baseline mean regressor, which predicts the mean value of the input data, using a paired t-test. Once the best performing

model was identified, an optimisation step was performed to find the optimal version of the model using a standard train-test set split (80:20) of the data. The Tree-structure Parzen Estimator (Bergstra et al., 2013; Bergstra et al., 2011) sampler is used which has been shown to outperform other optimisation techniques such as Bayesian optimisation, with pruning performed using Hyperband (L. Li, Jamieson, et al., 2017), running for 150 trials and minimising the Poisson deviance.

Measuring Importance

In **A-RQ2**, the focus is on what interaction metrics are important when inferring abandonment. To investigate this, a range of model interpretation techniques are applied to the optimised model for each sub-story. Permutation importance, Accumulated Local Effects (ALE), and Shapley Additive Explanations (SHAP) are calculated and estimated from the model and each provide different views on the model's decision-making. All importance measures are performed using the hold-out test set.

Permutation importance evaluates the relationship between the features and the target with a measure of how much the performance metric decreases when a single feature is randomly shuffled (Breiman, 2001). The mean Poisson deviance is used as the performance metric and each feature is shuffled 10 times for a range of performance results. Accumulated Local Affects (Apley and Zhu, 2020) provides a measure of the impact on the prediction for given feature values and are an alternative to a Partial Dependence Plot (Friedman, 2001). Shapely Additive Explanations (SHAP) measure how individual features contribute to the model's prediction (Lundberg and Lee, 2017). All these approaches are model-agnostic and work with black-box models, where there is no easily accessible or interpretable internal feature representation.

Abandonment & Engagement Relationship

To investigate the relationship between abandonment and engagement (**A-RQ3**), the interaction metrics for the subset of the audience that took part in the engagement survey are removed for the analysis presented above and a predicted distance is produced by the models for each sub-story. It is worth noting that the subset who took part in the engagement survey, by definition, will not have abandoned the show. The predicted distance is tested for normality using the Shapiro-Wilk test (with $\alpha = .05$); if the predicted distance is non-parametric then the Mann-Whitney U test is performed between the high and low engagement groups, with the common language effect size (f). If the distributions are parametric, then a one-way ANOVA is performed between the two groups. A difference in the predicted distance between high and low engagement would suggest that there is a relationship between the engagement of users and their perceived likelihood to abandon the experience.

4.2 Results

4.2.1 Sample

In total, 61,724 (Malawi = 21,169 and Cars = 40,555) users visited one of the two sub-stories (3,717 visited both, 17,452 only visited Malawi, and 36,838 only visited Cars). For the Malawi sub-story, a total of 1,472,865 events were captured, with users recording an average of 69.57 (STD = 265,05) events. Whilst, for the Cars sub-story, 3,226,503 events were recorded and users had an average of 79.55 (STD = 371.02) interaction events. The larger number of users visiting the Cars sub-story could be due to it being positioned before Malawi on the default path through the story or it could be a more popular topic.

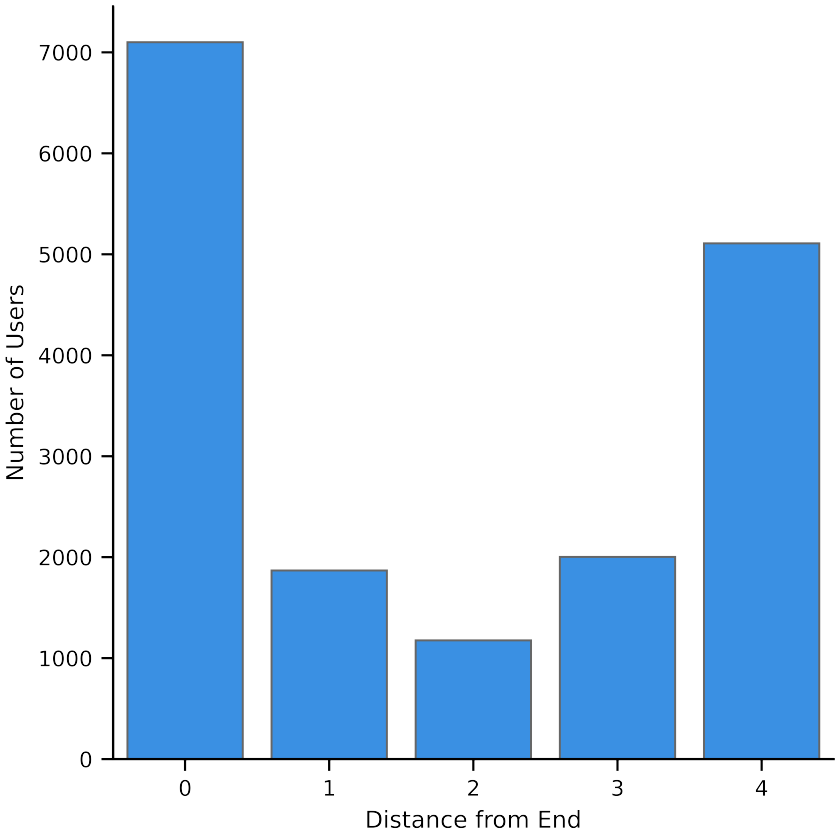
Abandonment Metric

Most users that visited both sub-stories reached the end and completed, as shown in Figure 4.2 where zero indicates completion. Abandoning in Malawi tended to occur most frequently at the start of the sub-story (the furthest point from the end, see Figure 4.2a), with a moderate rate of abandonment as the story progresses. A similar pattern is observed in the Cars sub-story, see Figure 4.2b, with a peak of users abandoning three or four nodes away from the end, which is the start of the sub-story.

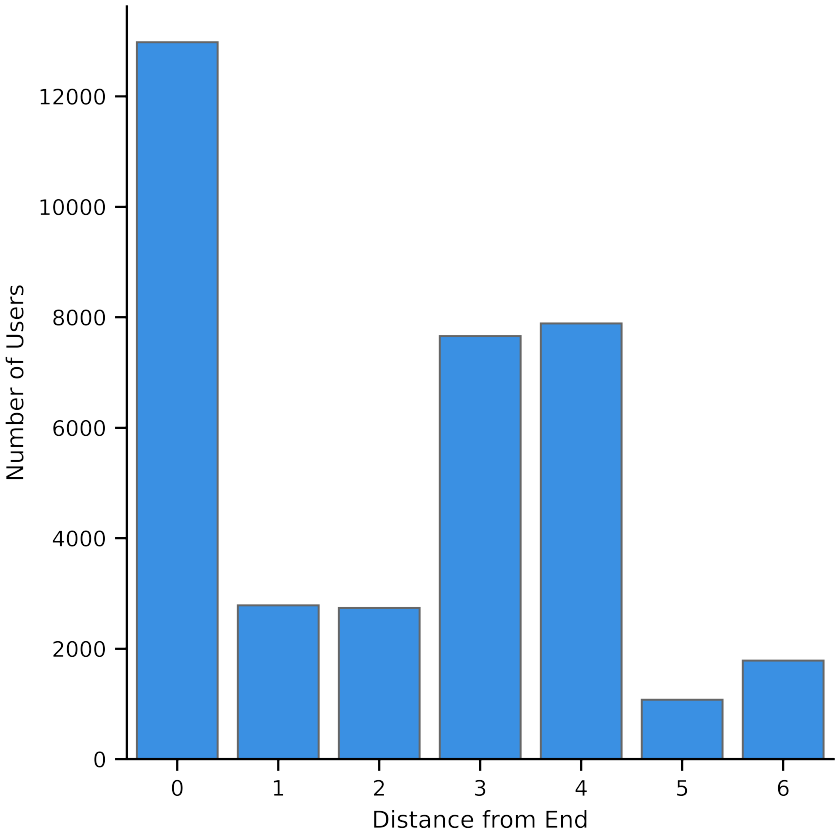
However, there is a specific limitation in the Cars sub-story. Due to the sub-story structure, a user can be further into the story but abandon with a larger minimum shortest path – there are more direct paths to the end of the sub-story at the beginning than part way through one segment (see Figure 4.1b). Experimentation was performed with alternative measures, such as weighting the shortest path by the number of narrative elements seen, but interpretability is lost, and as this limitation only affects the Cars sub-story and as there are a small number of users that fall into this category, this metric was taken forward and used for the remaining analysis.

4.2.2 Statistical Analysis

Across both sub-stories, there were a small number of interactions recorded by the users. For example, Table 4.2 shows that 14.47% (Cars) and 16.06% (Malawi) of users recorded at least one general video control event. The rates of interactions across both sub-stories were similar, with only minor differences in most cases. There was a noticeable difference in the use of forward navigation by users in the Malawi sub-story, with 16.1% of users recording these events compared to 12.3% in the Cars sub-story. An almost equal split between mobile and non-mobile users was observed – which is an inferred metric from the interaction data and is not explicitly captured – and the majority of users watched the show in horizontal mode.



(a) Malawi



(b) Cars

Figure 4.2. The distance from the end distributions for the Malawi and Cars sub-stories, calculated as per the description in Section 4.1.1. Zero means that the user reached the end of the sub-story.

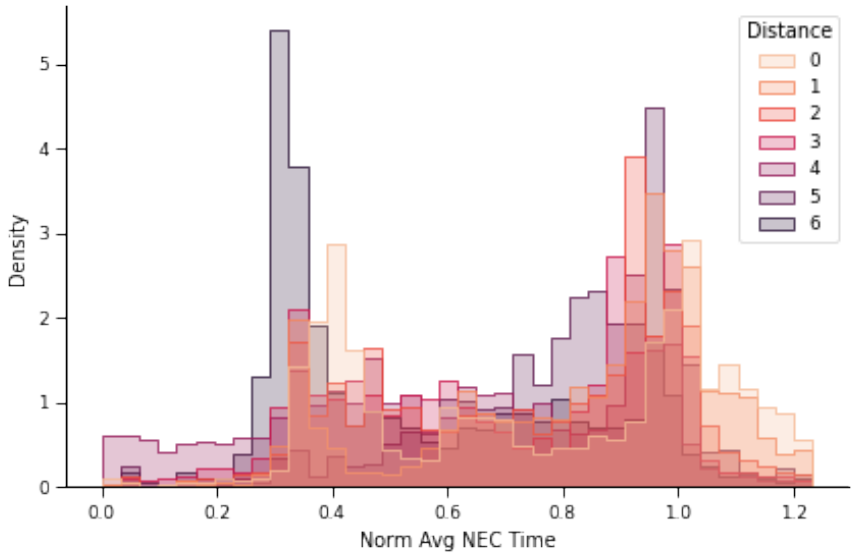
Table 4.2. Descriptive statistics for the interaction metrics derived from interaction data collected as part of the Click experience. The proportion of non-zero statistic describes the percentage of users that have registered at least one of the events.

	Cars				Malawi			
Interaction Metric	Mean	Std. Dev.	Median	Proportion non-zero	Mean	Std. Dev.	Median	Proportion non-zero
# Contextual Changes	0.74	1.72	0.00	37.76%	0.67	1.74	0.00	35.64%
# General Video Controls	0.40	1.89	0.00	14.47%	0.53	2.05	0.00	16.06%
# Visibility Changes	0.93	1.86	0.00	47.60%	0.89	1.63	0.00	47.72%
# Short Pauses	1.49	2.62	0.00	51.40%	1.56	2.71	1.00	53.56%
# Medium Pauses	0.79	1.45	0.00	40.18%	0.80	1.50	0.00	39.28%
# Long Pauses	0.37	0.79	0.00	26.33%	0.50	0.81	0.00	36.62%
# Very Long Pauses	3.40	3.05	3.00	86.21%	2.62	2.17	2.00	84.96%
Navigation - Backwards	0.32	2.11	0.00	5.33%	0.17	1.56	0.00	3.22%
Navigation - Forwards	0.67	2.93	0.00	12.34%	1.00	3.62	0.00	16.13%
# Link Choices (Pre)	7.13	6.22	5.00	97.24%	8.11	6.58	6.00	100%
# Narrative Element Changes (Pre)	6.04	0.48	6.00	100%	6.04	0.57	6.00	100%
Link Choice Ratio (Pre)	1.18	1.02	0.83	-	1.34	1.09	1.00	-
Avg NEC Time (Pre)	55.00	11.75	55.38	-	55.58	12.28	55.57	-
Norm Avg NEC Time (Pre)	0.71	0.13	0.75	-	0.72	0.12	0.75	-
Norm Avg NEC Time	0.70	0.29	0.72	-	0.66	0.30	0.69	-
Avg NEC Time	48.01	28.45	43.71	-	56.52	32.91	43.47	-
Navigation - Backwards (Proportion)	0.009	0.05	0.00	-	0.005	0.03	0.00	-
Navigation - Fowards (Proportion)	0.03	0.10	0.00	-	0.04	0.12	0.00	-
Contextual Changes (Proportion)	0.06	0.10	0.00	-	0.06	0.11	0.00	-
General Video Controls (Proportion)	0.01	0.06	0.00	-	0.02	0.08	0.00	-
	Mobile		Non-Mobile		Mobile		Non-Mobile	
Mobile (Inferred)	55.26%		44.73%		54.01%		45.98%	
	Horizontal		Vertical		Horizontal		Vertical	
Vertical Orientation (Pre)	98.93%		1.06%		99.59%		0.40%	

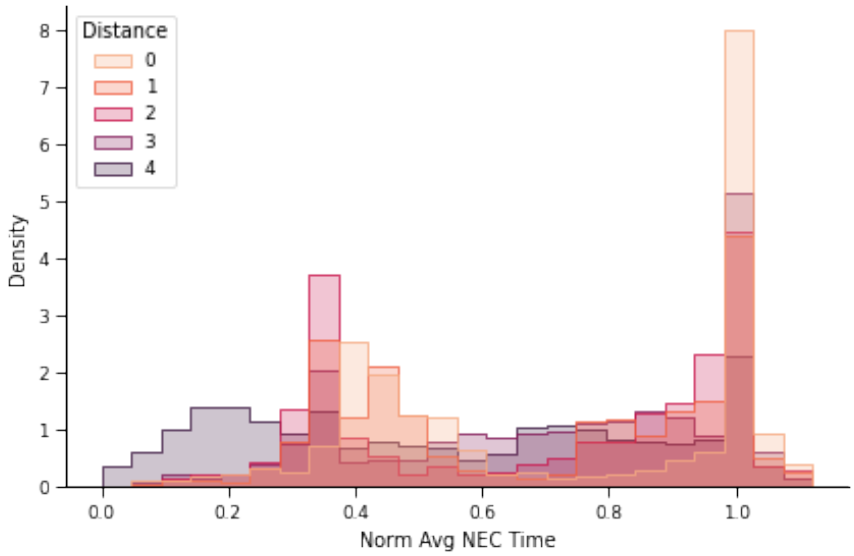
Figure 4.3a points towards users that do not reach the end of the Cars sub-story recording a low average time spend on narrative elements. A similar pattern is observed in the Malawi sub-story (Figure 4.3b), with users that abandon early recording low numbers for the average time spent on narrative elements. In contrast, users that reached the end or close to the end tend to spend more time on narrative elements (shown by users going beyond 100%), this could be due to revisiting narrative elements – this is not observed in the Malawi sub-story.

Minor correlations between interaction metrics and abandonment are shown in Figure 4.4. The proportion of visibility changes in both sub-stories is positively correlated with distance, indicating that a higher proportion of visibility changes is associated with leaving the experience earlier; this is also demonstrated in the number of visibility change events. The number of very long pauses has a negative correlation with distance, in both sub-stories, which suggests that users who interact little and watch content either abandon later or complete.

Testing for significant differences between interaction metrics and abandonment produced significant results across all interaction metrics except the number of narrative elements seen prior to the sub-story. Figure 4.5a shows effect plots for the Cars sub-story, while Figure 4.5b shows the same plots but for the Malawi sub-story. There are notable differences in some of the interaction metrics across both sub-stories, for example, users that complete record a significantly lower number of contextual changes compared to users that abandoned. This is also true for the number of visibility changes; however, the scale on which these differences are found is small. For example, the scale for the number of contextual changes is between 0.5 – 1.2 (for the Cars sub-story), means that most users record a small



(a) Cars



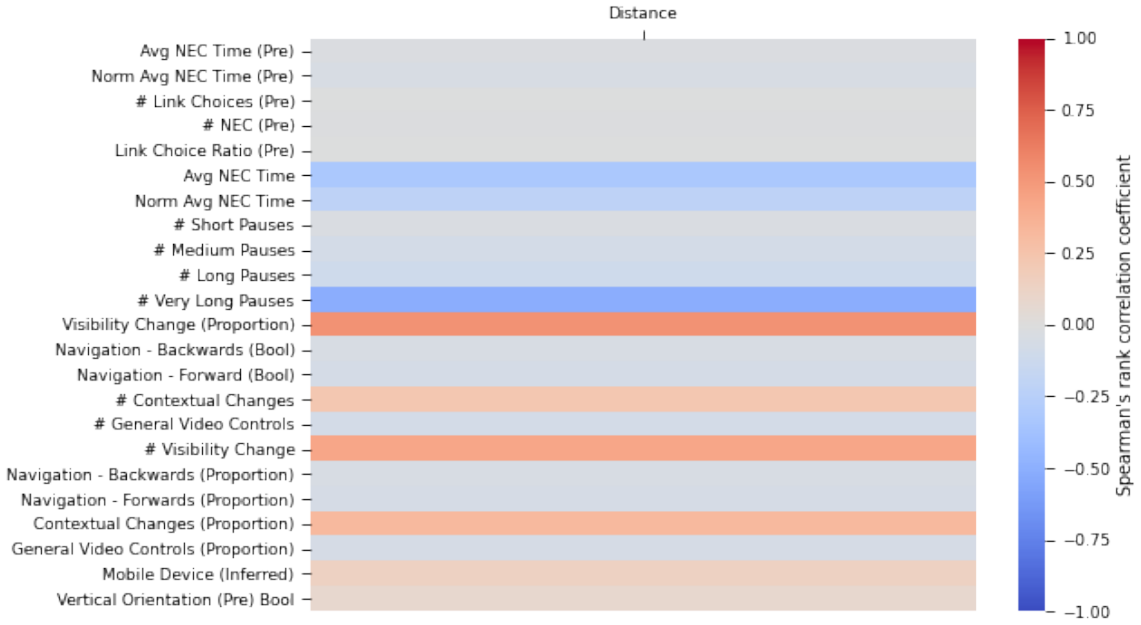
(b) Malawi

Figure 4.3. The distribution of the normalised average narrative element time across abandonment distances for both sub-stories. A normalised average time of 1.0 indicates that a user consumed, on average, 100% of the narrative element. The plot demonstrates, in the case for the Cars sub-story, that a high proportion of users with lower distances consume more than 100% of narrative elements on average, perhaps through revisiting

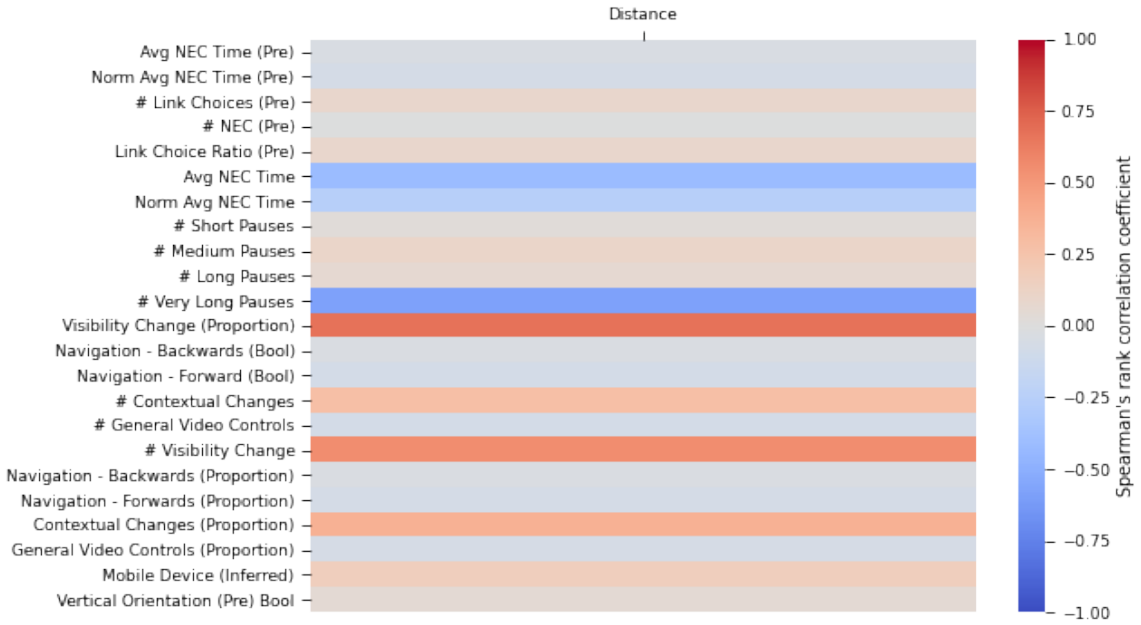
number of these events and in most cases, none - shown in Table 4.2.

4.2.3 Modelling Abandonment

When predicting abandonment, the three models trained all significantly outperformed a dummy regression model – which constantly predicts the mean value of the target – as shown in Table 4.3. The Gradient Boosted Regression Tree with histogram-based learning (HBGR) model performed best compared to all other models across all evaluation metrics, significantly outperforming the Poisson Regression (Malawi: $t(15749) = -46.87, p < .001$; Cars: $t(33684) = -41.15, p < .001$) and Decision Tree (Malawi: $t(15749) = -16.15, p < .001$; Cars: $t(33684) = -23.19, p < .001$) models in both sub-stories. As a result, the



(a) Cars

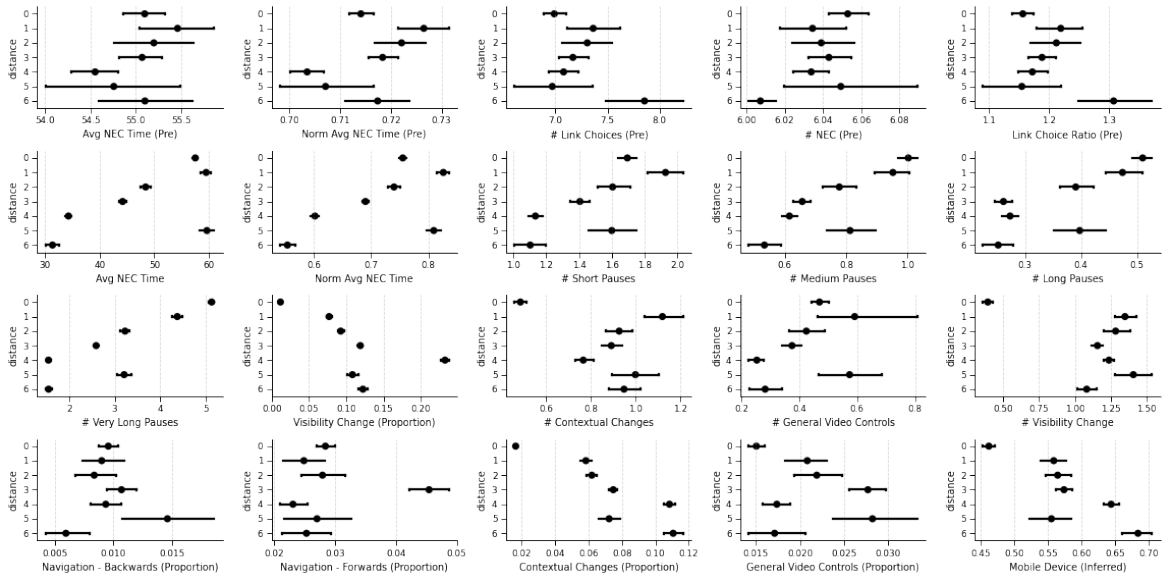


(b) Malawi

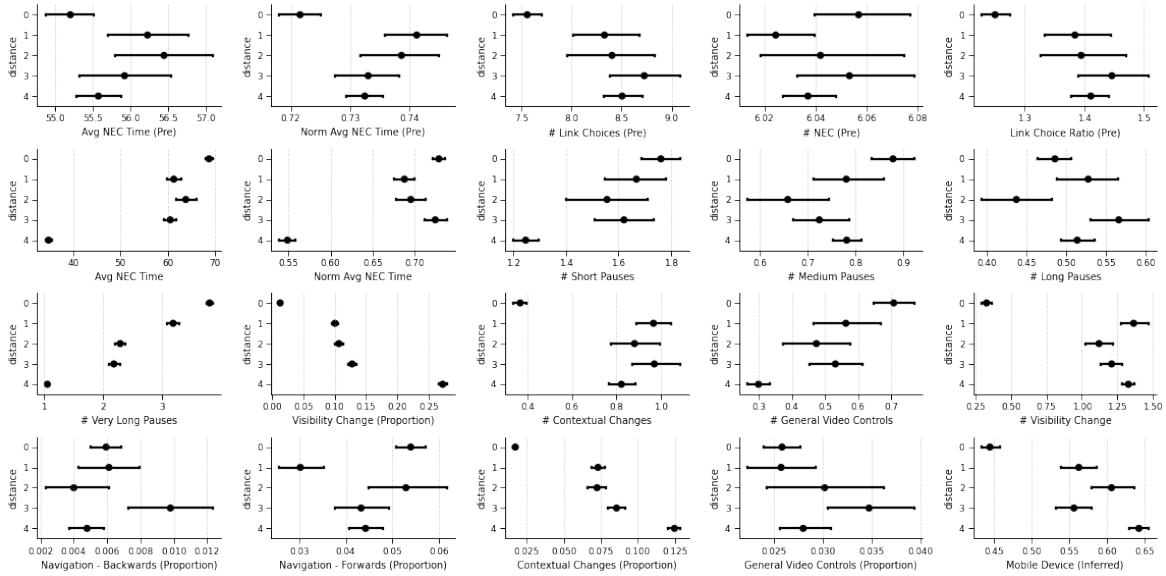
Figure 4.4. The correlations between interaction metrics and abandonment distances. Correlations were calculated using Spearman's rank correlation coefficient.

HGBR model was taken forward and used in the rest of the investigation.

To tune the model to the data, an optimisation step was performed and the following hyperparameters were tuned: learning rate, maximum number of trees, leaf nodes, and depth for each tree, the minimum number of samples per leaf, L_2 regularisation, and tolerance. A model was optimised and trained for each sub-story using a Poisson deviance loss function, and the optimal hyperparameter configuration for Malawi ($PD = 0.39$, $MSE = 0.49$, $MAE = 0.40$) was: learning rate $= 0.09$, maximum number of trees $= 180$, leaf nodes $= 220$, and depth $= 17$, minimum number of samples per leaf $= 18$, L_2 regularisation $= 90.74$, and tolerance $= 1.08^{-05}$. Whilst the optimal configuration for Cars ($PD = 0.67$,



(a) Cars

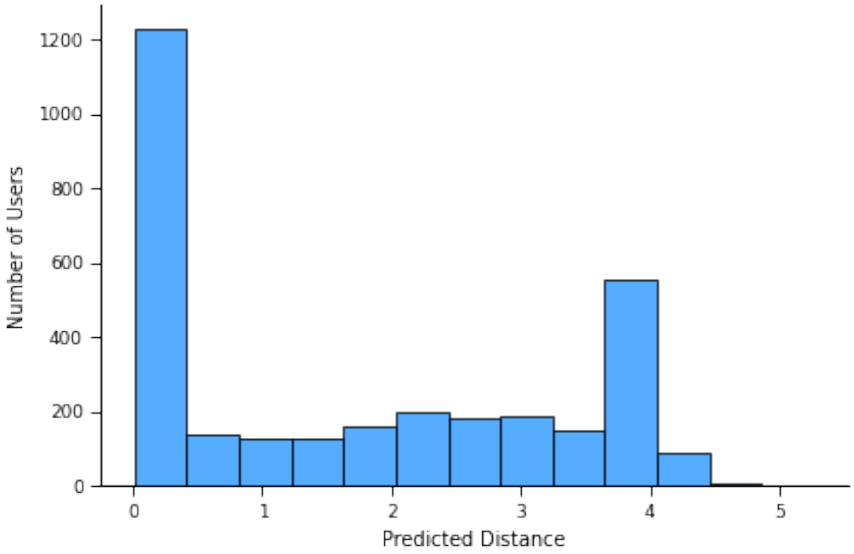


(b) Malawi

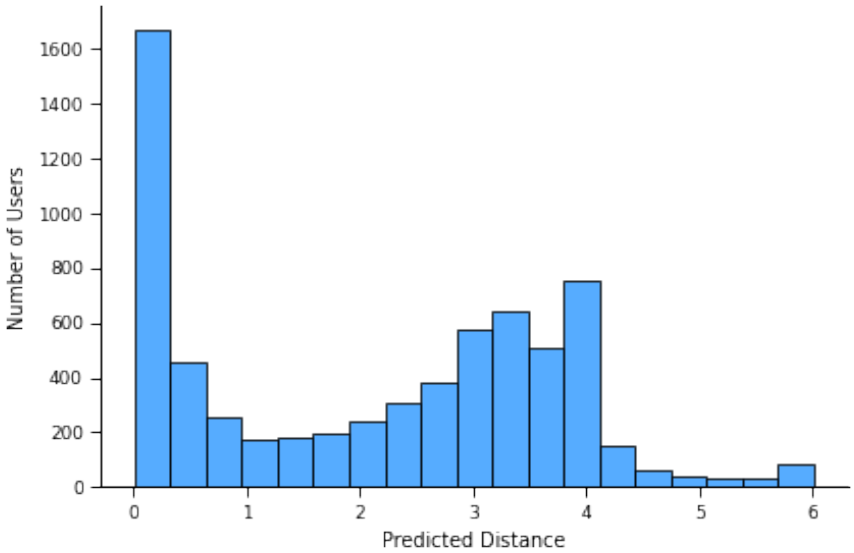
Figure 4.5. The significant differences between the interaction metrics and abandonment distances shown through effect plots

Table 4.3. The results of initial modelling. Significance is determined by a paired t-test (significance level: ***= $p < .001$, *= $p < .05$) between the Dummy Regression and the three other models and is reported based on the Poisson Deviance (PD), with Mean Squared Error (MSE) and Mean Absolute Error (MAE) also reported.

	Malawi			**Cars**		
	PD	**MSE**	**MAE**	**PD**	**MSE**	**MAE**
Dummy Regression	2.20 (0.14)	3.02 (0.03)	1.63 (0.01)	2.20 (0.18)	3.56 (0.08)	1.69 (0.03)
Poisson Regression	1.32 (0.10)***	1.51 (0.04)	1.06 (0.01)	1.61 (0.11)***	2.37 (0.05)	1.29 (0.01)
Decision Tree	0.88 (0.07)***	1.21 (0.07)	0.52 (0.01)	1.44 (0.08)***	2.26 (0.11)	0.85 (0.02)
HGBR	**0.42** (0.05)***	**0.52** (0.04)	**0.40** (0.02)	**0.67** (0.06)***	**0.95** (0.04)	**0.62** (0.01)



(a) Malawi



(b) Cars

Figure 4.6. The predicted distance distributions for the Malawi and Cars sub-stories. The plots demonstrate that the real distance is broadly captured by the model.

$MSE = 0.95$, $MAE = 0.64$) was: learning rate $= 0.04$, maximum number of trees $= 183$, leaf nodes $= 106$, and depth $= 18$, minimum number of samples per leaf $= 17$, L_2 regularisation $= 5.66$, and tolerance $= 5.14^{-05}$.

Figure 4.6 shows the distributions of the distances predicted by the Malawi and Cars models and demonstrates that the distributions are of a similar shape, including the same peaks, to the actual distances (shown in Figure 4.2). These results suggest that the models can capture the distribution of distances and that abandonment can be modelled and accurately predicted using interaction data collected from an interactive media experience.

4.2.4 Feature Importance

Permutation Importance

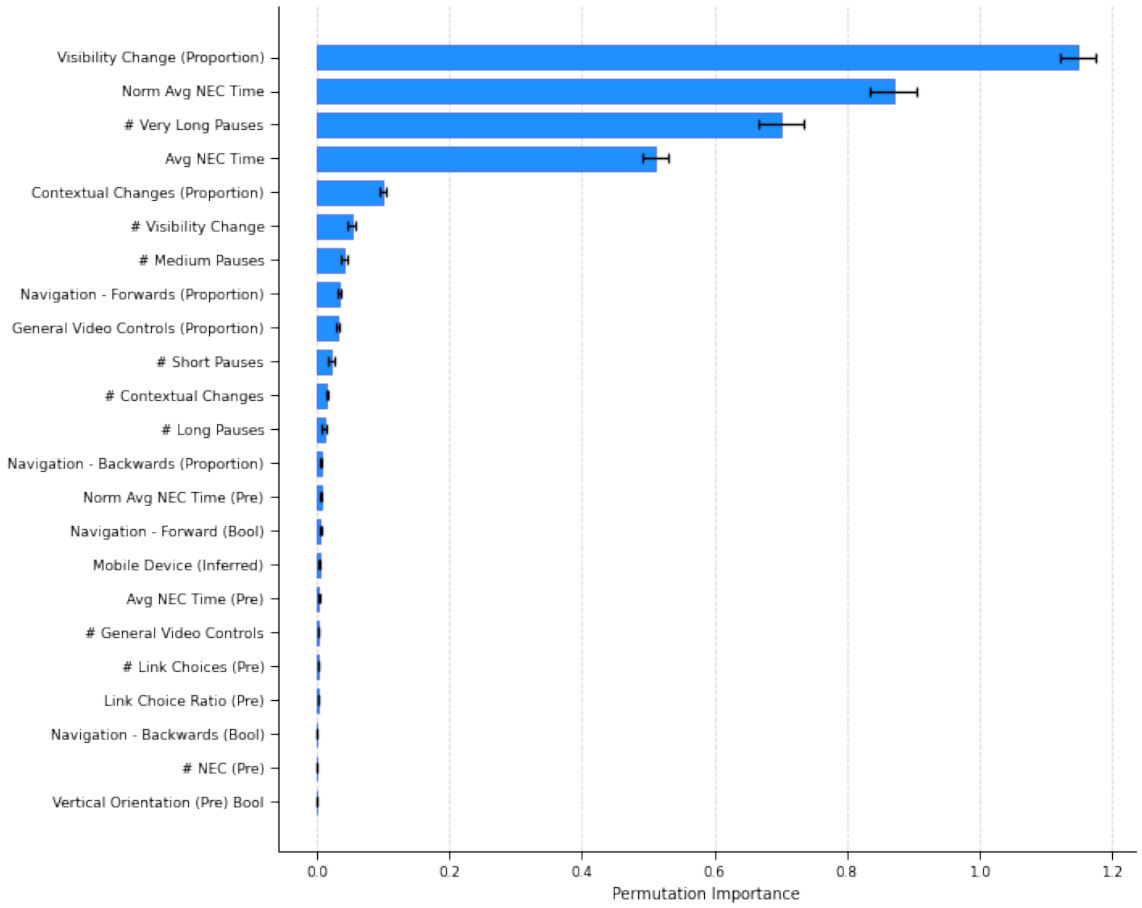
The results of performing permutation importance are presented in Figure 4.7, which shows that the proportion of visibility changes was the most important interaction metric across both sub-stories with little variation between tests. The normalised and non-normalised average narrative element time had differing levels of importance, with the model for Malawi placing more importance on the former and the Cars model on the latter. The number of very long pauses, a metric representing pauses between interaction events, is important for both sub-stories, further demonstrating the reliance on temporal interaction metrics. The proportion of contextual changes and the number of browser visibility changes have a moderate level of importance in the accurate prediction of abandonment in both sub-stories, with the remaining metrics contributing a smaller effect to the prediction and registering a smaller importance.

Accumulated Local Effects

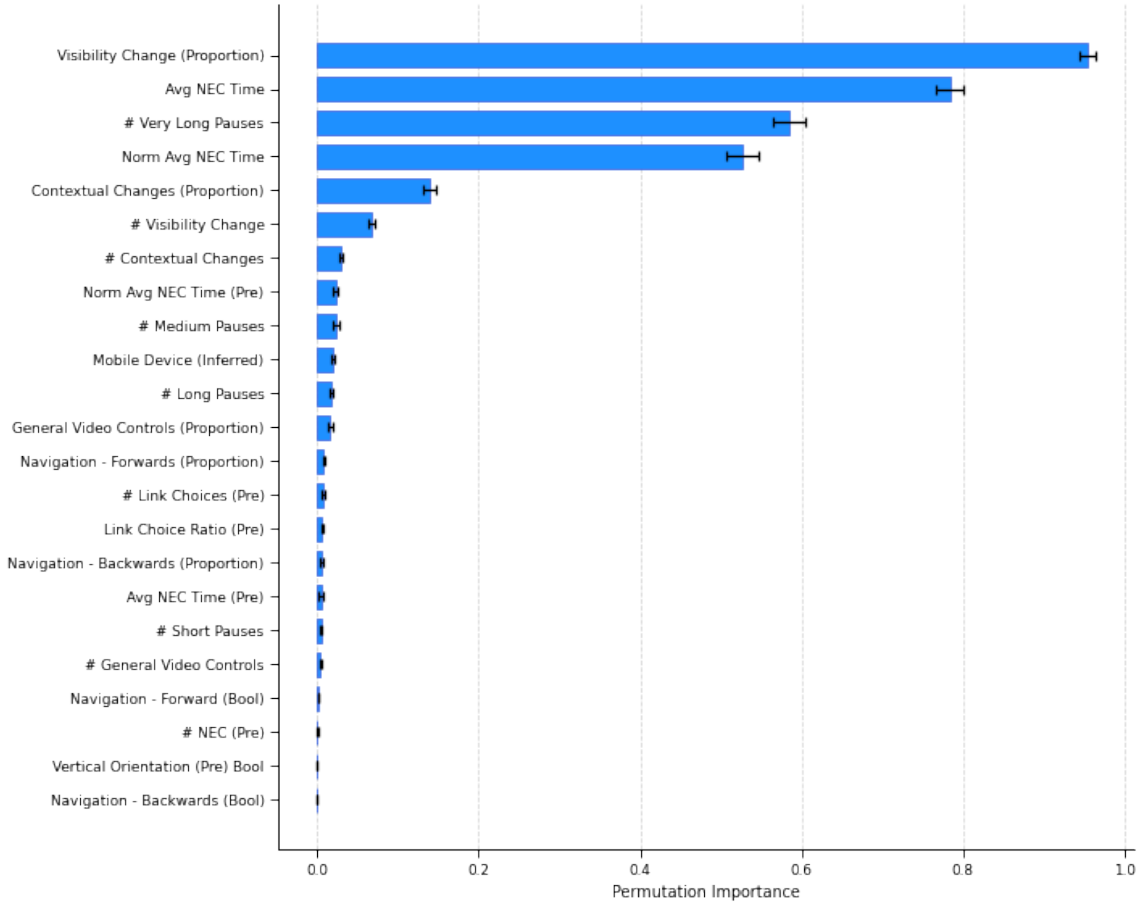
While the previous measure of importance provides an idea of what metrics the models rely on to make accurate predictions, accumulated local effects provide a measure of how different metric values effect the prediction, providing a notion of effect direction for metric values. Figures 4.8 and 4.9 show the results of calculating the Accumulated Local Effects (ALE) using the Malawi and Car models, respectively – the test dataset was used in both cases. As shown in the figures, interaction metrics both positively and negatively effect the model prediction, and in some cases, have no effect.

The metric with the strongest effects on the prediction in both sub-stories is the proportion of browser visibility changes, shown in Figures 4.10a and 4.10b. As the proportion of visibility changes increases, the effect on the model towards a positive prediction also increases and is shown in both sub-stories, with a stronger effect in the Malawi model. In contrast, there is an opposite effect when considering the number of browser visibility changes, with an increase negatively effecting the prediction of the models.

Shown in Figure 4.11, the proportion of contextual changes – including full screen and orientation change events – has a similar effect to the proportion of visibility changes, and suggests that if users have a higher proportion of these events then they are predicted to be more likely to leave the sub-stories earlier. For the metric that is a count of these events, there is an initial negative effect on the Malawi model (Figure 4.11a) and as the value increases, beyond 10, then the effect becomes positive. Whilst for the Cars model (Figure 4.11b), there is a strong negative effect on the relative prediction which positively increases as the number of contextual changes increases. Much like with the visibility change metrics, there is a distinction between the proportion of these metrics and the number of times they occur, which could demonstrate a difference between user intentions.



(a) Malawi



(b) Cars

Figure 4.7. The results from calculating permutation importance. The scores presented are averages, with the standard deviation shown by the error bares, and are in descending order.

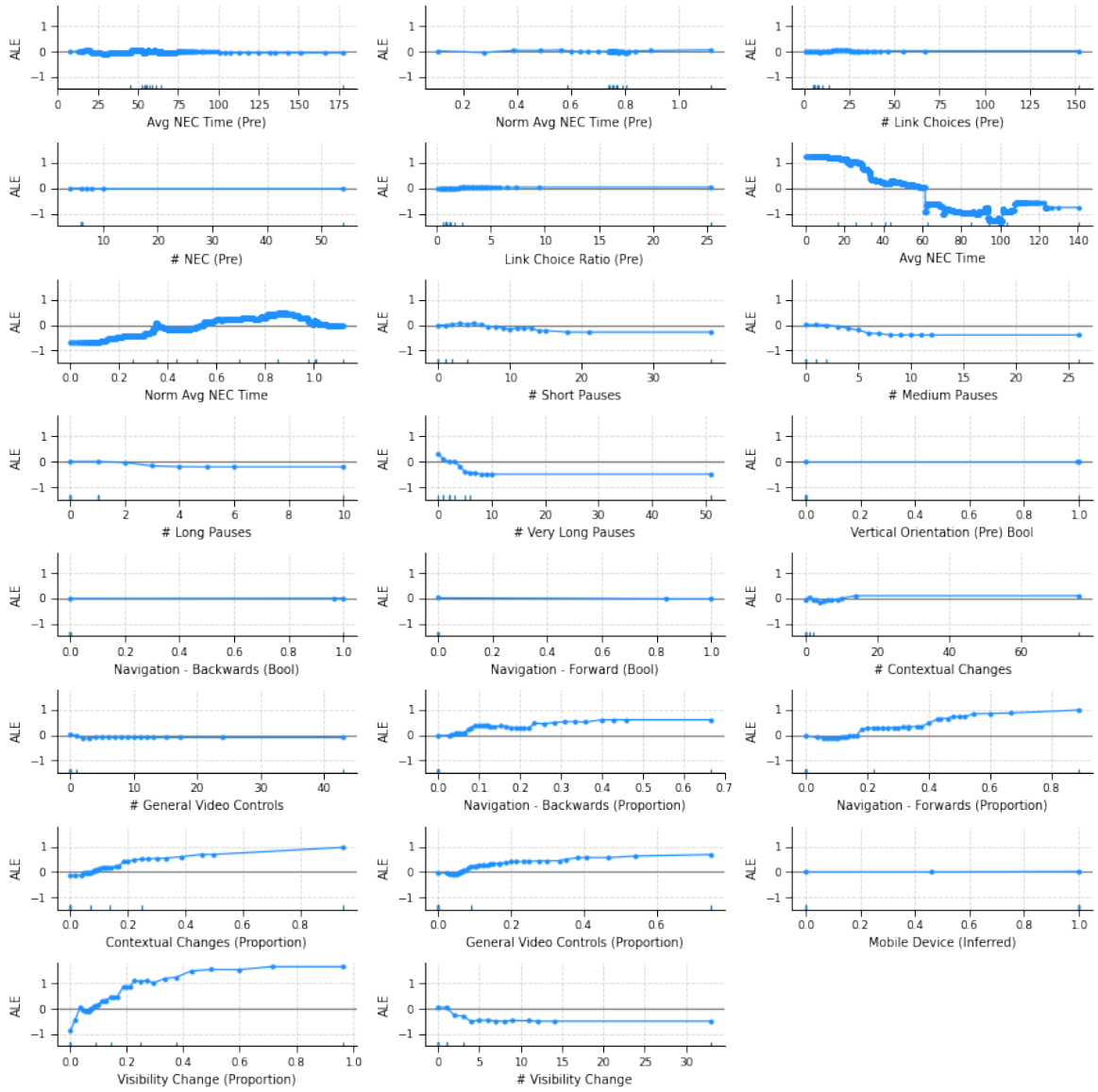


Figure 4.8. The results from calculating the accumulated local effects for all metrics used as input into the Malawi abandonment model.

Pauses are calculated based on the time between interaction events and were previously important in the modelling of engagement. For abandonment and in both sub-stories, a higher number of pauses of any type positively affects the models with varying strength, as shown in Figures 4.12a and 4.12b. An increase in the number of pauses demonstrates that the user is interacting more, in the case of shorter pauses, and interesting less, in the case of longer pauses, and is suggestive of non-abandonment.

For the proportion of backwards navigation events, as shown in Figure 4.13, an increased number positively affects the models – with a strong effect in the Malawi model – suggesting that users who move back through the content are more likely to abandon. Similarly, for Malawi, a large proportion of forward navigation events has a strong positive effect on the prediction, whereas for Cars, there is a minimal, weak effect which changes between positive and negative that suggests a difference between the two sub-stories. The count of these metrics was turned into binary representations due to low numbers and register no effect on the model prediction, as shown in Figures 4.8 and 4.9. Regarding the impact of general

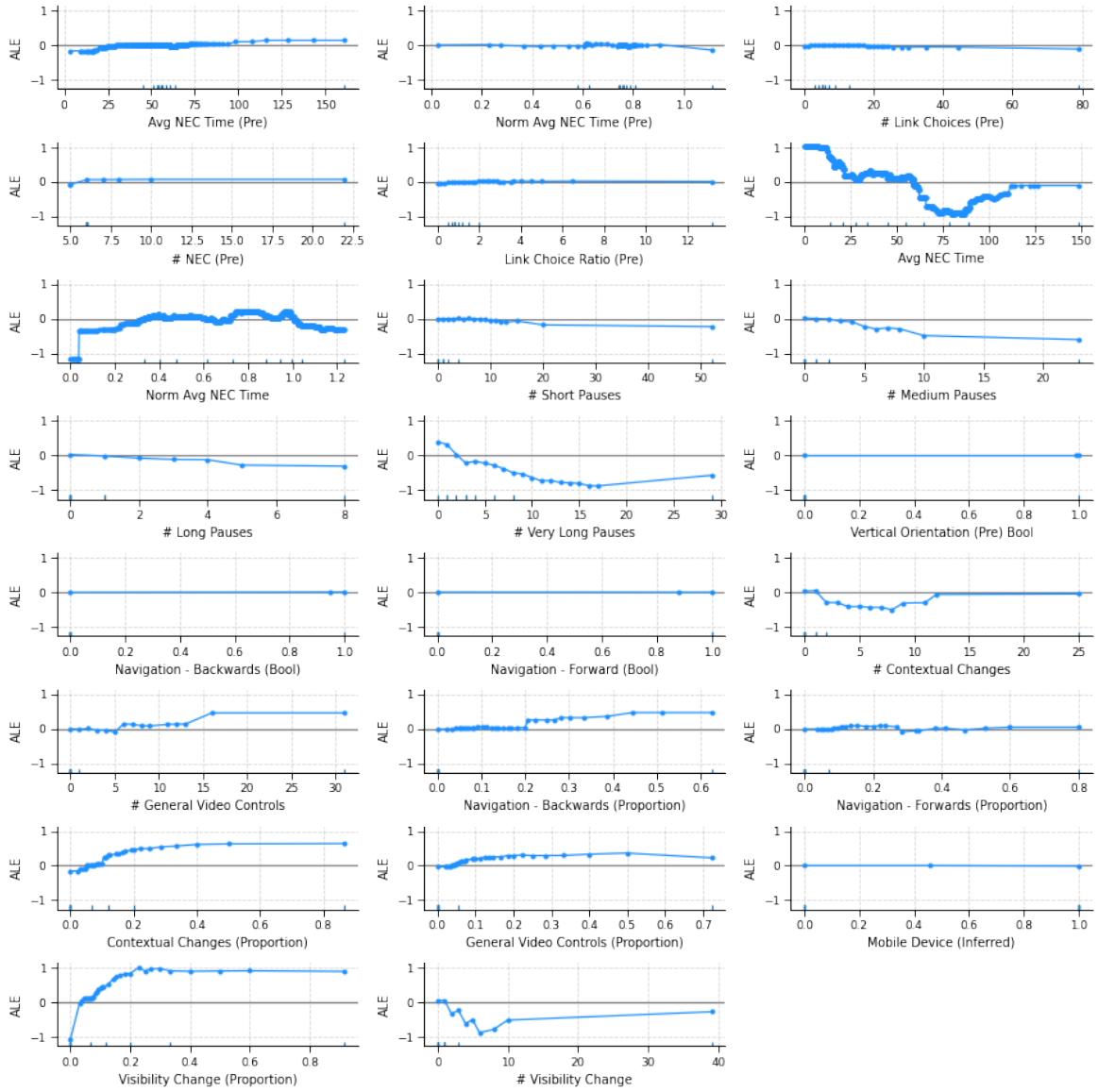


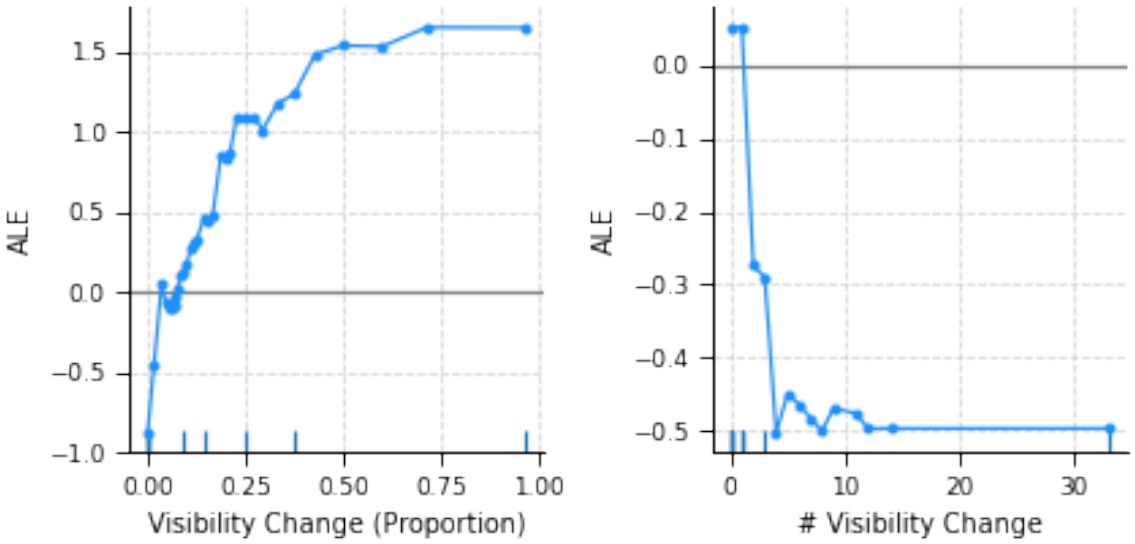
Figure 4.9. The results from calculating the accumulated local effects for all metrics used as input into the Cars abandonment model.

video controls, there is a positive effect (strongest for the Malawi sub-story) on the prediction as the proportion of general video controls increases (see Figure 4.13).

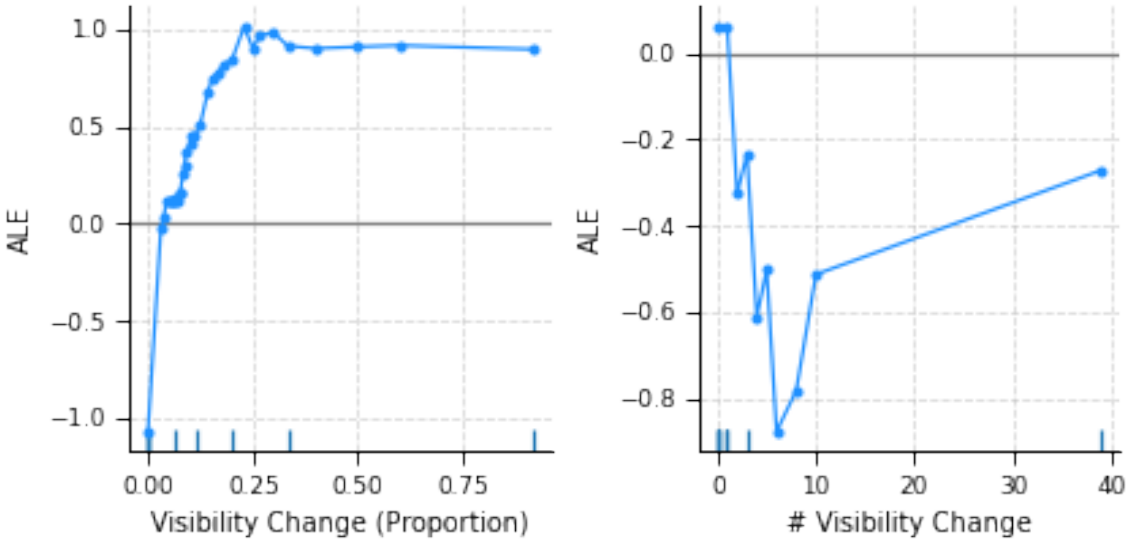
Shapely Additive Values

To further explore how the interaction metrics contribute to the prediction of abandonment, we calculated SHAP values for each of the sub-stories models – the results are shown in Figure 4.14.

The results, which are ordered on the y-axis by their mean SHAP value (the higher magnitude, the more impact on the prediction), show that there are distinct groupings of users for the proportion of visibility changes in both models: low feature values (indicated by the light blue colour) all negatively impact the prediction and pushes it towards lower values – values closer to zero, where zero is zero nodes from the end. In contrast, values with an increasing feature value (indicated by the transition from blue to red) all positively impact the



(a) Malawi

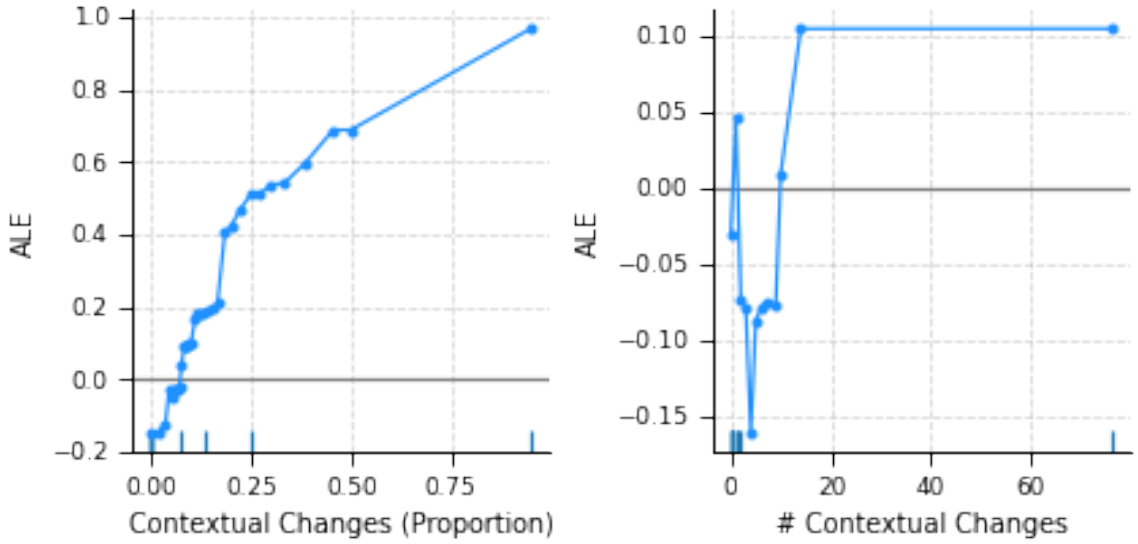


(b) Cars

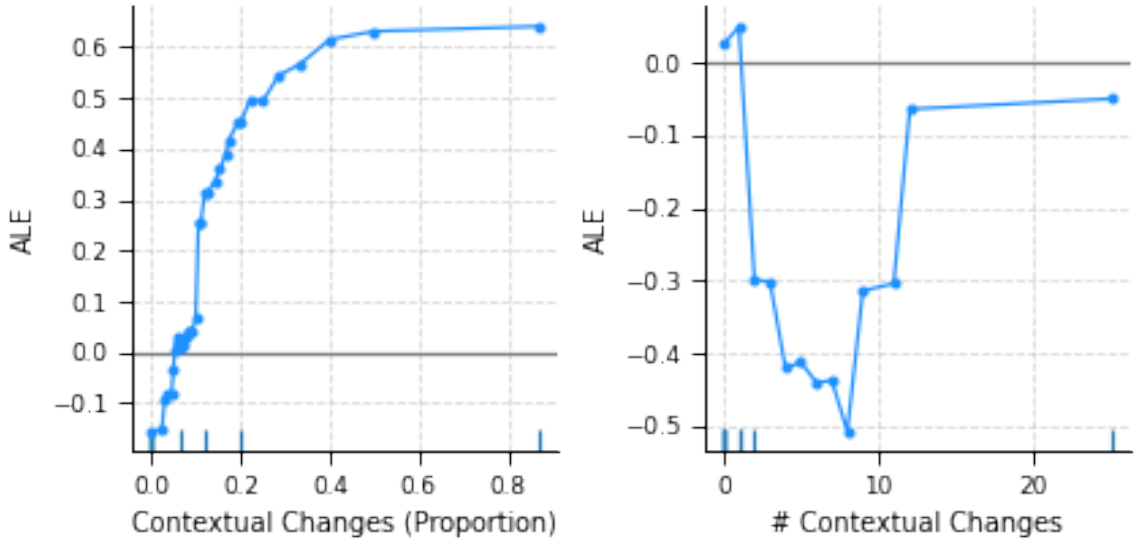
Figure 4.10. The accumulated local effects for the browser visibility change metric (both proportion and counts). The results demonstrate that a high proportion of visibility changes positively effects the model prediction (therefore, weighting the prediction towards higher abandonment distances), while the opposite is true for the counts.

prediction and push the model towards predicting higher values – predicting values further from the end.

The second most important feature for both sub-story models, the number of very long pauses, show that a high number negatively impacts the model and pushes the prediction towards smaller values, corroborating the findings in the previous section (specifically, Figure 4.12). The importance of forward navigation events was evidenced in the application of ALE and is further demonstrated here, with the metric being one of the most important in the prediction of abandonment for the Malawi sub-story but not for Cars. In addition to very long pauses, higher values for the number of medium pauses push the prediction towards lower values for the Cars model, with the metric not registering as important for the Malawi sub-story, again following a similar trend to that presented in the previous section.



(a) Malawi

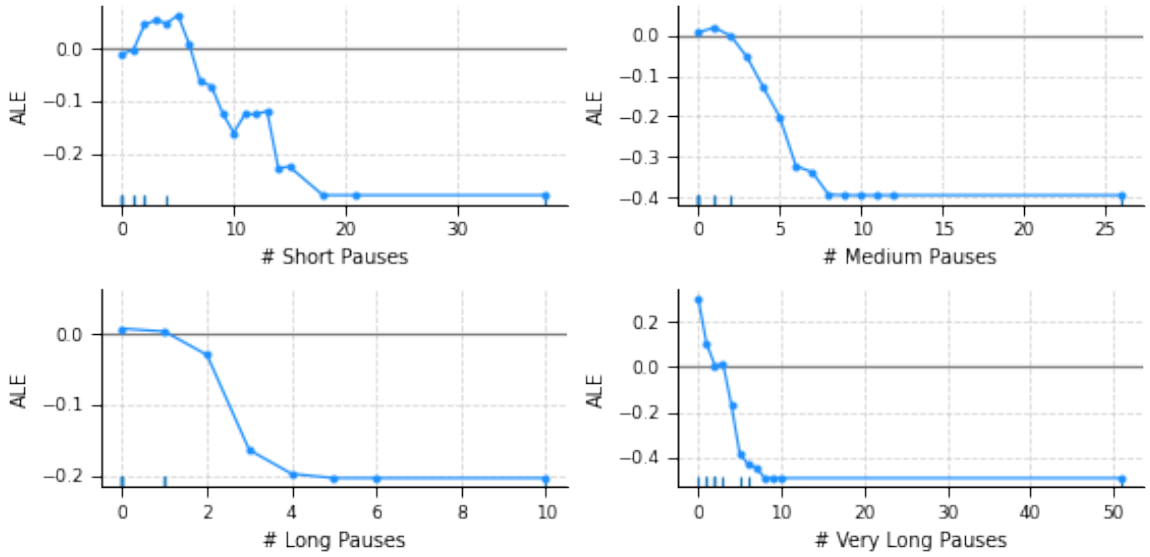


(b) Cars

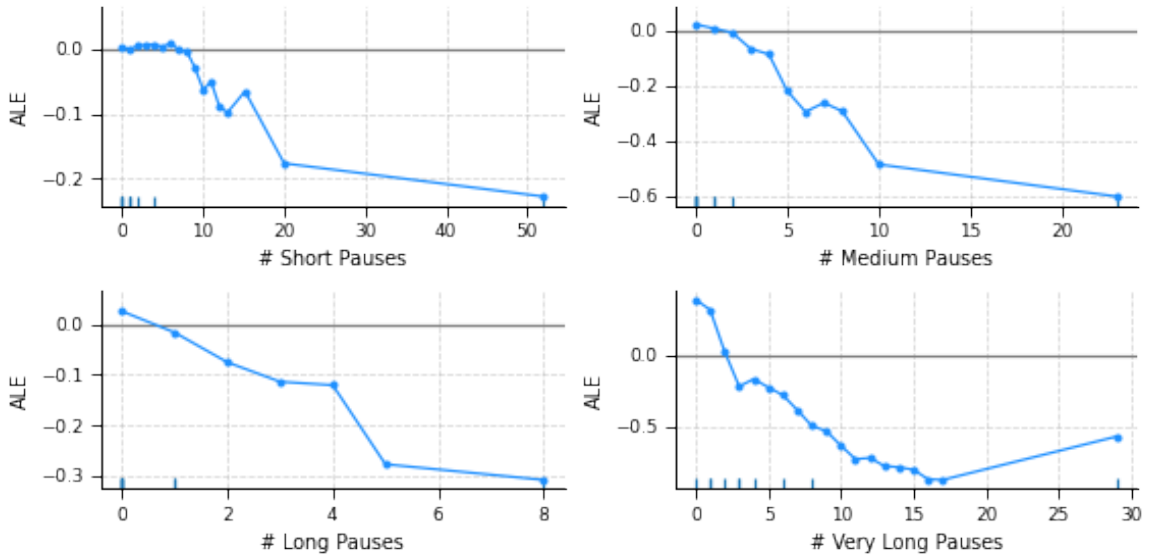
Figure 4.11. The accumulated local effects for the contextual change metric (both proportion and counts). The results demonstrate that a high proportion of contextual changes positively effects the model prediction (therefore weighting the prediction towards higher abandonment distances).

4.2.5 The Relationship between Abandonment & Engagement

Out of the 500 users that took part in the engagement survey, 235 visited both sub-stories, 97 only visited Malawi (and not Cars), and 167 only visited Cars. The distributions of engagement scores between the three groups are shown in Figure 4.15. In terms of differences between engagement scores, there was a significant difference between those that visited both and only the Malawi sub-story ($U(N_{both} = 235, N_{malawi} = 97) = 14384.0, p < .001, f = .63$)) and users that visited both and only the Cars sub-story ($U(N_{both} = 235$, $\mathrm{N}_{cars} = 168) = 24021.5, p < .001, f = .61$)). This difference in engagement suggests that users who watched both sub-stories were more engaged than those that only watched one sub-story.



(a) Malawi

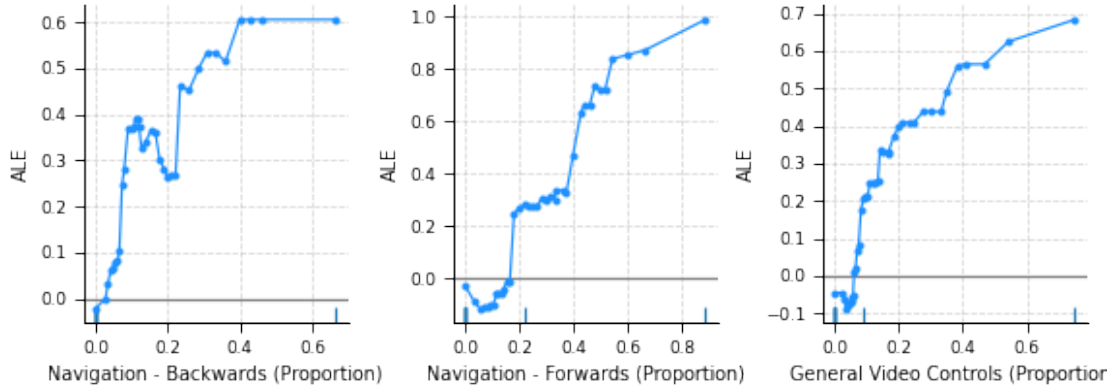


(b) Cars

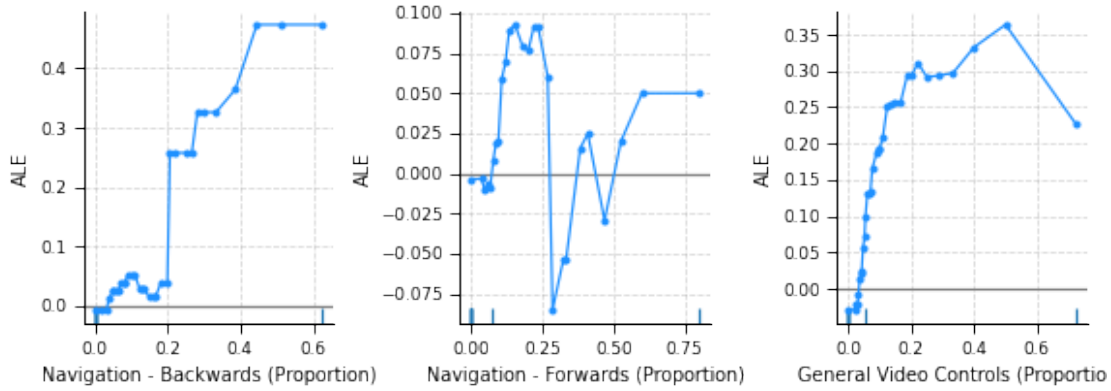
Figure 4.12. The accumulated local effects for the four types of pauses. The results demonstrate that an increase in all pauses have a negative effect on the model prediction (therefore weighting the prediction towards lower abandonment distances).

Both models are accurate in their prediction of distance, Malawi: PD $= 0.31$, MSE $= 0.08$, and MAE $= 0.15$; and Cars: PD $= 0.64$, MSE $= 0.27$, and MAE $= 0.32$. As all users in the engagement dataset finished the Click experience in its entirety – none dropped out of the sub-stories – all of their distance values were zero. The prediction distance for both sub-stories is shown in Figure 4.16 which demonstrate that the majority of predictions are zero or close to zero and were found to be non-parametric.

The predicted distance between the *high* and *low* engagement groups were significantly different in both the Malawi ($U(N_{high} = 149, N_{low} = 183) = 15800, p < 0.05, f = 0.57$) and Cars ($U(N_{high} = 201, N_{low} = 201) = 24887, p < .001, f = 0.61$) sub-stories, with a stronger effect in the latter. Users that register low engagement in both sub-stories are predicted higher distance values (show in Figure 4.17) which suggests there are signals



(a) Malawi



(b) Cars

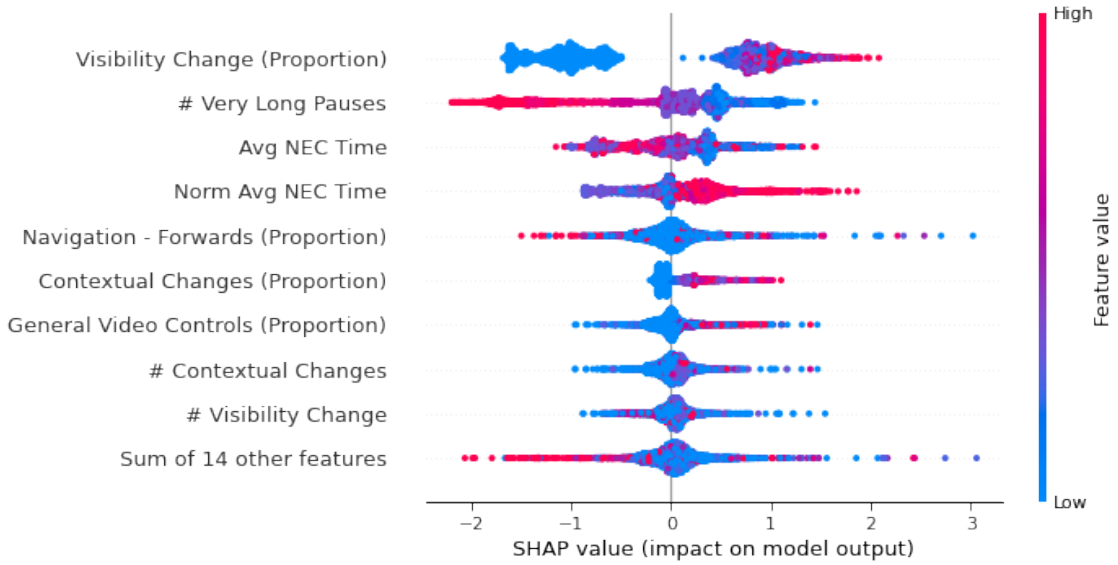
Figure 4.13. The accumulated local effects for the navigation and general video control metrics. The results show that an increase in all of these events has a positive effect on the model prediction, excluding the proportion of forward navigation events (in Cars) which fluctuates but remains at a low level of effect.

in the interaction data indicating that low engagement users are more likely to leave the experience and/or sub-stories.

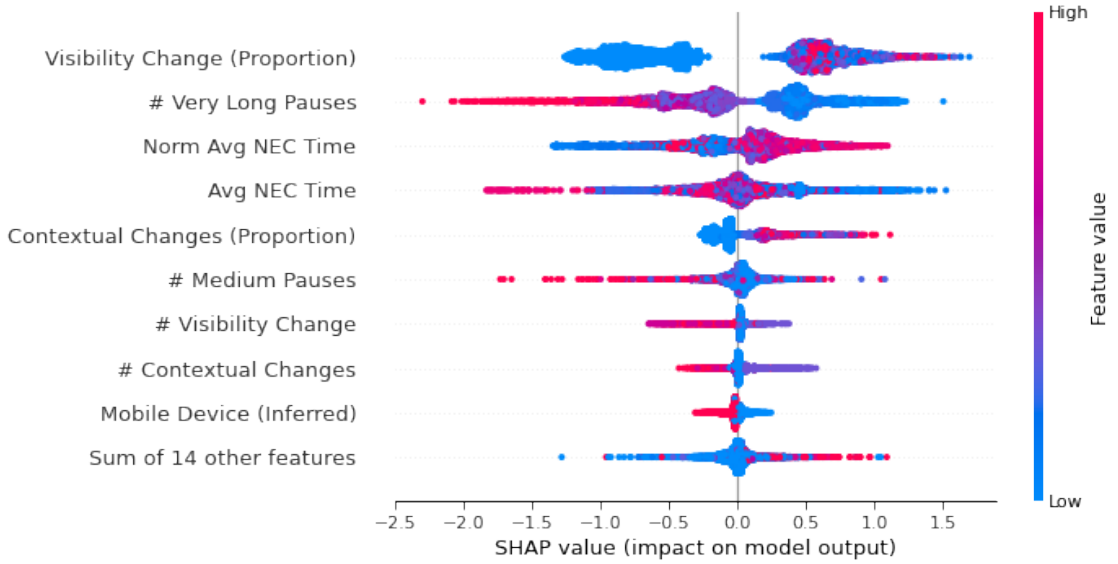
4.3 Summary

This chapter aimed to address the following objective: *to investigate whether there are signals in interaction data that are predictive of abandonment*. Abandonment, much like engagement in the previous chapter, is a facet of success (or non-success) that media creators are interested in. It can give a deeper understanding of the user experience and enable them to further develop best practice in the field of interactive media experiences. This type of media content can have complex narrative structures, where multiple sub-stories make up a show with each following their own structure and presenting their own content. This means that a more nuanced understanding of abandonment (beyond the basic measure of whether an audience member reached the end or not) will be of much more use to media creators. The graph-based structure of the interactive experience was used to create a distance-based abandonment metric and using interaction metrics, abandonment was modelled and predicted in the two main sub-stories from the Click experience.

In investigating **A-RQ1**, it was found that abandonment could be accurately modelled us-



(a) Malawi



(b) Cars

Figure 4.14. Interaction metric importance based on the contribution to the model in predicting abandonment (SHAP values). The position on the y-axis for the metrics is ordered by their mean SHAP value.

ing Poisson-based models. There was little variation across validation steps, with the model trained for Malawi performing slightly better than the Cars model – this may be due to the Cars sub-story being first on the default path through the experience, with user action being necessary to view Malawi first, so this group is self-selecting against those who do not interact much, providing more informative features for the model. The models were able to capture the distribution of the distance metric and suggest that abandonment can be predicted using interaction data collected from interactive media experiences.

To explore and understand what interaction metrics are important in the prediction of abandonment (**A-RQ2**), permutation importance, Accumulated Local Effects, and Shapely Additive Values were extracted, with each providing a different view and interpretation of the model's decision making. As with the engagement model, temporal metrics were found to be important, particularly the average time spent on narrative elements and very long

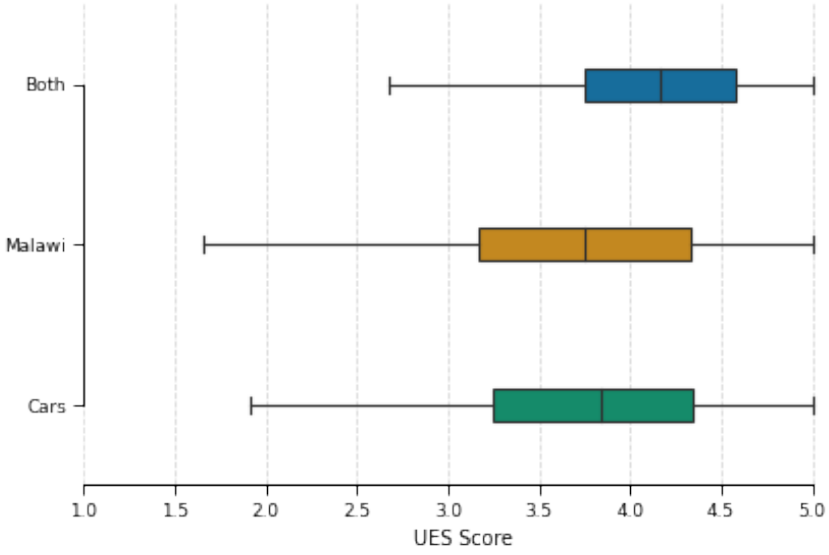


Figure 4.15. The User Engagement Scores of users that visited both sub-stories or only one of them

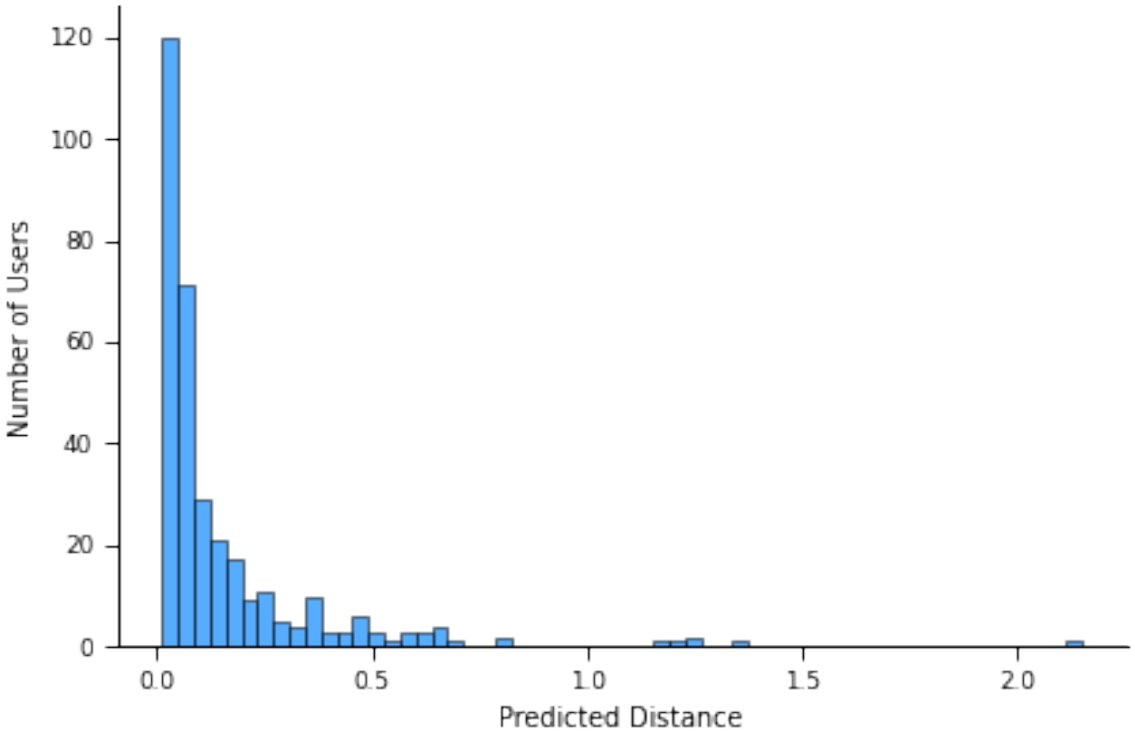
pauses. The accumulated local effects and SHAP both demonstrated the importance of visibility changes, an increase in the proportion pushes the prediction towards higher values (abandonment) and two distinct groups of users were found by SHAP. Low values indicate non-abandonment, while high values indicate abandonment. There is a limitation with the metric, however, as if a user only registers a single event in their entire session, then the metric could account for a 100% proportion.

The proportion of navigation events – backwards and forwards – suggest that an increase in such is likely to lead to abandonment. When investigating engagement, more next events were linked to lower engagement, suggesting a relationship between low engagement and early abandonment. The effect is more prominent in the Malawi sub-story than in the Cars sub-story, suggesting a difference between the two. Similarly, the number of contextual changes negatively effects the Cars model but turns positive as the number increases. Potentially demonstrating a difference between users and rationale for looking at abandonment from a sub-story perspective.

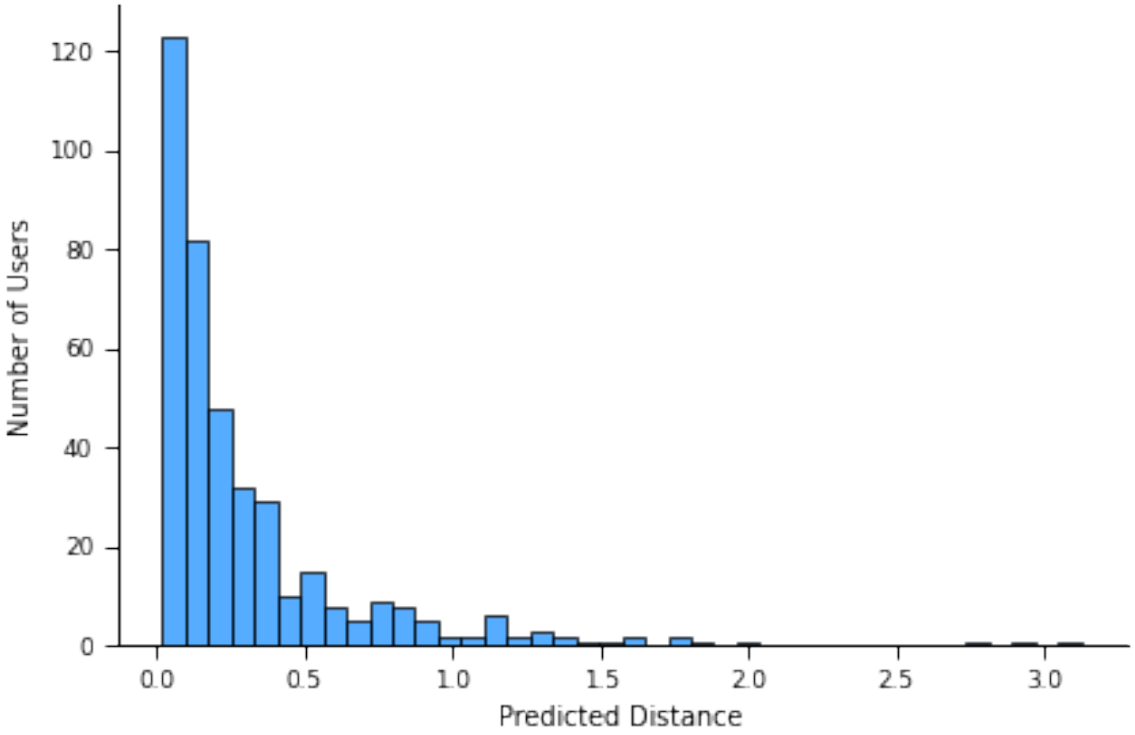
The previous chapter found that engagement was predictable from interaction metrics and **A-RQ3** investigated whether there is a link between abandonment and engagement, with the hypothesis that users who abandon felt lower engagement with the experience. There was a significant difference in engagement scores between users that visited both sub-stories as opposed those who just visited a single sub-story. The models that were previously trained were able to accurately predict abandonment for the users that took part in the engagement survey. When grouping users by *low* and *high* engagement – as per the previous chapter – there was a significant difference between the distance metric between the two groups. Users in the *low* engagement group are predicted higher values, suggesting that they are more likely to abandon in both sub-stories (with a stronger effect in the Cars sub-story). The finding demonstrates that abandonment and engagement are linked and that by inferring abandonment, it provides some level of inference about the engagement of users. However, there are other reasons, external to the experience, that may result in abandonment.

The work in this chapter has built on the prior chapter, expanding the investigation into user behaviour to a larger, real-world dataset collected from interactive media. The findings demonstrate that abandonment can be inferred and predicted from interaction data collected in a production-quality experience and that there are metrics that could enable the real-time monitoring of abandonment. The models suggest that people are more likely to be engaged and less likely to abandon if they are watching in a relatively passive way: perhaps one orientation change and a small amount of interaction with the video controls and then next or back, which can be thought of as settling down to give the content attention. To understand the relationship between user behaviours and abandonment in more depth, beyond modelling, then exploring alternative means could be fruitful, for example, qualitative user-testing, which would allow for the validity to be assessed.

To summarise, the following are the key takeaways from this chapter. Abandonment can be predicted from the interactions of users when posed as a graph-based metric. An increase in the browser changing from hidden to visible is an indicator for abandonment (the higher, the more likely abandonment will occur). The longer a user spends on a narrative element, the less likely they are to abandon. Abandonment is related to low engagement with the content.

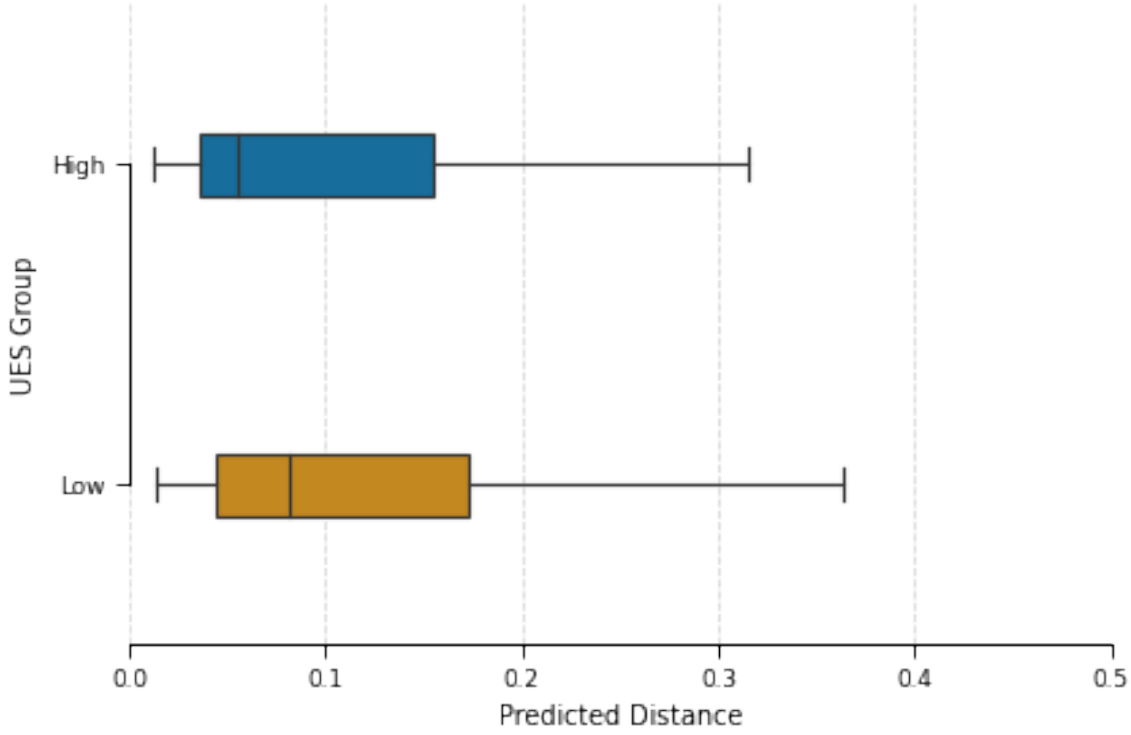


(a) Malawi

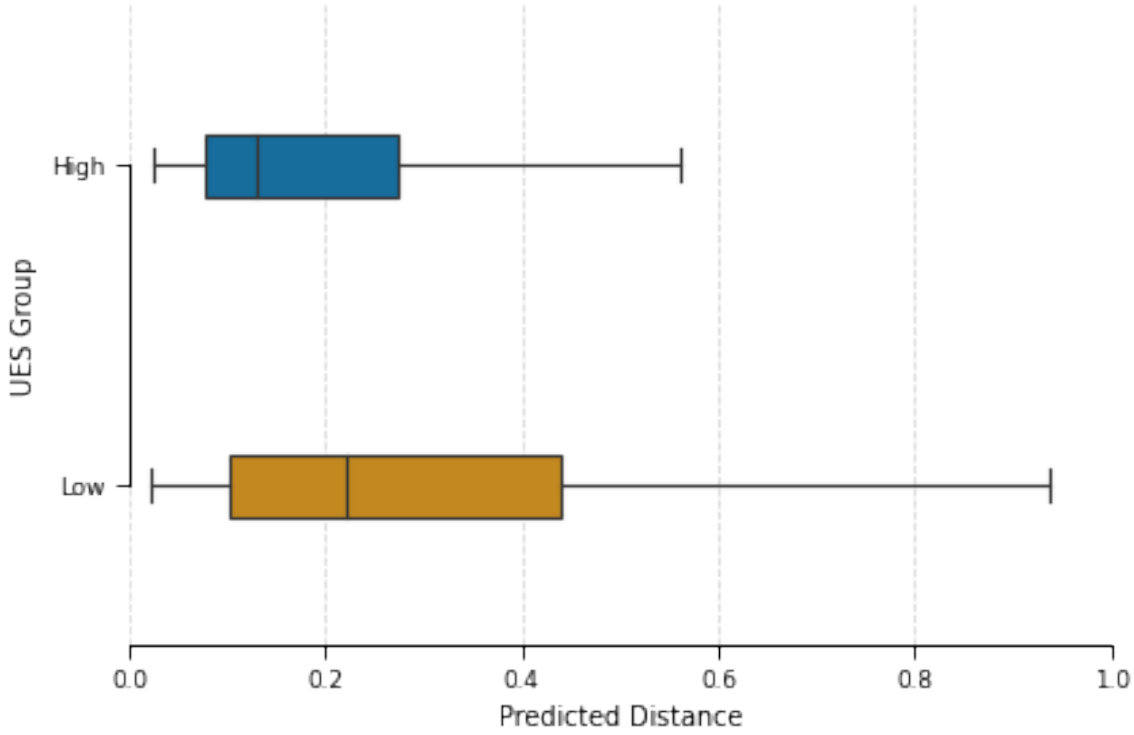


(b) Cars

Figure 4.16. The predicted distance for users that completed the UES survey



(a) Malawi



(b) Cars

Figure 4.17. Predicted Distance for *high* and *low* engagement groups

Chapter 5

Towards a Framework for the Analysis of Interaction Data

As presented in Chapter 2, interaction data has been successfully applied in a range of domains and settings to model and understand user behaviour. Often, interaction data is generated at a high frequency, meaning that analysis is carried out to create descriptive representations – statistics or sequences, for example – of how each user interacted throughout their session. As in this thesis, statistical tests or machine learning techniques are applied to these representations to uncover latent relationships or to predict high-level user characteristics – for example, engagement, abandonment, or satisfaction.

The literature review (2) highlighted that many approaches and techniques are used when working with interaction data. The stakeholders that use the data come from a range of disciplines and therefore, clear and consistent communication is important. However, the processes used were often unclear, decisions made when handling interaction data not explained, and the use of terminology – what constitutes a behaviour, for example – is inconsistent across the literature. Such issues were particularly prevalent when carrying out the empirical work presented in the previous two chapters (Chapters 3 and 4). These issues introduce communication barriers between stakeholders: researchers (those carrying out the analysis) and consumers (those wanting to reproduce, apply, or review the work), which reduces transparency and hinder reproducibility. So, a fresh analysis of the literature was performed to understand these issues in greater detail. From the analysis and drawing on the experiences of carrying out empirical work where these issues were prevalent, a framework is presented to move towards providing methodological support when using interaction data to understand user behaviours, enabling clear communication between stakeholders.

This work, which presents a move towards providing methodological support, is presented at this point in the thesis as a progression from the prior three chapters. More specifically, the investigation into creating a framework is motivated by the issues identified in the original literature review (Chapter 2) and challenges in carrying out the empirical work (Chapters 3 and 4). Additionally, experiences from the empirical work aids in framing the analysis and the layers presented as part of the framework.

5.1 Method

5.1.1 Sampling Articles

Two domains that make significant use of interaction data – Human-Computer Interaction and Data Science – were used to source the papers for analysis. From these two domains, three popular, highly rated, and general-purpose venues that attract a wide range of submissions and are reflective of the field were chosen. To determine the venues, Google Scholar rankings[1] were used – these were recorded at the time of collection and valid as of October 19, 2020.

For the Human-Computer Interaction venues, ACM CHI (Ranked 1st), the International Journal of Human-Computer Studies (IJHCS; Ranked 7th), and ACM Transactions of Computer-Human Interaction (TOCHI; Ranked 10th) from the Human-Computer Interaction sub-category on Google Scholar were chosen. Whilst for the Data Science venues, ACM WebConf (Ranked 2nd; Databases & Information Systems), ACM KDD (Ranked 1st; Data Mining & Analysis), and ACM RecSys (Ranked 6th; Data Mining & Analysis) were chosen. The top three venues in each category were not chosen as the aim was to pick venues that are general-purpose, with higher ranking venues focusing on particular sub-domains within their respective field. For clarity, the following venues were in the top ten for the Human-Computer Interaction category: 1) ACM CHI, 2) ACM CSCW, 3) IEEE Transactions on Affective Computing, 4) ACM/IEEE HRI, 5) ACM UbiComp, 6) IJHCS, 7) IEEE Transactions on Human-Machine Systems, 8) ACM IMWUT, 9) ACM PACMHCI, and 10) ACM UIST; Databases & Information Systems: 1) ACM WebConf, 2) IEEE Transactions on Knowledge and Data Engineering, 3) ACM SIGMOD, 4) VLBD, 5) ACM SIGIR, 6) ACM WSDM, 7) ACM CIKM, 8) Information Processing & Management, 9) IEEE ICDE, and 10) AAAI ICWSM; Data Mining & Analysis: 1) ACM KDD, 2) IEEE Transactions on Knowledge and Data Engineering, 3) AISTATS, 4) ACM WSDM, 5) IEEE Data Mining, 6) ACM RecSys, 7) Knowledge and Information Systems, 8) IEEE Big Data, 9) Journal of Big Data, and 10) ACM TIST.

Papers were collected from their respective online libraries, which in all but IJCHS (Elsevier[2]) were the ACM Digital Library[3]. For each, a search was performed for papers that contained the terms interaction and behaviour in their title or abstract. The search result was then filtered to only include full papers, extended abstracts, and short papers. The results were then sorted by the date of publication and the top 50 were downloaded (or all of the results if less than 50) to collect a broad sample of papers from each venue.

Prior to performing a full analysis on the collected papers, manual filtering was carried out to remove papers that were not relevant; this involved an analysis of the title, abstract, and keywords to determine whether the paper collected interaction data and/or inferred be-

[1] Google Scholar metrics: https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng
[2] https://www.elsevier.com/en-gb
[3] https://dl.acm.org/

haviours from the interaction data. With a reduced sample of papers, each paper was analysed in full to ensure it met the criteria.

5.1.2 Analysis

Using the final set of papers, a full analysis was performed. To develop a clearer picture of how interaction data is used across the two fields, what techniques and approaches are used, and to identify themes in the literature, each paper was evaluated individually with the aim of answering the following questions:

- What type of interaction data is collected or used?
- What type of processing is performed? What form is their data in: statistics, sequences, or something else?
- How are human behaviours, motives, actions, or decisions represented? Are they inferred through analysis? What techniques are used to determine the differences or similarities?
- What outcome are the authors focusing on; understanding, modelling, or exploring?
- What evaluation or validation is performed?

5.2 Results

A total of 282 articles were collected from the sources listed previously. For each of the HCI venues, as well as KDD and WebConf, 50 articles were collected, while for RecSys, 32 articles were sourced. From carrying out the initial screening, where titles, abstract, and keywords were evaluated, 204 papers were excluded which left 78 for full review. Of the remaining 78, an additional 32 papers that did not use interactions and/or were not inferring/extracting behaviours were removed, leaving a total of 46 articles. These articles are the focus of the remaining analysis. From this, we found that the collected samples of articles cover a range of different themes, from understanding users to proposing new types of modelling approaches.

5.2.1 Notes from the Analysis

The following are examples taken from the analysis. For the complete set of notes, we refer the reader to the reference in the introduction (see Section 1.4).

(Tanjim et al., 2020) A new modelling approach is proposed to capture user intent from the sequential interactions of users on a shopping website. The idea is that by inferring intent then better next-item and next-interaction recommendations can be served.

Data Data is sourced from two open-source repositories and are contextualised sequences of actions, e.g., click, add-to-favourite, add-to-cart, and purchase events. The actions in the second dataset are not described. There is a third dataset used for comparison purposes and is a smaller version of the first dataset.

Metrics Their data is used in its collected form. The authors propose a novel neural network-based model architecture.

User Behaviours Their work is motivated by being able to differentiate between a user in 'discovery' and 'purchase' mood and how they capture this user intent in their model architecture. However, the actual intent of users is not captured or validated – they do not know that a user is in a 'discovery' mood. Latent intent is learnt from the interaction as part of the modelling process, the authors do not capture an explicit notion of intent (e.g., by asking the users).

Outcome of Interest The authors are focused on advancing the state-of-the-art and frame their research questions in this way.

Validation Validation is not performed to assess whether their model captures intent. The authors state that their model captures intent based on its performance.

(L. Guo et al., 2019) The paper proposes a new modelling approach that captures both long-term user intentions and short-term preferences, i.e., the model remembers items that the user has clicked on in the past but also uses new information about their interactions. The model is designed to predict the next items that the users are most likely to click on.

Data Binary indications whether the user clicked on an item within a session. The LastFM dataset is used, along with a dataset from a check-in website called Gowalla.

Metrics Their data is organised into sequences. Their model proposes a way to use both historical click events and new click events to optimise the set of items recommended to users. The model performs better than the baselines.

User Behaviours The authors refer to their data as 'behaviours' consistently throughout. They also state that due to their method outperforming another it indicates that user intent is being captured.

Outcome of Interest They are interested in next-item recommendations, specifically advancing the state-of-the-art.

Validation They do not validate the intentions of the user – in this setting, intention is a click on an item. The authors do not know if their model captures the actual intent of the users (they have no ground truth or proxy), but that their model performs better for the specific task it was designed for.

(Chen et al., 2018) The paper presents a new dataset and modelling approach for predicting the attractiveness of videos. The dataset consists of video segments and descriptive statistics. The authors term their statistics as user behaviours and include statistics such as view, fast-forward, and fast-rewind counts. The view counts are the target for video attractiveness, which are predicted using a neural network model. Throughout the paper, the language is inconsistent with the statistics being referred to as engagement indicators or user engagement behaviours.

Data Their dataset contains video segments and statistics derived from interactions that users perform on the videos.

Metrics The statistics included in the data are: exit, start of fast-forward, end of fast-forward, start of fast-rewind, end of fast-rewind, fast-forward skips, fast-rewind skips, bullet screens, and likes. All of which are counts. View counts are also included as used as their modelling target. Various deep learning-based models are trained, with features removed in some cases to compare the performance with and without.

User Behaviours The authors do not attempt to understand the behaviours of users, even though they present the general trend of the statistics. They do however consistently refer to their statistics as behaviours. The view count is referred to as ground truth attractiveness which is not validated or captured directly from the users.

Outcome of Interest They are interested in predicting the attractiveness of videos, which is represented by the view count for each video – the higher, the more users found that video or segment attractive.

Validation Validation is performed on the model to test its ability to predict view count. Validating their measure – the view count as a proxy for attractiveness – is not performed.

(Z. Li et al., 2018) The paper presents a new next-item recommendation model which incorporates two separate models for historical and present user 'behaviour'.

Data Actions that are performed on a shopping website, i.e., click, add-to-cart, and purchase, are collected.

Metrics The actions are organised into sequences. The authors exclude 'click' actions from the input into the user preference model and only include those more closely associated with preference, such as collect, add-to-cart, and purchase actions.

User Behaviours The authors frame their research as capturing the historical and present behaviours of users to perform next-item recommendation. They report

that because of an increase in performance of the model, the behaviours are being successfully captured by the model. They use the terms 'behaviour', 'behaviour sequences', 'sequential behaviours', or 'preference behaviours' to refer to the sequences of actions.

Outcome of Interest They are interested in the state-of-the-art next-item recommendation.

Validation No validation is performed beyond comparing their model with other state-of-the-art models. The authors do not evaluate if their model actually captures user behaviour or motivation, nor is it validated with the users directly.

5.2.2 Interaction Data

Across the papers, four types of interaction data were identified: application-specific interaction, lower-level interaction events, logs of items on websites click on by users, and interaction data that is already in statistical form. These are described in more detail below:

Application-specific interactions: These are a commonly collected form of interaction data (Ekstrand et al., 2015; L. Li, Deng, et al., 2017; Z. Li et al., 2018; Lin et al., 2020; Z. Liu et al., 2020; Macha et al., 2020; Natarajan et al., 2013; Tanjim et al., 2020; Wan and McAuley, 2018; W. Wang et al., 2020; Xu et al., 2020; H. Zhang et al., 2020; Y. Zhang et al., 2020; X. Zhao et al., 2020). They are descriptive of actions that a user can perform on the system in question, for example, add-to-cart-type events on an e-commerce website (Tanjim et al., 2020), visits and ratings (H. Zhang et al., 2020), and picture editing events (Z. Liu et al., 2020).

Lower-level interaction events: A fine-grained form of interaction data which includes cursor movements and keystrokes and tends to be more commonly collected by those focusing on understanding users (Apaolaza and Vigo, 2019; Vigo and Harper, 2017; Vizer and Sears, 2017; Wampfler et al., 2020).

Clicks on items: The collection of clicks on items on websites is popular interaction data type (Belletti et al., 2019; Cen et al., 2020; Fan et al., 2019; Ge et al., 2020; L. Guo et al., 2019; He et al., 2017; Jia et al., 2020; Ma et al., 2020; Niu et al., 2020; Pasricha and McAuley, 2018; D. Wang et al., 2020; Yao et al., 2020; Yuan et al., 2020; J. Zhao et al., 2019; Zhong et al., 2012), which consist of examples such as: video IDs (Belletti et al., 2019), new articles (Ge et al., 2020), and movies (Zhong et al., 2012).

Pre-processed form: The final data type collected is interaction data that is already in pre-processed form (Chen et al., 2018; Curmi et al., 2017; Grinberg, 2018; Hariri et al., 2014; Kulp et al., 2020; Lamba and Shah, 2019; Trattner and Elsweiler, 2017; Zagermann et al., 2020). For example, the number of times particular actions on the website that are performed by the user (Chen et al., 2018; Grinberg, 2018) or the cumulative time spent on the website (Curmi et al., 2017).

5.2.3 Data Representation

Broadly, articles tend to work with sequences of interactions or statistical representations extracted from interaction data, with some focusing on more specific approaches, for example, embeddings (translating high-dimensional data into low-dimensional space) (He et al., 2017; X. Zhao et al., 2020) and graph-based representations (Ge et al., 2020; Jia et al., 2020; Niu et al., 2020; D. Wang et al., 2020; W. Wang et al., 2020; J. Zhao et al., 2019; Zhong et al., 2012).

Sequences are a popular representation method for interaction data in the sampled literature, particularly with papers that focus on modelling (Apaolaza and Vigo, 2019; Belletti et al., 2019; Cen et al., 2020; L. Guo et al., 2019; Z. Li et al., 2018; Lin et al., 2020; Ma et al., 2020; Natarajan et al., 2013; Wan and McAuley, 2018; Yao et al., 2020; Yuan et al., 2020; H. Zhang et al., 2020; X. Zhao et al., 2020). These are representations that temporally order the interactions of users, for example, (Cen et al., 2020; Tanjim et al., 2020) creates sequences of application-specific events, (Belletti et al., 2019) uses sequences of video identification numbers that the users have interacted with, and (Wampfler et al., 2020) uses 2D heatmaps of interactions on a webpage. There are some examples where sequences are used by papers focusing on understanding users, but these tend to have a modelling component (M. Liu et al., 2019; Macha et al., 2020; Tanjim et al., 2020; Teo et al., 2016; Wampfler et al., 2020).

Statistical representations – summaries – of interactions performed over the course of a user session are the main alternative method of data representation. The approach of summarising interaction sessions is common across the sampled literature but has more prominence in papers attempting to understand users than sequences (Curmi et al., 2017; Grinberg, 2018; Müller et al., 2017; Trattner and Elsweiler, 2017; Vigo and Harper, 2017; Vizer and Sears, 2017; Xu et al., 2020; Youngmann and Yom-Tov, 2018; Zagermann et al., 2020). For example, the number of times a user scrolled (Vigo and Harper, 2017; Youngmann and Yom-Tov, 2018), the dwell time (Grinberg, 2018; Vizer and Sears, 2017), and counts or descriptive statistics of interaction events (Trattner and Elsweiler, 2017; Xu et al., 2020; Zagermann et al., 2020).

5.2.4 Behaviour Representation

Often, interaction data - and the representations extracted - are used to develop an understanding or infer user behaviours from. A tendency that is exclusive to papers focusing on modelling is making the implicit presumption that their data is representative of user behaviours, with authors referring to their data as behaviours (Belletti et al., 2019; Chen et al., 2018; Jia et al., 2020; Lin et al., 2020; Ma et al., 2020; Pasricha and McAuley, 2018; D. Wang et al., 2020; Yuan et al., 2020; H. Zhang et al., 2020; Y. Zhang et al., 2020; J. Zhao et al., 2019) – none of the papers that focus on understanding users do this. This tendency and inconsistency in language use is demonstrated in Section 5.2.1, with both Cen et al., 2020,

and L. Guo et al., 2019, referring to their sequences of actions as 'behaviour sequences', 'click sequences', and 'click behaviours', and Chen et al., 2018, referring to the statistics extracted from interaction data as 'behaviours'. In some cases, sub-sequences and combining descriptive statistics extracted from interaction data are used to represent a high-level behaviour, while in other instances, the term 'behaviour' is used to refer to the data itself (as per the examples previously). Similarly, the implicit representation of behaviours in the modelling process and the resulting claims about the models ability to capture behaviours is again mostly exclusive to papers focusing on modelling. Commonly, authors state that their proposed modelling approach captures a behaviour or high-level user trait based on the performance of their model over other state-of-the-art approaches. Articles by L. Guo et al., 2019; Z. Li et al., 2018; Tanjim et al., 2020, in Section 5.2.1 all do this, with others stating that there is a latent representation of user behaviour or traits such as intent (Fan et al., 2019; Niu et al., 2020; Wan and McAuley, 2018; X. Zhao et al., 2020), preference (Ge et al., 2020; He et al., 2017), or interest (Yao et al., 2020; Zhong et al., 2012). In contrast, the dominant approach used by those focusing on understanding users is to uncover behaviours through combining statistics extracted from interaction data (Apaolaza and Vigo, 2019; Curmi et al., 2017; Grinberg, 2018; M. Liu et al., 2019; Macha et al., 2020; Mehrotra et al., 2020; Müller et al., 2017; Trattner and Elsweiler, 2017; Vigo and Harper, 2017; Xu et al., 2020; Youngmann and Yom-Tov, 2018). For example, the application of correlation analysis or testing for statistical differences between metrics derived from interactions or groups of users (Curmi et al., 2017; M. Liu et al., 2019; Müller et al., 2017; Trattner and Elsweiler, 2017; Xu et al., 2020). An alternative approach to understanding behaviours shown in the literature is through video recording and coding behaviours, providing an accurate measure and understanding of physical behaviours performed by people (Kulp et al., 2020; Zagermann et al., 2020).

5.2.5 Outcomes

Most articles fit into one of two outcome focuses: improving the state-of-the-art or understanding the users.

Improving the state-of-the-art: As shown in Section 5.2.1, all four examples focus on improving the state-of-the-art in next-item recommendation (Cen et al., 2020; L. Guo et al., 2019; Tanjim et al., 2020) and video attractiveness (Chen et al., 2018). These present the most recent stage in the development of technology; developing models and expanding on the latest techniques. This strong focus on aiming to improve the state-of-the-art are found in papers proposing modelling approaches (Belletti et al., 2019; Cen et al., 2020; Fan et al., 2019; Ge et al., 2020; L. Guo et al., 2019; Hariri et al., 2014; He et al., 2017; Jia et al., 2020; Lamba and Shah, 2019; L. Li, Deng, et al., 2017; Z. Li et al., 2018; Lin et al., 2020; Ma et al., 2020; Natarajan et al., 2013; Niu et al., 2020; Pasricha and McAuley, 2018; Tanjim et al., 2020; Wan and McAuley, 2018; D. Wang et al., 2020; W. Wang et al., 2020; Yuan et al., 2020; H. Zhang et al., 2020;

Y. Zhang et al., 2020; J. Zhao et al., 2019; X. Zhao et al., 2020; Zhong et al., 2012).

Understanding the users: Articles concentrating on understanding users look for differences, similarities, or specific user traits such as satisfaction (Apaolaza and Vigo, 2019; Curmi et al., 2017; Grinberg, 2018; Kulp et al., 2020; M. Liu et al., 2019; Macha et al., 2020; Mehrotra et al., 2020; Trattner and Elsweiler, 2017; Vigo and Harper, 2017; Vizer and Sears, 2017; Wampfler et al., 2020; Xu et al., 2020; Youngmann and Yom-Tov, 2018; Zagermann et al., 2020). Some articles do, however, explore new approaches or tools, for example, segmenting image editing logs (Z. Liu et al., 2020) and a tool to assist users in applying pattern mining (Apaolaza and Vigo, 2019).

5.2.6 Evaluation or Validation

The articles focusing on the state-of-the-art evaluate the performance of their proposed model. There is a tendency by articles who make claims relating to user traits to not perform evaluation or validation on those claims (Cen et al., 2020; Chen et al., 2018; Fan et al., 2019; L. Guo et al., 2019; Hariri et al., 2014; He et al., 2017; Jia et al., 2020; Lamba and Shah, 2019; L. Li, Deng, et al., 2017; Z. Li et al., 2018; Lin et al., 2020; Ma et al., 2020; Natarajan et al., 2013; Niu et al., 2020; Pasricha and McAuley, 2018; Wan and McAuley, 2018; D. Wang et al., 2020; W. Wang et al., 2020; Y. Zhang et al., 2020; J. Zhao et al., 2019; X. Zhao et al., 2020; Zhong et al., 2012), instead focusing on the model performance. The example articles in Section 5.2.1 demonstrate this, with Tanjim et al., 2020, Z. Li et al., 2018, and L. Guo et al., 2019, relating an increase in performance to behaviours and traits being successfully captured by their proposed models. As additional examples, claims about user preference are made by Teo et al., 2016, but are not tested prior to in-the-wild deployment of their model and both Tanjim et al., 2020, and Xu et al., 2020, do not test whether their model captures the intent of users. In some cases, however, articles capture a form of ground truth about a user trait and use it as a prediction target or grouping method, which results in their conclusions being framed around a more empirical notion of that user trait. This is something that is more prominent in papers that focus on understanding users (M. Liu et al., 2019; Vizer and Sears, 2017; Wampfler et al., 2020; Zagermann et al., 2020). Similarly, some articles evaluate their proposed model, system, or findings through user studies or with other datasets (Apaolaza and Vigo, 2019; Grinberg, 2018; Macha et al., 2020) – testing their approach on unseen data or in new settings, providing evidence for the veracity of their findings.

5.2.7 Summary

The analysis has shown there are a mixture of approaches used by researchers when working with interaction data, for example, data representation after processing and the various analytical approaches applied. Most prominent is inconsistencies in what constitutes a behaviour across the literature. A large proportion of researchers focusing on state-of-the-art

user modelling implicitly presume that user behaviours are represented by their data or latently in the inner workings of complex models, neither of which are evaluated or tested – demonstrated by the examples in Section 5.2.1. For example, that a sequence of video IDs a user has viewed, or a sequence of clicks, are representative of user behaviours. These are more akin to sequences of actions and not behaviours. To add to the confusion, researchers also take the approach of uncovering behaviours through combinations of descriptive measures. The results further expose a tendency by researchers in this space to make claims about their findings that require additional validation, for example, that some high-level user trait is latently captured by their model as it outperforms other state-of-the-art approaches. All these issues create a communication barrier and hinder transparency and reproducibility. To address this, in the next section, a framework is proposed to provide structure and facilitate clear communication of research using interaction data.

5.3 Proposed Framework

As demonstrated in the analysis, there is a communication issue with research in the interaction data analysis space. The clear communication of research is evidently important, it means that those wishing to consume research fully understand the process undertaken by the researchers, it enables the reproduction of results, and encourages open research. For those wishing to reproduce, apply, or review the work – the consumers – inconsistency of language and implicit presumptions made about data or complex models obfuscates the information, limiting their ability to do so. This obfuscation is by no means intentional by researchers but is rather introduced due to the lack of a consistent and structured method to describe and discuss their work. As such, a framework which could be applied to all the literature analysed and is intended to improve communication between stakeholders is proposed.

5.3.1 Structure

To define the structure of the framework, a set of definitions are provided in this section - shown graphically in Figure 5.1. These definitions are based on the literature and are designed for use within this framework, with their evolution being open to the community.

Interaction Data: The data collected from systems in its raw format: a collection of singular actions performed by users (e.g., a mouse click or movement) and/or generated by the system itself (e.g., an event recording the change of song in a music listening context).

Interaction Metrics: An abstraction from the interaction data collected and the analytics that are applied to that abstraction. For example, the frequency of changing songs in a music listening context.

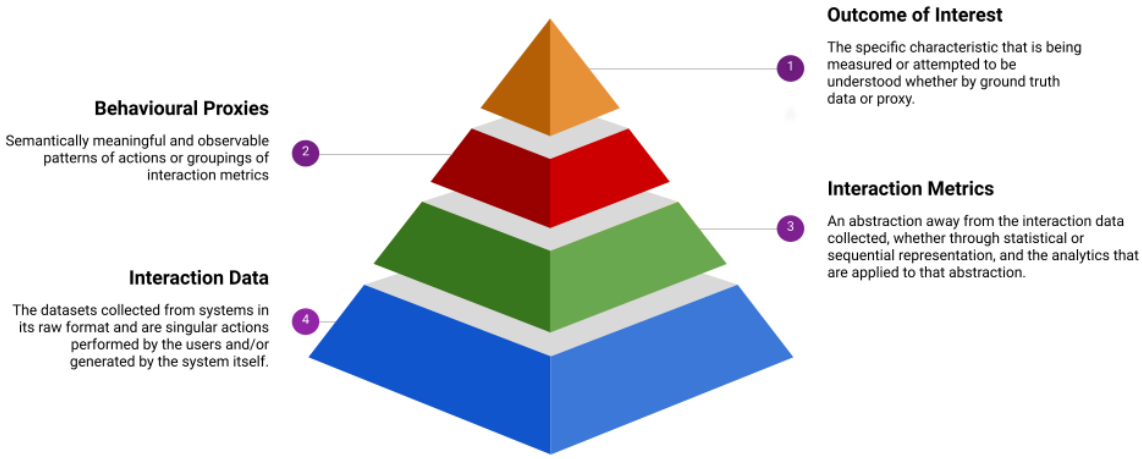


Figure 5.1. The Framework

Behavioural Proxies: The dictionary definition of a behaviour is: *"an instance or way of behaving. Now usu. of animals or people as objects if study; an observable pattern of actions, a response to a stimulus" Oxford Dictionary of English (3 ed.)* 2010; in this context, we define it as a semantically meaningful and observable pattern of actions or groupings of interaction metrics. For example, a high song change frequency in a short time period indicates the user skipping through songs. It is worth noting that due to the dictionary definition of 'behaviour', an event sequence could be literally considered a behaviour. But, as demonstrated in the analysis of the literature, researchers are not being literal, and are using the term as a proxy for something higher level.

Outcome of Interest: The specific characteristic that is being measure, investigated, or attempted to be understood. For example, user satisfaction with recommended song playlists.

The framework is designed such that when authors use it, they should *build-up*; so when one layer is talked about, the authors should then always provide a complete picture of the lower layers which provide context and clarity. For example, when discussing the behavioural proxies layer, both the interaction data and interaction metrics layer should also be clearly discussed and presented.

5.3.2 Use-Cases

To demonstrate how the framework can be applied to existing literature, how each article's components – their interaction data and analysis – fit into each layer, and how the work is structured relative to the framework, use-cases are drawn from the prior analysis. To further demonstrate the framework in use, it is applied to the empirical work presented in this thesis.

Z. Liu et al., 2020, explored using interaction data to automatically segment image editing tasks into distinct sub-tasks with the goal of supporting editors in their work.

Interaction Data: collected during a think-aloud study with experts who were asked to carry out three tasks: poster creation, portrait retouching, and special effect creation. Interactions were collected, containing four attributes: layer ID, command (e.g. move), image, and timestamp.

Interaction Metrics: derived from the data and categorised into interaction and contextual, which included: command similarity, layer similarity, duration, working region overlap, and image diff. A two-stage modelling approach was used, first predicting boundaries for low-level task chunks and then using the output to predict high-level sub-tasks which make up the task as a whole (for example, position images).

Behavioural Proxies: inferred using the predicted boundaries, sub-tasks groupings, and expert generated labels for the tasks. The authors explored variations in individual editors, finding commonalities and differences: they often make mistakes, they interleave different sub-tasks together, perform experimentation through trial and error, and task boundaries are intrinsically fuzzy.

Outcome of Interest: the authors frame the work around automatically segment image editing tasks, with the aim of supporting editors with additional tools.

An investigation of user behaviours while searching for medical symptoms and how interactions change due to anxiety was performed by Youngmann and Yom-Tov, 2018.

Interaction Data: collected from a production search engine only when a user mentioned a medical symptoms in their query. The data was mouse cursor movements and consisted of time-stamped horizontal and vertical coordinates of the cursor location.

Interaction Metrics: summarised the user's interaction on the search engine result page and consisted of descriptive statistics for x and y points, number of local y minimums, duration of the session, total mouse distance, rank of deepest clicked result, fraction of displayed results clicked, number of clicks below 1^{st}, 3^{rd}, or 5^{th} index, and percentage of the screen seen. A learn-to-rank model was trained and the severity of symptoms described by the user in their query were predicted.

Behavioural Proxies: a comparison of user query ranks was performed using two interaction metrics: the percentage of the screen seen and rank of the deepest click result. They found that users with a higher inferred anxiety focus on a smaller proportion of the result page and click higher ranked results.

Outcome of Interest: how anxiety affects search behaviour was the focus, with a measure being provided by medical professionals asked to rank symptoms described by users in their queries. To assure a level of confidence in their measure, lay-people were also asked to rank queries based on if they were experiencing the same symptoms, finding that both were highly correlated.

Both use-cases above broadly follow the structure of the framework. Using the information presented in each article, the framework can be fitted with relative ease. For the second use-case, however, the use of the framework would introduce consistency in the language used. The authors use the term 'behaviours', but using 'behavioural proxies' would be more accurate. The differences found between users point towards high-level and untested behaviours, which the authors could hypothesise about while referring to them as behavioural proxies.

Moving away from the related literature examples and taking an overall view of the empirical work presented in this thesis, the work can be framed within the scope of the framework.

Interaction Data: collected from interactive media experiences produced by the BBC. Application-specific events were captured and consisted of a user ID, timestamp, the type of action performed, the name of the action performed, and additional metadata about the event (for example, hidden or visible states for a browser visibility change event).

Interaction Metrics: descriptive statistics were derived from the data and consisted of session length, time to completion, hidden time, pauses between interaction events (short, medium, long, and very long), event counts, proportions of events, and the total number of events. Statistical tests (correlation analysis and difference testing) were performed and models were trained to predict engagement (as per Chapter 3) and abandonment (as per Chapter 4), with metric importance being extracted from the models and model-agnostic methods applied to the models.

Behavioural Proxies: there is an interest in inferring behaviours as these might provide an understanding of the link between the content or context and outcome, to open up the opportunity of content adaption or user interface modification. Some proxies are speculated about through the results of the analysis but are not tested, for example, skipping-type behaviours associated with low engagement.

Outcome of Interest: a measure of abandonment was derived from the story structure and engagement was captured through an established survey in two user studies.

The examples above describe their method and results in a manner where fitting the framework and using the provided definitions only requires minor readjustments to terminology used. But, as found in the analysis of the literature, there are several issues with the way in which research is communicated, one of which is the inconsistent use of language. The framework is designed to introduce consistency, to facilitate clearer communication between the two stakeholders; if authors were to describe and present results in line with this framework, the potential confusion caused by the language discrepancies currently found in the literature could be significantly reduced. Section 5.2.1 presents some examples of such discrepancies, and illustrates how the framework could be used to address the problem. For example, the inconsistency when describing data representations extracted from interaction data, with Chen et al., 2018, referring to statistics extracted from their interaction data

as ‘behaviour’; and Z. Li et al., 2018, referring to their sequences as either ‘behaviours’, ‘behaviour sequences’, ‘sequential behaviours’, or ‘preference behaviours’. These terms broadly describe the same data type and referring to them as interaction metrics introduces a consistency in the language and brings all data representations under the same umbrella term. By consistently describing the forms of data (extracted from interaction data) as interaction metrics, it prevents the tendency to refer to them as behaviours (incorrectly in these instances), which are associated with a higher level of the framework.

There is also a trend to use an increase in the performance of a model over the current state-of-the-art as evidence that a behavioural proxy or high-level trait is being captured. For example, in Section 5.2.1 articles by Tanjim et al., 2020, and L. Guo et al., 2019, say this of intent and Z. Li et al., 2018, more generally refers to user behaviours being captured. While it is possible that the proposed models capture this, the claims positions their results into the top layers of the framework, requiring further validation with respect to a form of ground truth, which the authors may not intend. Using the framework, the authors can be more consistent in their language and express a degree of confidence in their results, for example, while the results may point towards a latent behavioural proxy being captured or represented, further validation would be needed to boost confidence in the validity of the results.

In some cases, articles use consistent language and frame their results within layers of the framework, but earlier parts are missing, hindering reproducibility. For example, Mehrotra et al., 2020, investigates user receptivity to divergent music recommendations. The authors state that their measure of receptivity is a possible proxy for behaviour and that additional work is needed, which is good. However, the interaction metrics used in the investigation are described, but not in full with only those associated with user characteristics included, and the interaction data to create those metrics is not described or presented in the article. With these components missing, the results cannot be reproduced, and while the reason for their exclusion may be commercial or limits imposed by the publication template, the authors should publish the data separately and link to it or explicitly state the reasons for not including it.

5.3.3 Checklist for Application

In the previous section, the framework was applied to several use-cases where its ability to address the issues found in the literature around language use and reproducibility was demonstrated. This section provides a checklist for authors to follow when reporting research that uses interaction data to understand or model users. Its application should ensure that the work is properly described, can be understood by researchers across disciplines, and is reproducible.

It is recommended that the framework is followed consistently, with the correct terms used throughout the work to describe the data, the interaction metrics derived from that data, behavioural proxies used to refer to the output of analysis and any inference made about user

behaviours, and defining the outcome of interest. When discussing and presenting results it is important for authors to carefully consider in which layer their contribution lies; the framework should help these contributions be framed in a way that ensures they are clearly stated and robustly supported by the data. Similarly, application of the framework could also make it clearer that when discussing layers above the main contribution, these statements are essentially hypotheses forming the basis for future work. To aid in the application of the framework, the layers are first restated:

Interaction Data: The data collected from systems in its raw format: a collection of singular actions performed by users (e.g., a mouse click or movement) and/or generated by the system itself (e.g., an event recording the change of song in a music listening context).

Interaction Metrics: An abstraction from the interaction data collected and the analytics that are applied to that abstraction. For example, the frequency of changing songs in a music listening context.

Behavioural Proxies: The dictionary definition of a behaviour is: *"an instance or way of behaving. Now usu. of animals or people as objects if study; an observable pattern of actions, a response to a stimulus" Oxford Dictionary of English (3 ed.)* 2010; in this context, we define it as a semantically meaningful and observable pattern of actions or groupings of interaction metrics. For example, a high song change frequency in a short time period indicates the user skipping through songs. It is worth noting that due to the dictionary definition of 'behaviour', an event sequence could be literally considered a behaviour. But, as demonstrated in the analysis of the literature, researchers are not being literal, and are using the term as a proxy for something higher level.

Outcome of Interest: The specific characteristic that is being measure, investigated, or attempted to be understood. For example, user satisfaction with recommended song playlists.

The following checklist is provided as a list of prospective elements for authors to consider and use when applying the framework to their work:

- ☐ Interaction data is presented in full and in a clear, interpretable, and transparent manner, hosted on an open data repository[4], or provided in supplementary material;
  - ☐ How the data was collected is described;
  - ☐ The attributes of each event are reported;
  - ☐ If data cannot be included, the reason why is explicitly stated;
- ☐ The omission as well as the inclusion of events are reported (whether related to the goal of the research, ethical or privacy concerns, or limitations of the system);

[4] A list of example open data repositories can be found here: https://www.nature.com/sdata/policies/repositories

- ☐ The processing applied to the interaction data is clearly described, including the step(s) taken to create the abstraction and what its final form is;
- ☐ Abstractions from the interaction data (e.g., statistics or sequences) are referred to as interaction metrics throughout;
- ☐ If multiple interaction metrics are derived from the data, then each are described;
- ☐ The analysis, whether statistical test, modelling, or another approach, that is performed on the interaction metrics is described;
- ☐ Behavioural proxies are referred to as behavioural proxies throughout and are not referred to as behaviours;
- ☐ Validation against a form of ground truth is performed and reported when establishing how representative of a behaviour as result is;
- ☐ Speculation about results is made clear, for example, results fitting different levels of the framework;

5.4 Summary

In this chapter issues previously identified in the literature were explored in greater detail. Through a fresh analysis of the literature, it was found that there were issues in the way that research is communicated, with the term 'behaviours' being used inconsistently, and that there is an unintentional tendency for researchers to make unsupported claims about their results. To address these issues, a framework was proposed to allow for clearer communication between two stakeholders, the consumers of research and the researchers, and to provide methodological support for those carrying out user behaviour research using interaction data. The framework is composed of four layers, providing a structure and consistent language foundation for research in the space. The studies reported in this thesis (Chapters 3 and 4) use the framework: interaction data is presented in full (and open-sourced), the extracted statistics referred to as interaction metrics, behavioural proxies are inferred, and the outcome is clearly stated in each (engagement and abandonment). Across domains there is a wider trend towards open research, particularly around the transparent collection and use of datasets (Gebru et al., 2021). The framework presented here is one such solution to the problems found with reproducible research and it is open to evolution by the community, but the problem is clear and applying a framework like this would improve the quality of science and science communication in the area. The framework is a move towards providing methodological support and does, however, require validation to ensure that it is both usable and practical. The validation of the framework is discussed in the future work (see Section 7.1 in Chapter 7).

Chapter 6

Discussion

Media is changing, with a move towards content which is more tailored to the needs of the individual audience member. One important and common aspect of personalisation is interactivity, where viewers or listeners are given the opportunity to directly or indirectly influence the content they consume. This interactivity presents an opportunity to collect user interactions which we might be able to evaluate the user experience and provide a measure of success for the content. Therefore, this thesis investigated two aspects of success – user engagement and abandonment – and whether user interactions collected from interactive media experiences could be used to understand and model them. In reviewing the literature and investigating both engagement and abandonment, it was found that the literature around interaction data analysis, specifically for modelling and understanding users, had inconsistencies. As such, a framework was also presented to aid in the clear communication of research in the space, to facilitate reproducibility, and to increase transparency.

Engagement

User engagement with interactive media experience was the first point of investigation. Developing an understanding of user engagement (and abandonment) for media producers would allow retrospective insights to be fed into the creative process to improve future media. By leveraging user interactions, there is a possibility for engagement to be monitored at scale and in real-time. Before moving into the realms of real-time monitoring of engagement, an understanding of user engagement within the context of interactive media experiences first needs to be established. Therefore, the following aim was proposed: *to investigate whether there are signals in interaction data that are predictive of engagement*. To develop this understanding and to direct the research, the following three research questions were also proposed:

E-RQ1: Can user engagement with interactive media experiences be inferred from interaction data?

E-RQ2: What interaction-derived metrics are important when inferring engagement?

E-RQ3: Are there commonalities between interactive media experiences?

Two differing interactive media experiences were the focus: one an adaptive tutorial and the other an interactive branching narrative; each having a different structure, interaction opportunities for the audience, and content. For both experiences, user studies were carried out to collect interaction data and a measure of engagement. In the studies, models of engagement were trained using interaction metrics derived from the data and evaluated.

In evaluating the models to assess the extent to which engagement can be predicted (**E-RQ1**), it was found that both engagement and the individual factors (perceived usability, focused attention, aesthetic appeal, and reward) could be predicted with reasonable accuracy in the adaptive tutorial. However, the inclusion of demographics – which had a statistically significant effect on engagement and perceived usability and a data type not collected as part of the interactive branching narrative – tended to reduce the model's predictive capability. These results are counterintuitive and demonstrate that the results can not be relied on. A potential explanation is that it is due to the lack of interaction by users that found the experience less usable and therefore less engaging. Whereas, those with higher engagement (along with more origami making and technology experience) made more use of the interaction opportunities.

For the interactive branching narrative, where participants were divided into two engagement groups (high and low), it was found that group membership could be accurately predicted. These findings demonstrate that engagement, as captured by the User Engagement Scale, a close to ground truth measure, can be predicted using interaction metrics derived from two differing interactive media experiences.

The second research question (**E-RQ2**) was to understand what interaction metrics were important in modelling engagement, and whether there were any signals of behavioural proxies. To answer this, the internal representation of importance and model-agnostic approaches were extracted and calculated.

From the models trained to predict engagement and the individual factors in the adaptive tutorial, it was found that most relied on a single interaction metric to make predictions. Relying on a single metric limits the inference that can be made to understand whether there are behavioural proxies, and suggests that there were not reliable signals in the data.

External factors were found to affect the engagement of participants in the adaptive tutorial. Participants with no prior experience making origami and with less experience in using technology found the tutorial less usable. The interaction metrics showed that these groups of participants used the experience in different ways, with differences in the number of times participants switched camera angles between high and low usability and engagement. These suggest that media producers need to consider the backgrounds of their audience, for example, by taking a direct approach by including questions up-front which could then alter the tutorial to better suit the individual.

For the interactive branching narrative, behavioural proxies relating to different levels of engagement were found. In the high engagement group, we found an increase in the time

it takes a user to finish the show, a higher number of narrative element changes, and more very long pauses. This suggests that these users were sitting back and consuming the content, which can be considered a consumption-type behaviour. In the low engagement group, we found a higher next button count, narrative element change frequency, and shorter pauses between interactions. The combination of these metrics suggest that the users were skipping around the content and exploring what was available, which can be considered a skipping-type behaviour. By monitoring the interaction metrics that make up these behavioural proxies media producers could do a number of things, from retrospectively assessing the success of the content to the real-time measurement of engagement. The latter could further enable media producers to react to changing engagement levels, perhaps through a change in story structure or by nudging users in a direction that may increase their engagement.

A shared importance of interaction metrics or similar behavioural proxies across multiple experiences would allow for media producers to have a set of common evaluation methods (**E-RQ3**). It was found that the metrics that were predictive of engagement differed between the two experiences. The differing content in each of the experiences is likely to be a factor - there are different interaction opportunities in each - along with the nature of how the data was collected: one a more controlled study and the other an in-the-wild study. Additionally, the goals of the users are likely to be different, which affects their interactions. More widely, however, it demonstrates that forms of interactive media experiences need to be evaluated in slightly different ways to understand engagement. It is possible that in experiences sharing a similar format – another branching narrative with a different subject matter – users may express more closely related or the same behavioural signals, enabling a common set of evaluation metrics to be used, but this is left for future work.

The investigation into engagement has demonstrated that it can be predicted using interaction data, that there are signals in the interactions of users to differentiate between engagement levels, but these relationships differ between the experiences studied. These results suggest that this is an avenue of research that is worth pursuing, and that analysis of interaction data offers production teams a potential route for large-scale unobtrusive measurement of audience engagement.

There are, however, limitations. Drawing actionable conclusions from the findings is constrained by the small sample size in both studies which limits the generalisability. Another major limitation is the specific nature of the experiences - we have only explored two. As demonstrated by the adaptive tutorial, the wider context of the user is important in understanding their engagement. In the interactive branching narrative, no information about the users were collected which limits what can be said about the audience drawn to the experience. Also, the non-random sampling of those who took the engagement survey, whether audience members visited the show or completed the survey more than once was not collected; a technique has since been implemented to avoid this in future studies while ensuring the privacy of the audience is respected. Finally, for the interactive branching narrative, it is not clear as to why users are skipping - are they unable to engage with the content because they are skipping or are they skipping because they are not engaged with the content?

Future observational studies could address this by collecting additional data on user intent before using an interactive media experience.

Abandonment

The following aim was proposed: *to investigate whether there are signals in interaction data that are predictive of abandonment*. To assess whether predictive signals exist and to direct the research, the following questions were also proposed:

A-RQ1: Can abandonment in interactive media experiences be modelled using interaction data?

A-RQ2: What interaction-derived metrics are important when inferring abandonment?

A-RQ3: Is there a relationship between abandonment and user engagement?

The dataset used was the wider dataset collected from the deployment and release to the public of the interactive branching narrative. To simplify the analysis, the two main sub-stories of the Click TV show were the focus. The show is a single story, with two main sub-stories where one focuses on technology use in Malawi and the other on self-driving cars. A model was build that used the interaction metrics to predict how far from the end of the story a user would abandon.

It was found that it can be accurately modelled, and that the distribution of the distance-based metric could be captured by the model (**A-RQ1**). Between the two sub-stories, abandonment could be more accurately predicted in the Malawi sub-story, likely due to the users having to interact with the experience to enter the sub-story and therefore providing more information to the model.

A notable factor in the dataset were the low levels of interaction with the experience, limiting the usefulness of most interaction-based metrics – for most types of interaction, a high proportion of users recorded zero events, and the median counts were below one. Therefore, when investigating how the models made their choices and what interaction metrics were relied upon (**A-RQ2**), temporal metrics were found to be important and relied on by the models – as with engagement previously. More specifically, the average amount of time that a user spent on narrative elements and the number of very long pauses (longer than 30 seconds of no interaction between interactions) were important in predicting abandonment.

By applying more in-depth, model-agnostic interpretation techniques, it was found that an increase in the proportion of visibility changes – where the user's web browser tab becomes visible or hidden – increased the likelihood of the models' predicting abandonment. This could provide a metric for media producers to monitor as users move through an experience, continually evaluating their likelihood of abandonment and potentially opening opportunities to reduce or gather additional data on why. The proportion of navigation events (backwards and forwards) were indicative of an increased likelihood of abandonment. This

could demonstrate users exhibiting an exploratory-type behaviour where they are exploring the new form of content by moving backwards and forwards and then leaving. The relationship between the behaviour and abandonment was stronger in the Malawi sub-story, and could provide motivation for media producers to look at abandonment at a per-story level.

In a search context, there is a relationship between satisfaction (analogous to engagement) and abandonment. In exploring whether a similar relationship existed in this context (**A-RQ3**), the models trained previously were used to predict the distance (abandonment) for users that had taken part in the user engagement survey – all of which finished the story. In the evaluation of the predicted distances, it was found that low engagement users were predicted to have higher values and that there was a significant difference between the high and low engagement groups. This implies that users with lower engagement express signals in their data similar to those who abandoned earlier in the story. It establishes a link between engagement and abandonment in this context, and that through the modelling of abandonment a level of inference can be made about the engagement of users.

The investigation into abandonment has demonstrated that it can be predicted from the interactions of users, that temporal metrics are important in its prediction but there are also pointers to exploratory-type behavioural proxies, and that there is a link to engagement. Additionally, the utility of an abandonment metric that can be generated from any interactive media experience was also demonstrated.

However, there are limitations to the findings. The dataset was collected prior to an implemented solution to track users across multiple sessions, so we do not know if an abandonment was temporary (closed window and then came back to the experience later) or permanent. The distinction is important when assessing success. While it has been demonstrated that abandonment can be modelled, the reasons may be external to the experience which do not manifest themselves in the interactions of users, so some level of abandonment will be present without explanation by data.

The link between engagement and abandonment requires further study with additional data collected to strengthen the link. One potential way to achieve this, and to understand whether external factors are at play, is to take an approach similar to that in Diriye et al., 2012 where abandonment is detected and users were asked for rationale. This would allow for engagement-based questions to be asked and whether some external factors caused them to leave.

Framework

This is not the first piece of work that brings together approaches from two domains (human-computer interaction and machine learning) to understand the mapping between interaction data, user behaviours, and high level user traits. However, potential issues in the communication of research were identified and that there is a need to be more rigorous, creating reproducible and transparent research. As such, the following aim was proposed: *based on the above, we will investigate whether a generic framework can be established to guide and*

assist researchers when working with interaction data.

A fresh review of the literature was performed to further explore issues identified as part of the literature review. As a result, issues in the way research is communicated were found, namely the use of the term 'behaviour' and authors unintentionally making unsupported claims about their results – creating barriers to the reproducibility of the work and reducing transparency.

To address these issues and provide methodological support for those researching user behaviour through interaction data, a framework consisting of four layers and providing linguistic guidance was proposed. The framework enables reproducible research to be produced and encourages open-research in the area, with researchers able to frame their work in a consistent way and the consumers of the research able to better understand the work. It is, however, a possible approach to addressing the problems identified and it is open to evolution by the community; nonetheless, the problems identified are clear and by using a framework such as this the quality of science and science communication in the area should improve.

Summary

With a new form of media emerging, methods to evaluate the success are needed. Due to the interactive nature of these media experiences, the interaction data generated by users provides the opportunity to evaluate success at scale and without directly questioning the audience. However, its utility in this emerging domain is untested. The work in this thesis aimed to evaluate the usefulness of interaction data to gather an understanding of user engagement and abandonment, providing measures for two aspects of success for media producers. It was found that both engagement and abandonment could be predicted using interaction data and that there were suggestions of behavioural proxies exhibited by users which could be monitored by media producers to gauge levels of engagement and the likelihood of abandonment.

The findings demonstrate and establish that interaction data can be used to understand and model users in interactive media experiences, providing a scalable and objective measure of success for media producers. Alongside the empirical work presented, a framework to enable clear communication to aid in producing reproducible and open research was also presented. The application of such a framework should improve cognate activity.

Chapter 7

Conclusions

Media content is moving from a traditional one-size-fits-all narrative to a more personalised experience which can adapt to audience context, knowledge, and needs. In the same way that movie pioneers had to discover and develop the craft of film-making, creators of interactive content must learn what works and what does not in this new domain. To do so, it is essential to understand how different aspects of their production – the interaction design, the ways in which it adapts, and the content itself – impact on audience experience, of which two key measures will be engagement with, and the abandonment of, content.

Measuring both aspects of success is challenging, but an approach from other domains that is relatively cheap and works at scale is to capture user interactions and infer engagement and modelling abandonment from these. The thesis aimed to test whether this approach can be applied in this domain.

Signals in interaction were found that can differentiate between levels of engagement and be predictive of abandonment. It was found that engagement can be predicted, that there were behavioural proxies which could differentiate between levels of engagement, and that the user context affects engagement. Whilst for abandonment, it was demonstrated that it can be predicted from the interactions of users, that audience members also exhibit behavioural proxies indicative of abandonment, and that there is a link to engagement. The findings demonstrate that interaction is a viable method to evaluate the success of experiences, specifically in capturing signals for engagement levels and predicting abandonment. Overall, points of consideration for media creators were highlighted. Metrics which are useful to monitor and collect to understand engagement and abandonment rates were found, and behavioural proxies that link interaction metrics to engagement and abandonment were identified.

There are limitations in the work. For example, the users that took part in the engagement survey were self-selecting which limits the generalisability of the findings. However, while engagement measures were not collected for every audience member, a further limitation, the findings demonstrating the link between those with low engagement and behaviours indicative of abandonment provide promising avenues for future work.

Metrics which were found to be useful in the work have been implemented into production environments at the BBC and are actively being collected as part of the analytics[1]. These

[1] https://github.com/bbc/storyplayer/blob/main/docs/analytics.md#segment_completion

include the duration, event counts, amount of the time paused, the hidden time, visible time, and playing time, which are all collected in real-time on the client-side at the narrative element level. The work demonstrated that there is value in the client-side analysis of interactions and sending summaries back to the analytics database, rather than large quantities of data, which has improved the quality of data in the backend and increases privacy. By performing client-side analysis, it means that the story could react and adapt to the metrics. The changes identified by the work impacted both the data collected – the analytics – as well as the data model of the story itself, particularly in the case of a default length (for the narrative element) becoming a core component of the model. Taking another step forward, the collection of these metrics allows for analytical tools to be developed from this data. For example, an analytical dashboard presenting a retrospective summary of success and real-time insights into experience performance.

Meta-issues relating to the communication of research in the interaction data analysis space were identified, with the use of language and lack of transparency (owing to the cross-disciplinary nature of the field) hinder reproducibility and fully understanding the research. As a result, and via a fresh analysis of the literature, a framework was proposed that addressed the issues and provides methodological support for those undertaking research in the area. It is conjectured that using such a framework should improve the quality of science and science communication in the area, with more reproducible and transparent research being produced.

7.1 Future Work

There are several opportunities for further research, building on the empirical work presented in Chapters 3 and 4; this section presents and discusses these possible directions. In the first instance, a reflection on the work is presented and how it might change if additional data were available, along with immediate future work mentioned in previous chapters, is presented. Following this, two alternative, and potentially fruitful, analysis methods which slightly reframe the modelling of abandonment and the users are considered. Finally, natural step in the analytics pipeline, i.e., deployment, is presented; some of the challenges associated with this are discussed and some approaches that could be used to address these are outlined.

It was shown that user context affects engagement and when expanding the work to a wider sample, contextual data was not collected, nor did we have access to a broader user profile for each audience member. Having access to a broader profile would allow for contextual data to be introduced into the analysis process. For example, historical data on what content the user has watched could provide a more detailed understanding for content preferences within a multi-story branching narrative. An analysis of the data could then be performed by exploring whether those who historically preferred the content shown were engaged, and how likely they were to abandon. One further aspect that could be investigated is whether those less familiar with the content were engaged and less likely to leave, a potential strong

signal of success for media producers.

The framework presented in Chapter 5 is a move towards providing methodological support for practitioners. However, whether it is useable and provides the support it aims to provide is currently unknown. Validation is required to ensure that it is both useable and practical. One method of validation is to discuss the framework with the practitioners for whom it is designed to assist. For example, by using semi-structured interviews to gather their thoughts on the identified issues, which motivated its creation, and asking them to apply the framework to their work to find out how their work might change as a result. These conversations could further confirm the identified issues and be used to adapt and alter the framework to make it more applicable.

7.1.1 Survival Analysis

Survival analysis is an approach to analysis where the outcome of interest is the time until an event occurs (Kleinbaum and Klein, 2010). The technique has had successful application within the clinical domain, where it originates, for example in the modelling of remission time for leukaemia patients or the survival time after receiving a heart transplant. Much like in other modelling approaches, important features from modelling the survival time can be extracted from the models which makes the approach useful when wanting to understand what factors affect survival. Most traditional survival analysis methods lie within the realm of non-/semi-parametric and parametric statistical approaches, such as Kaplan-Meier, Cox regression, and Accelerated Failure Time. However, machine learning methods have grown in popularity due to their ability to model non-linear relationships (P. Wang et al., 2019), and there are a range of established methods to use: survival trees, bayesian methods, neural networks, and support-vector machines.

Why might these methods be interesting to explore in the domain of interactive media experiences? User abandonment in media experiences could be a fertile place for this type of analysis, enabling modelling of the time until an audience member leaves an experience and investigating what interaction metrics are important in the prediction. Having a notion of when an audience member is likely to abandon, much like predicting their distance from the end, goes beyond a simpler measure such as a binary (complete/abandon) representation. The results could provide media creators with an understanding as to whether a user is likely to leave part-way through, what types of behavioural proxies are observed before abandonment, and dynamically reorganising the content to suit the predicted time left on the experience.

Due to the adaptive nature of interactive experiences, where their length is variable depending on the content or the choices the user makes, modelling the time to an event poses a challenge. In interactive media experiences the relationship between time spent in an experience and whereabouts in it a user is not constant. This is due to experience being personalised to the user, the enablement of back-tracking in the narrative, or allowing the audience to choose their own path, all of which impact the measurement of time. Censoring is also a

challenge, as is the case in most survival analysis problems. Briefly, censoring is where the event of interest, a user abandoning a piece of media for example, is not observed for some audience members. There are three types of censoring: right-censoring (the most common) where the period of observation finishes before the event is recorded; left-censoring where the event occurs at some point before being observed (exact time of infection prior to a positive test result, for example); and interval-censoring which is where the event occurs during a known time window (Kleinbaum and Klein, 2010; P. Wang et al., 2019).

Depending on how survival analysis is applied to the problem within the context of interactive media experiences, there are different challenges introduced based on the two issues previously discussed. Investigating an experience as a whole, for example modelling the survival of audience members in the Click TV show, introduces issues around variable time – the show can have a variable time based on their path and there is no accurate notion of how long the show lasts, beyond that of the default path. An alternative solution is to look at the components of the show, similar to how sub-stories in the Click TV show were focused on in Chapter 4, as there is less variation in the amount of time users could spend. This would introduce a challenge with censoring, as those that do finish the sub-story may abandon right after, but it is a smaller limitation in comparison to time. A potentially prosperous approach to both is training a global, multi-level model which allows for group-level effects to be modelled, for example the story visited or path taken, and which may have an effect on the outcome of interest (Austin, 2017).

7.1.2 Sequential Modelling

Through the application of statistical tests and modelling, behavioural proxies were found when investigating both engagement and abandonment. Some of these proxies point towards having a sequential nature, for example, skipping through the content (engagement) and exploration (abandonment). This raises a question as to whether sequence-based modelling would effectively capture the dynamics of the audience interaction, leading to more accurate models and an alternative method to understand user differences between engagement levels or abandonment distances.

The sequential modelling of interactions is a popular method, with many state-of-the-art approaches using sequences of clicks or interaction events as their data of choice (Quadrana et al., 2018) - as shown in the review of the literature that led to the proposal of the framework (Chapter 5). In large-scale systems, users generate a large quantity of interaction events, for example, 100 million distinct users' generating interaction-based events in a music listening context (Anderson et al., 2020). Due to the scale of the data, Neural Network-based approaches are popular and an often the deployed model of choice, with variations of Recurrent Neural Network architectures used which can effectively capture the dynamics of sequential data (Lipton et al., 2015; Pouyanfar et al., 2018) and yield highly accurate models.

With these types of approaches, however, there is a trade-off. Whilst highly accurate mod-

els can be trained, fully interpreting the factors in the data that contribute to the prediction is difficult – it is harder to gain a full understanding of why decisions are made. Techniques exist that can be applied to interpret models, for example anchors (Vizer and Sears, 2017) which identifies words that need to be present to ensure a prediction and integrated gradients (Curmi et al., 2017) which attributes an importance value to each input feature of a model based on the model's gradients. However, these add an additional layer of complexity, and the output relies on being able to infer the semantic connection between words that are identified as important – which may not be as straightforward when looking at interaction events.

Alongside methodological challenges, considerations need to be made about the data collected from the interactive media experiences. For example, there are an abundance of narrative element change events, where the audience member moves from one segment of content to another, both due to the nature of the experience and as audiences choose not to interact but rather watch the content. One solution is to encode the number of narrative elements seen into the event itself, for example, rather than a sequence of eight narrative element change events, it is compressed into a single event with the number eight part of it (i.e, NEC_8). Whilst there are several challenges and considerations, the application of sequential modelling could yield highly accurate models – important in an application context, with the caveat that there is an understanding of how decisions are made – and supplement the statistical-based findings presented in the thesis, giving media producers further information about differences in engagement and the actions leading up to abandonment.

7.1.3 Real-Time Modelling

The empirical work presented in the thesis focuses on how the results from modelling engagement and abandonment can be used to assess the success (or failure) of interactive experiences a posteriori. However, the adaptable nature of the experiences – their underlying construction – mean there is an opportunity to create a more dynamic feedback loop, where predictions based on the interactions of users are fed back into the content in a real-time manner. Investigating this area would be a natural progression of the work in Chapters 3 and 4; putting into practise the models that have been trained.

Predicting engagement or abandonment in real-time could enable a multitude of different possibilities for the content, either through further personalisation, nudging the audience member, or to gather additional data to have a clearer picture about user engagement or abandonment rationales. Further personalising the content could include re-organising on-the-fly – previously mentioned as a motivating factor for investigating engagement and abandonment – for example, if the user is likely to abandon or is predicted to be exhibiting low levels of engagement then the user could be given a shorter or simpler version of the narrative. Alternatively, if the audience member is predicted to have low levels of engagement or more likely to abandon, nudging them towards another path in the story could boost their engagement or reduce the likelihood of them abandoning the content early.

Different from the other two research directions, using the models in real-time to predict engagement or abandonment also presents the opportunity to gather more detailed data about the user. For example, when predicting the engagement of audience members, a subset could be asked to provide feedback on their experience to gather more data on why they appear not to be engaged with the content or whether the prediction is accurate (you appear to be highly engaged with this content, would you agree?). Gathering this type of data would allow media creators to have a more in-depth view of their audience and evaluate the models' predictions. This approach has found success in the search domain, with Diriye et al., 2012, investigating rationales for abandonment by directly questioning the users using a pop-up survey.

This direction of future research does, however, have a number of practical challenges associated with it. In the deployment of models, there will always be audience members predicted to be highly engaged with a piece of media but who actually have low levels of engagement. Similarly, a user predicted to have a high likelihood of abandonment may not, in fact, be about to leave the experience. Handling such misclassifications is crucial, as they could be damaging to the narrative as well as the user experience. One approach could be to use prediction thresholds where action, for example nudging the user, should only be taken when the likelihood is over a given threshold. Alternatively, taking an ensemble approach to the problem, where multiple models predict the likelihood of abandonment or engagement, with a prediction threshold could provide additional confidence. Subtle changes to the story, for example, presenting shorter or simpler versions to the audience based on their predicted likelihood of abandoning, may not have an adverse effect on the user experience, but nudging the user or asking the user questions could detract from the media creator's goal with the story. Additionally, nudging the user, could itself have a negative impact on their experience and increase their likelihood to abandon, and there arises a challenge in explaining to the audience their abandonment likelihood or that they appear to not be overly engaged with the content. There should be clarity to the predictions being made and the approach would have to be balanced and carefully considered alongside the ultimate goal of the media creator and narrative.

A challenge that affects the deployment of predictive models is the inconsistent format of interactive media experiences, which is natural given the newness of the area. The differing formats and interaction opportunities for audiences mean that training and deploying a one-size-fits-all model would pose challenges. As historical data is needed from similar experiences to train the models, the differences in underlying content will result in a mismatch of signals where the importance of a metric in one experience may be different in another experience. For example, with engagement being dependent on external demographic factors in the Make-along: Origami Frog experience (Chapter 3.1) and temporal interaction metrics being important in predicting engagement with the Click TV show.

A potential approach which addresses some of the challenges listed above is to develop a centralised federated learning approach to model deployment and continual training. Federated learning is a collaborative model training method where a global model is trained,

distributed to the device (often and routinely used on mobile devices ("Federated Learning: Collaborative Machine Learning without Centralized Training Data", 2017)) which then improves the model by learning from data local to the device, and then updates are propagated to the global model (Q. Li et al., 2021). In the context of interactive media experiences, federated learning could be applied by training a global model for engagement or abandonment on all data captured in previously deployed experiences, providing a starting point for predictions. When a new experience is published, the global model is attached to the experience and becomes the local model which continuously learns from the data captured as part of the deployment, sending updates back to the global model. There is an engineering overhead, as well as a potential issue with performance of the global model, but it could be achievable as interaction data is already automatically captured from the experiences.

References

Agichtein, E., Brill, E., & Dumais, S. (2019). Improving Web Search Ranking by Incorporating User Behavior Information. *SIGIR Forum*, *52*(2), 11–18. https://doi.org/10.1145/3308774.3308778

Anderson, A., Maystre, L., Anderson, I., Mehrotra, R., & Lalmas, M. (2020). Algorithmic Effects on the Diversity of Consumption on Spotify. *Proceedings of the web conference 2020* (pp. 2155–2165). Association for Computing Machinery. https://doi.org/10.1145/3366423.3380281

Apaolaza, A., Harper, S., & Jay, C. (2015). Longitudinal Analysis of Low-Level Web Interaction through Micro Behaviours. *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, 337–340. https://doi.org/10.1145/2700171.2804453

Apaolaza, A., & Vigo, M. (2019). Assisted pattern mining for discovering interactive behaviours on the web. *International Journal of Human-Computer Studies*, *130*, 196–208. https://doi.org/https://doi.org/10.1016/j.ijhcs.2019.06.012

Apley, D. W., & Zhu, J. (2020). Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *82*(4), 1059–1086.

Arapakis, I., Barreda-Ángeles, M., & Pereda-Baños, A. (2019). Interest as a Proxy of Engagement in News Reading: Spectral and Entropy Analyses of EEG Activity Patterns. *IEEE Transactions on Affective Computing*, *10*(1), 100–114. https://doi.org/10.1109/TAFFC.2017.2682089

Arapakis, I., Lalmas, M., & Valkanas, G. (2014). Understanding Within-Content Engagement through Pattern Analysis of Mouse Gestures. *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 1439–1448. https://doi.org/10.1145/2661829.2661909

Arapakis, I., & Leiva, L. A. (2016). Predicting User Engagement with Direct Displays Using Mouse Cursor Information. *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 599–608. https://doi.org/10.1145/2911451.2911505

Attfield, S., Kazai, G., Lalmas, M., & Piwowarski, B. (2011). Towards a science of user engagement (position paper). *WSDM Workshop on User Modelling for Web Applications*, 9–12.

Austin, P. C. (2017). A tutorial on multilevel survival analysis: Methods, models and applications. *International Statistical Review*, *85*(2), 185–203.

Barbieri, N., Silvestri, F., & Lalmas, M. (2016). Improving Post-Click User Engagement on Native Ads via Survival Analysis. *Proceedings of the 25th International Conference on World Wide Web*, 761–770. https://doi.org/10.1145/2872427.2883092

BBC: Click 1000. (2019). Retrieved January 26, 2022, from https://www.bbc.co.uk/taster/pilots/click1000

BBC: Cook-Along Kitchen Experience. (2016). Retrieved January 24, 2022, from https://www.bbc.co.uk/rd/projects/cake

BBC: Instagramification. (2019). Retrieved January 24, 2022, from https://www.bbc.co.uk/taster/pilots/instagramification

BBC: Make-along: Origami Jumping Frog. (2017). Retrieved January 24, 2022, from https://www.bbc.co.uk/taster/pilots/origamimakealong

Belle, A., Hobson, R., & Najarian, K. (2011). A physiological signal processing system for optimal engagement and attention detection. *2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*, 555–561. https://doi.org/10.1109/BIBMW.2011.6112429

Belletti, F., Chen, M., & Chi, E. H. (2019). Quantifying Long Range Dependence in Language and User Behavior to Improve RNNs. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1317–1327.

Bergstra, J., Yamins, D., & Cox, D. D. (2013). Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, I–115–I–123.

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for Hyper-Parameter Optimization. *Proceedings of the 24th International Conference on Neural Information Processing Systems*, 2546–2554.

Black Mirror: Bandersnatch. (2018). Retrieved January 24, 2022, from https://www.imdb.com/title/tt9495224/

Breiman, L. (2001). Random Forests. *Machine learning*, *45*(1), 5–32.

Brown, E. T., Ottley, A., Zhao, H., Lin, Q., Souvenir, R., Endert, A., & Chang, R. (2014). Finding Waldo: Learning about Users from their Interactions. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12), 1663–1672. https://doi.org/10.1109/TVCG.2014.2346575

Brückner, L., Arapakis, I., & Leiva, L. A. (2020). Query Abandonment Prediction with Recurrent Neural Models of Mouse Cursor Movements. *Proceedings of the 29th*

acm international conference on information & knowledge management (pp. 1969–1972). Association for Computing Machinery. https://doi.org/10.1145/3340531.3412126

Cen, Y., Zhang, J., Zou, X., Zhou, C., Yang, H., & Tang, J. (2020). Controllable multi-interest framework for recommendation. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2942–2951. https://doi.org/10.1145/3394486.3403344

Chen, X., Chen, J., Ma, L., Yao, J., Liu, W., Luo, J., & Zhang, T. (2018). Fine-grained video attractiveness prediction using multimodal deep learning on a large real-world dataset. *Companion Proceedings of the The Web Conference 2018*, 671–678. https://doi.org/10.1145/3184558.3186584

Chien, I., Enrique, A., Palacios, J., Regan, T., Keegan, D., Carter, D., Tschiatschek, S., Nori, A., Thieme, A., Richards, D., Doherty, G., & Belgrave, D. (2020). A Machine Learning Approach to Understanding Patterns of Engagement With Internet-Delivered Mental Health Interventions. *JAMA Network Open*, *3*(7), e2010791–e2010791. https://doi.org/10.1001/jamanetworkopen.2020.10791

Constantinides, M., & Dowell, J. (2018). A Framework for Interaction-Driven User Modeling of Mobile News Reading Behaviour. *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, 33–41. https://doi.org/10.1145/3209219.3209229

Curmi, F., Ferrario, M. A., & Whittle, J. (2017). Biometric data sharing in the wild: Investigating the effects on online sports spectators. *International Journal of Human-Computer Studies*, *105*, 56–67. https://doi.org/https://doi.org/10.1016/j.ijhcs.2017.03.008

Das Sarma, A., Gollapudi, S., & Ieong, S. (2008). Bypass Rates: Reducing Query Abandonment Using Negative Inferences. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 177–185. https://doi.org/10.1145/1401890.1401916

Dedieu, A., Mazumder, R., Zhu, Z., & Vahabi, H. (2018). Hierarchical Modeling and Shrinkage for User Session Length Prediction in Media Streaming. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 607–616. https://doi.org/10.1145/3269206.3271700

Diriye, A., White, R., Buscher, G., & Dumais, S. (2012). Leaving so Soon? Understanding and Predicting Web Search Abandonment Rationales. *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, 1025–1034. https://doi.org/10.1145/2396761.2398399

Doherty, K., & Doherty, G. (2018). Engagement in HCI: Conception, Theory and Measurement. *ACM Comput. Surv.*, *51*(5). https://doi.org/10.1145/3234149

Dupret, G., & Lalmas, M. (2013). Absence Time and User Engagement: Evaluating Ranking Functions. *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, 173–182. https://doi.org/10.1145/2433396.2433418

Ekstrand, M. D., Kluver, D., Harper, F. M., & Konstan, J. A. (2015). Letting users choose recommender algorithms: An experimental study. *Proceedings of the 9th ACM Conference on Recommender Systems*, 11–18. https://doi.org/10.1145/2792838.2800195

Fan, S., Zhu, J., Han, X., Shi, C., Hu, L., Ma, B., & Li, Y. (2019). Metapath-guided heterogeneous graph neural network for intent recommendation. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2478–2486. https://doi.org/10.1145/3292500.3330673

Federated Learning: Collaborative Machine Learning without Centralized Training Data. (2017). Retrieved January 12, 2022, from https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of statistics*, 1189–1232.

Ge, S., Wu, C., Wu, F., Qi, T., & Huang, Y. (2020). Graph enhanced representation learning for news recommendation. *Proceedings of The Web Conference 2020*, 2863–2869. https://doi.org/10.1145/3366423.3380050

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., III, H. D., & Crawford, K. (2021). Datasheets for Datasets. *Commun. ACM*, *64*(12), 86–92. https://doi.org/10.1145/3458723

Gledson, A., Asfiandy, D., Mellor, J., Faraj Ba-Dhfari, T. O., Stringer, G., Couth, S., Burns, A., Leroi, I., Zeng, X., Keane, J., Bull, C., Rayson, P., Sutcliffe, A., & Sawyer, P. (2016). Combining Mouse and Keyboard Events with Higher Level Desktop Actions to Detect Mild Cognitive Impairment. *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, 139–145. https://doi.org/10.1109/ICHI.2016.22

Grinberg, N. (2018). Identifying modes of user engagement with online news and their relationship to information gain in text. *Proceedings of the 2018 World Wide Web Conference*, 1745–1754. https://doi.org/10.1145/3178876.3186180

Guo, L., Yin, H., Wang, Q., Chen, T., Zhou, A., & Quoc Viet Hung, N. (2019). Streaming session-based recommendation. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1569–1577. https://doi.org/10.1145/3292500.3330839

Guo, Q., & Agichtein, E. (2012). Beyond Dwell Time: Estimating Document Relevance from Cursor Movements and Other Post-Click Searcher Behavior. *Proceedings of the 21st International Conference on World Wide Web*, 569–578. https://doi.org/10.1145/2187836.2187914

Guo, Q., Yuan, S., & Agichtein, E. (2011). Detecting Success in Mobile Search from Interaction. *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1229–1230. https://doi.org/10.1145/2009916.2010133

Gupta, S., & Maji, S. (2020). Predicting Session Length for Product Search on E-Commerce Platform. *Proceedings of the 43rd international acm sigir conference on research and development in information retrieval* (pp. 1713–1716). Association for Computing Machinery. https://doi.org/10.1145/3397271.3401219

Halawa, S., Greene, D., & Mitchell, J. (2014). Dropout prediction in MOOCs using learner activity features. *Proceedings of the second European MOOC stakeholder summit*, *37*(1), 58–65.

Hariri, N., Mobasher, B., & Burke, R. (2014). Context adaptation in interactive recommender systems. *Proceedings of the 8th ACM Conference on Recommender Systems*, 41–48. https://doi.org/10.1145/2645710.2645753

He, R., Kang, W.-C., & McAuley, J. (2017). Translation-Based Recommendation. *Proceedings of the Eleventh ACM Conference on Recommender Systems*, 161–169.

Hong, L., & Lalmas, M. (2020). Tutorial on Online User Engagement: Metrics and Optimization. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3551–3552. https://doi.org/10.1145/3394486.3406472

Huang, J., White, R. W., & Dumais, S. (2011). No Clicks, No Problem: Using Cursor Movements to Understand and Improve Search. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1225–1234. https://doi.org/10.1145/1978942.1979125

Jia, X., Zhao, H., Lin, Z., Kale, A., & Kumar, V. (2020). Personalized image retrieval with sparse graph representation learning. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2735–2743. https://doi.org/10.1145/3394486.3403324

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 3149–3157.

Kim, Y., Hassan, A., White, R. W., & Zitouni, I. (2014). Modeling Dwell Time to Predict Click-Level Satisfaction. *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, 193–202. https://doi.org/10.1145/2556195.2556220

Kleinbaum, D. G., & Klein, M. (2010). *Survival Analysis* (Vol. 3). Springer.

Kulp, L., Sarcevic, A., Zheng, Y., Cheng, M., Alberto, E., & Burd, R. (2020). Checklist design reconsidered: Understanding checklist compliance and timing of interactions. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3313831.3376853

Lagun, D., & Lalmas, M. (2016). Understanding User Attention and Engagement in Online News Reading. *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, 113–122. https://doi.org/10.1145/2835776.2835833

Lalmas, M., O'Brien, H., & Yom-Tov, E. (2014). *Measuring user engagement* (Vol. 6). Morgan & Claypool Publishers.

Lamba, H., & Shah, N. (2019). Modeling dwell time engagement on visual multimedia. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1104–1113. https://doi.org/10.1145/3292500.3330973

Lehmann, J., Lalmas, M., Yom-Tov, E., & Dupret, G. (2012). Models of User Engagement. *Proceedings of the 20th International Conference on User Modeling, Adaptation, and Personalization*, 164–175. https://doi.org/10.1007/978-3-642-31454-4_14

Li, A., Thom, J., Chandar, P., Hosey, C., Thomas, B. S., & Garcia-Gathright, J. (2019). Search Mindsets: Understanding Focused and Non-Focused Information Seeking in Music Search. *The World Wide Web Conference*, 2971–2977. https://doi.org/10.1145/3308558.3313627

Li, J., Huffman, S., & Tokuda, A. (2009). Good Abandonment in Mobile and PC Internet Search. *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 43–50. https://doi.org/10.1145/1571941.1571951

Li, L., Deng, H., Dong, A., Chang, Y., Baeza-Yates, R., & Zha, H. (2017). Exploring query auto-completion and click logs for contextual-aware web search and query suggestion. *Proceedings of the 26th International Conference on World Wide Web*, 539–548. https://doi.org/10.1145/3038912.3052593

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. (2017). Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization. *J. Mach. Learn. Res.*, *18*(1), 6765–6816.

Li, Q., Wen, Z., Wu, Z., Hu, S., Wang, N., Li, Y., Liu, X., & He, B. (2021). A Survey on Federated Learning Systems: Vision, Hype and Reality for Data Privacy and Protec-

tion. *IEEE Transactions on Knowledge and Data Engineering*, 1–1. https://doi.org/10.1109/TKDE.2021.3124599

Li, Z., Zhao, H., Liu, Q., Huang, Z., Mei, T., & Chen, E. (2018). Learning from history and present: Next-item recommendation via discriminatively exploiting user behaviors. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1734–1743. https://doi.org/10.1145/3219819.3220014

Lin, J., Pan, W., & Ming, Z. (2020). Fissa: Fusing item similarity models with self-attention networks for sequential recommendation. *Fourteenth ACM Conference on Recommender Systems*, 130–139. https://doi.org/10.1145/3383313.3412247

Lipton, Z. C., Berkowitz, J., & Elkan, C. (2015). A Critical Review of Recurrent Neural Networks for Sequence Learning.

Liu, C., White, R. W., & Dumais, S. (2010). Understanding Web Browsing Behaviors through Weibull Analysis of Dwell Time. *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 379–386. https://doi.org/10.1145/1835449.1835513

Liu, M., Mao, J., Liu, Y., Zhang, M., & Ma, S. (2019). Investigating cognitive effects in session-level search user satisfaction. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 923–931. https://doi.org/10.1145/3292500.3330981

Liu, Z., Liu, Z., & Munzner, T. (2020). Data-driven multi-level segmentation of image editing logs. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3313831.3376152

Lu, H., Zhang, M., & Ma, S. (2018). Between Clicks and Satisfaction: Study on Multi-Phase User Preferences and Satisfaction for Online News Reading. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 435–444. https://doi.org/10.1145/3209978.3210007

Lu, H., Zhang, M., Ma, W., Shao, Y., Liu, Y., & Ma, S. (2019). Quality Effects on User Preferences and Behaviorsin Mobile News Streaming. *The World Wide Web Conference*, 1187–1197. https://doi.org/10.1145/3308558.3313751

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–4777.

Ma, J., Zhao, Z., Yi, X., Yang, J., Chen, M., Tang, J., Hong, L., & Chi, E. H. (2020). Off-Policy Learning in Two-Stage Recommender Systems. *Proceedings of The Web Conference 2020*, 463–473. https://doi.org/10.1145/3366423.3380130

Macha, M., Venkitachalam, S., & Pai, D. (2020). Creos: Identifying critical events in online sessions. *Companion Proceedings of the Web Conference 2020*, 337–342. https://doi.org/10.1145/3366424.3382185

Mehrotra, R., Lalmas, M., Kenney, D., Lim-Meng, T., & Hashemian, G. (2019). Jointly Leveraging Intent and Interaction Signals to Predict User Satisfaction with Slate Recommendations. *The World Wide Web Conference*, 1256–1267. https://doi.org/10.1145/3308558.3313613

Mehrotra, R., Shah, C., & Carterette, B. (2020). Investigating listeners' responses to divergent recommendations. *Fourteenth ACM Conference on Recommender Systems*, 692–696. https://doi.org/10.1145/3383313.3418482

Mehrotra, R., Zitouni, I., Hassan Awadallah, A., Kholy, A. E., & Khabsa, M. (2017). User Interaction Sequences for Search Satisfaction Prediction. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 165–174. https://doi.org/10.1145/3077136.3080833

Miroglio, B., Zeber, D., Kaye, J., & Weiss, R. (2018). The Effect of Ad Blocking on User Engagement with the Web. *Proceedings of the 2018 World Wide Web Conference*, 813–821. https://doi.org/10.1145/3178876.3186162

Müller, J., Oulasvirta, A., & Murray-Smith, R. (2017). Control theoretic models of pointing. *ACM Trans. Comput.-Hum. Interact.*, *24*(4). https://doi.org/10.1145/3121431

Nakano, Y. I., & Ishii, R. (2010). Estimating User's Engagement from Eye-Gaze Behaviors in Human-Agent Conversations. *Proceedings of the 15th International Conference on Intelligent User Interfaces*, 139–148. https://doi.org/10.1145/1719970.1719990

Natarajan, N., Shin, D., & Dhillon, I. S. (2013). Which app will you use next? collaborative filtering with interactional context. *Proceedings of the 7th ACM Conference on Recommender Systems*, 201–208. https://doi.org/10.1145/2507157.2507186

Niu, X., Li, B., Li, C., Xiao, R., Sun, H., Deng, H., & Chen, Z. (2020). A dual heterogeneous graph attention network to improve long-tail performance for shop search in e-commerce. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3405–3415. https://doi.org/10.1145/3394486.3403393

Object-Based Media. (2017). Retrieved January 24, 2022, from https://www.bbc.co.uk/rd/object-based-media

O'Brien, H. (2016). Theoretical Perspectives on User Engagement. In H. O'Brien & P. Cairns (Eds.), *Why engagement matters: Cross-disciplinary perspectives of user engagement in digital media* (pp. 1–26). Springer International Publishing. https://doi.org/10.1007/978-3-319-27446-1_1

O'Brien, H. L., Cairns, P., & Hall, M. (2018). A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies*, *112*, 28–39. https://doi.org/https://doi.org/10.1016/j.ijhcs.2018.01.004

O'Brien, H. L., & Lebow, M. (2013). Mixed-methods approach to measuring user experience in online news interactions. *Journal of the American Society for Information Science and Technology*, *64*(8), 1543–1556. https://doi.org/https://doi.org/10.1002/asi.22871

O'Brien, H. L., & Toms, E. G. (2010). The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology*, *61*(1), 50–69. https://doi.org/https://doi.org/10.1002/asi.21229

Oxford Dictionary of English (3 ed.) (3rd ed.). (2010). Oxford University Press. https://doi.org/10.1093/acref/9780199571123.001.0001

Pasricha, R., & McAuley, J. (2018). Translation-Based Factorization Machines for Sequential Recommendation. *Proceedings of the 12th ACM Conference on Recommender Systems*, 63–71. https://doi.org/10.1145/3240323.3240356

Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M. P., Shyu, M.-L., Chen, S.-C., & Iyengar, S. S. (2018). A Survey on Deep Learning: Algorithms, Techniques, and Applications. *ACM Comput. Surv.*, *51*(5). https://doi.org/10.1145/3234150

Quadrana, M., Cremonesi, P., & Jannach, D. (2018). Sequence-Aware Recommender Systems. *ACM Comput. Surv.*, *51*(4). https://doi.org/10.1145/3190616

Ramesh, A., Goldwasser, D., Huang, B., Daumé III, H., & Getoor, L. (2013). Modeling learner engagement in MOOCs using probabilistic soft logic. *NIPS workshop on data driven education*, *21*, 62.

Roth, C., & Koenitz, H. (2019). Bandersnatch, Yea or Nay? Reception and User Experience of an Interactive Digital Narrative Video. *Proceedings of the 2019 ACM International Conference on Interactive Experiences for TV and Online Video*, 247–254. https://doi.org/10.1145/3317697.3325124

Schoenau-Fog, H. (2011). Hooked!–evaluating engagement as continuation desire in interactive narratives, 219–230.

Shen, C., Chen, Y., Guan, X., & Maxion, R. A. (2020). Pattern-growth based mining mouse-interaction behavior for an active user authentication system. *IEEE Transactions on Dependable and Secure Computing*, *17*(2), 335–349. https://doi.org/10.1109/TDSC.2017.2771295

Song, Y., Shi, X., White, R., & Awadallah, A. H. (2014). Context-Aware Web Search Abandonment Prediction. *Proceedings of the 37th International ACM SIGIR Conference*

on Research & Development in Information Retrieval, 93–102. https://doi.org/10.1145/2600428.2609604

Su, N., He, J., Liu, Y., Zhang, M., & Ma, S. (2018). User Intent, Behaviour, and Perceived Satisfaction in Product Search. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 547–555. https://doi.org/10.1145/3159652.3159714

Tanjim, M. M., Su, C., Benjamin, E., Hu, D., Hong, L., & McAuley, J. (2020). Attentive Sequential Models of Latent Intent for Next Item Recommendation. *Proceedings of the web conference 2020* (pp. 2528–2534). Association for Computing Machinery. https://doi.org/10.1145/3366423.3380002

Taylor, C., Veeramachaneni, K., & O'Reilly, U.-M. (2014). Likely to stop? Predicting Stopout in Massive Open Online Courses.

Teo, C. H., Nassif, H., Hill, D., Srinivasan, S., Goodman, M., Mohan, V., & Vishwanathan, S. (2016). Adaptive, personalized diversity for visual discovery. *Proceedings of the 10th ACM Conference on Recommender Systems*, 35–38. https://doi.org/10.1145/2959100.2959171

Thomas, P. (2014). Using Interaction Data to Explain Difficulty Navigating Online. *ACM Trans. Web*, *8*(4). https://doi.org/10.1145/2656343

Tian, Y., Zhou, K., & Pelleg, D. (2021). What and How Long: Prediction of Mobile App Engagement. *ACM Trans. Inf. Syst.*, *40*(1). https://doi.org/10.1145/3464301

Trattner, C., & Elsweiler, D. (2017). Investigating the healthiness of internet-sourced recipes: Implications for meal planning and recommender systems. *Proceedings of the 26th International Conference on World Wide Web*, 489–498. https://doi.org/10.1145/3038912.3052573

Ursu, M. F., Thomas, M., Kegel, I., Williams, D., Tuomola, M., Lindstedt, I., Wright, T., Leurdijk, A., Zsombori, V., Sussner, J., Myrestam, U., & Hall, N. (2008). Interactive TV Narratives: Opportunities, Progress, and Challenges. *ACM Trans. Multimedia Comput. Commun. Appl.*, *4*(4). https://doi.org/10.1145/1412196.1412198

Vasiloudis, T., Vahabi, H., Kravitz, R., & Rashkov, V. (2017). Predicting Session Length in Media Streaming. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 977–980. https://doi.org/10.1145/3077136.3080695

Vigo, M., & Harper, S. (2017). Real-time detection of navigation problems on the world 'wild' web. *International Journal of Human-Computer Studies*, *101*, 1–9. https://doi.org/https://doi.org/10.1016/j.ijhcs.2016.12.002

Vizer, L. M., & Sears, A. (2017). Efficacy of personalized models in discriminating high cognitive demand conditions using text-based interactions. *International Journal of Human-Computer Studies*, *104*, 80–96. https://doi.org/https://doi.org/10.1016/j.ijhcs.2017.03.001

Wampfler, R., Klingler, S., Solenthaler, B., Schinazi, V. R., & Gross, M. (2020). Affective state prediction based on semi-supervised learning from smartphone touch data. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3313831.3376504

Wan, M., & McAuley, J. (2018). Item Recommendation on Monotonic Behavior Chains. *Proceedings of the 12th ACM Conference on Recommender Systems*, 86–94. https://doi.org/10.1145/3240323.3240369

Wang, D., Jiang, M., Syed, M., Conway, O., Juneja, V., Subramanian, S., & Chawla, N. V. (2020). Calendar Graph Neural Networks for Modeling Time Structures in Spatiotemporal User Behaviors. *Proceedings of the 26th acm sigkdd international conference on knowledge discovery & data mining* (pp. 2581–2589). Association for Computing Machinery. https://doi.org/10.1145/3394486.3403308

Wang, P., Li, Y., & Reddy, C. K. (2019). Machine Learning for Survival Analysis: A Survey. *ACM Comput. Surv.*, *51*(6). https://doi.org/10.1145/3214306

Wang, W., Zhang, W., Liu, S., Liu, Q., Zhang, B., Lin, L., & Zha, H. (2020). Beyond clicks: Modeling multi-relational item graph for session-based target behavior prediction. *Proceedings of The Web Conference 2020*, 3056–3062. https://doi.org/10.1145/3366423.3380077

Webster, J., & Ho, H. (1997). Audience Engagement in Multimedia Presentations. *SIGMIS Database*, *28*(2), 63–77. https://doi.org/10.1145/264701.264706

Williams, K., Kiseleva, J., Crook, A. C., Zitouni, I., Awadallah, A. H., & Khabsa, M. (2016). Detecting Good Abandonment in Mobile Search. *Proceedings of the 25th International Conference on World Wide Web*, 495–505. https://doi.org/10.1145/2872427.2883074

Williams, K., & Zitouni, I. (2017). Does That Mean You're Happy? RNN-Based Modeling of User Interaction Sequences to Detect Good Abandonment. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 727–736. https://doi.org/10.1145/3132847.3133035

Xu, X., Hassan Awadallah, A., T. Dumais, S., Omar, F., Popp, B., Rounthwaite, R., & Jahanbakhsh, F. (2020). Understanding user behavior for document recommendation. *Proceedings of The Web Conference 2020*, 3012–3018. https://doi.org/10.1145/3366423.3380071

Yan, A., Lee, M. J., & Ko, A. J. (2017). Predicting abandonment in online coding tutorials. *2017 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, 191–199. https://doi.org/10.1109/VLHCC.2017.8103467

Yao, J., Dou, Z., Xu, J., & Wen, J.-R. (2020). Rlper: A reinforcement learning model for personalized search. *Proceedings of The Web Conference 2020*, 2298–2308. https://doi.org/10.1145/3366423.3380294

Yom-Tov, E., Lalmas, M., Baeza-Yates, R., Dupret, G., Lehmann, J., & Donmez, P. (2013). Measuring inter-site engagement. *2013 IEEE International Conference on Big Data*, 228–236. https://doi.org/10.1109/BigData.2013.6691579

You vs. Wild. (2019). Retrieved June 24, 2022, from https://www.imdb.com/title/tt10044952/

Youngmann, B., & Yom-Tov, E. (2018). Anxiety and information seeking: Evidence from large-scale mouse tracking. *Proceedings of the 2018 World Wide Web Conference*, 753–762. https://doi.org/10.1145/3178876.3186156

Yuan, F., He, X., Jiang, H., Guo, G., Xiong, J., Xu, Z., & Xiong, Y. (2020). Future Data Helps Training: Modeling Future Contexts for Session-Based Recommendation. *Proceedings of The Web Conference 2020*, 303–313.

Zagermann, J., Pfeil, U., von Bauer, P., Fink, D., & Reiterer, H. (2020). "it's in my other hand!" – studying the interplay of interaction techniques and multi-tablet activities. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3313831.3376540

Zhang, H., Li, Y., Ding, B., & Gao, J. (2020). Practical Data Poisoning Attack against Next-Item Recommendation. *Proceedings of the web conference 2020* (pp. 2458–2464). Association for Computing Machinery. https://doi.org/10.1145/3366423.3379992

Zhang, Y., Zhu, Z., He, Y., & Caverlee, J. (2020). Content-collaborative disentanglement representation learning for enhanced recommendation. *Fourteenth ACM Conference on Recommender Systems*, 43–52. https://doi.org/10.1145/3383313.3412239

Zhao, J., Zhou, Z., Guan, Z., Zhao, W., Ning, W., Qiu, G., & He, X. (2019). Intentgc: A scalable graph convolution framework fusing heterogeneous information for recommendation. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2347–2357. https://doi.org/10.1145/3292500.3330686

Zhao, X., Louca, R., Hu, D., & Hong, L. (2020). The difference between a click and a cart-add: Learning interaction-specific embeddings. *Companion Proceedings of the Web Conference 2020*, 454–460. https://doi.org/10.1145/3366424.3386197

Zhong, E., Fan, W., Wang, J., Xiao, L., & Li, Y. (2012). Comsoc: Adaptive transfer of user behaviors over composite social network. *Proceedings of the 18th ACM SIGKDD*

International Conference on Knowledge Discovery and Data Mining, 696–704. https://doi.org/10.1145/2339530.2339641

Zhuang, M., Demartini, G., & Toms, E. G. (2017). Understanding Engagement through Search Behaviour. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 1957–1966. https://doi.org/10.1145/3132847.3132978

Zhuang, M., Toms, E. G., & Demartini, G. (2018). Can User Behaviour Sequences Reflect Perceived Novelty? *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 1507–1510. https://doi.org/10.1145/3269206.3269243