Evaluating Post-Hoc Explanations In Recommendation Algorithms Within Citizen Science

Noah Brooks

4th Year Project Report
Artificial Intelligence and Computer Science
School of Informatics
University of Edinburgh

2024

Abstract

This skeleton demonstrates how to use the `infthesis` style for undergraduate dissertations in the School of Informatics. It also emphasises the page limit, and that you must not deviate from the required style. The file `skeleton.tex` generates this document and should be used as a starting point for your thesis. Replace this abstract text with a concise summary of your report.

Research Ethics Approval

Instructions: *Agree with your supervisor which statement you need to include. Then delete the statement that you are not using, and the instructions in italics.*

Either complete and include this statement:

This project obtained approval from the Informatics Research Ethics committee.
Ethics application number: ???
Date when approval was obtained: YYYY-MM-DD
[If the project required human participants, edit as appropriate, otherwise delete:]
The participants' information sheet and a consent form are included in the appendix.

Or include this statement:

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Noah Brooks*)

Acknowledgements

Any acknowledgements go here.

Table of Contents

Chapter 1

Introduction

The preliminary material of your report should contain:

- The title page.
- An abstract page.
- Declaration of ethics and own work.
- Optionally an acknowledgements page.
- The table of contents.

As in this example `skeleton.tex`, the above material should be included between:

```
\begin{preliminary}
    ...
\end{preliminary}
```

This style file uses roman numeral page numbers for the preliminary material.

The main content of the dissertation, starting with the first chapter, starts with page 1. ***The main content must not go beyond page 40.***

The report then contains a bibliography and any appendices, which may go beyond page 40. The appendices are only for any supporting material that's important to go on record. However, you cannot assume markers of dissertations will read them.

You may not change the dissertation format (e.g., reduce the font size, change the margins, or reduce the line spacing from the default single spacing). Be careful if you copy-paste packages into your document preamble from elsewhere. Some LaTeX packages, such as `fullpage` or `savetrees`, change the margins of your document. Do not include them!

Over-length or incorrectly-formatted dissertations will not be accepted and you would have to modify your dissertation and resubmit. You cannot assume we will check your submission before the final deadline and if it requires resubmission after the deadline to conform to the page and style requirements you will be subject to the usual late penalties based on your final submission time.

1.1 Using Sections

Divide your chapters into sub-parts as appropriate.

1.2 Citations

Citations, such as Arimura (1997) or (Chang and Keisler, 1990), can be generated using `BibTeX`. We recommend using the `natbib` package (default) or the newer `biblatex` system.

You may use any consistent reference style that you prefer, including "(Author, Year)" citations.

Chapter 2

Background

2.1 Increasing User Participation in Citizen Science

In Citizen Science, the average user participates in a low number of projects and for a short period of time. Research has sought to understand what factors explain the extent of a user's involvement due to the long term sustainability of citizen science relying on continued participation. Consequently, these motivation factors can be split into explaining the initial and long term reasons for participation. Within the context of ecology-related citizen science, it was found that in the short term, personal interest was an important motivator. Strong motivation at the start of project participation indicated good long term involvement. Other important aspects related to how valued the participant felt by the scientists and project community Rotman et al. (2014). Likewise, Rotman et al. (2012) finds that due to the "temporal nature of user motivation" that a lack of explicitly recognising their motivations will lead to a decline in participation. Importantly, the motivations can vary due to the category of a project so targeting what works for one project may not be effective for the context of another Geoghegan et al. (2016). Another key difficulty is that the differences between the scientist's motivations and the participant may be an obstacle to collaboration, as it is said to have a particular effect with regards to their "decision to participate in a project, and the ensuing decision to participate in more tasks or continue participating in the project for an extended period of time after the initial task was completed" Rotman et al. (2012). Several different approaches have used such factors and problems as a starting point for improving user participation.

In Goncalves et al. (2013) the authors used a digital public display to try and involve passers-by in completing a non time bound task: identify images of infected blood cells. They displayed 4 different messages on the display describing the task to be completed using different motivating language. Notably, a control message versus an enjoyment based, a community based, and one of both the latter two. They found that the participants who saw the more motivational task description participated more while also completing the task more accurately and continued for longer than those who hadn't. Another example of increasing participant motivation is shown in Gonçalves et al. (2014). The authors sought to gather feedback on the topic of public transit from

its users. They sent 'motivational' messages of differing sentiment to users once a day to ascertain the effect of their content on the receiver's likelihood to continue to provide feedback. This resulted in users who received messages aligned with the idea of making them feel valued, participating significantly more than those who did not. Similarly, in Kamar et al. (2016) motivational messages are shown to users with the goal of keeping them engaged with the online platform they are completing their tasks on. The content of these messages follow mostly the same theme: emphasising user value and community. One other key difference between this study and the ones previously mentioned is that an algorithm is employed to predict the point of user disengagement. When these messages were shown at this predicted point, the user value messages were found to increase the number of tasks completed compared to no intervention. An interesting finding of this study is that randomly timed interventions were as effective as no intervention. Moreover, the community messages were not statistically significant (they contained a URL redirecting out of the task). This seems to suggest that although motivational prompts improve user motivation and thus participation, if they are disruptive, this may negate the improvement.

One of the important motivators for participants is genuine interest in the project and a chance to learn something new Rotman et al. (2014). However, the number of citizen science projects available has continued to grow due to its increased accessibility via the internet. This means that finding a project that aligns with the aforementioned motivations can be more difficult. This challenge could be a significant obstacle to participation: Ngo et al. (2022) finds that "the intention to participate in contributory citizen science highly correlates with the intention to participate in the sample projects" that they present to the respondents in their survey. A common solution used in other domains, such as e-commerce, is recommendation systems based on user data. Considering that 'SciStarter?' holds demographic data and that West et al. (2021) found that people from different demographic groups were likely to hold different clusters of motivation, this could be used as a predictive tool for catering to users' motivations. In fact, recommendation systems have been used for citizen science, in Ben Zaken (2021) they employed matrix factorisation of each user's past project interaction data to make recommendations for future projects. This method yielded an increase in the average number of times users were active in a project for an extended period, suggesting increased long term motivation. One issue that arose from this method was popularity bias: 90% of projects recommended over a 6 month period were from the top 20% of projects. Interestingly, the solution to this, penalising popularity in the recommendation algorithm, ended up increasing overall participation Sultan et al. (2022). This again suggests the importance of variety or a new topic as motivation for participation.

2.2 The Use of Explanations in Recommendation Systems

The increasing use of explanations in recommendations follows a rise in complexity of recommendation algorithms. This is accompanied by more research showing that users are more likely to adopt and trust an algorithm's recommendations if it can convey some

reasoning behind them which has the effect of increasing a user's perceived personalisation Zhang and Curley (2018). Many forms of explanations exist and the manner of generating them can also differ greatly. Zhang and Chen (2020) splits the types of explanations into six categories: similarity between different users, recommendation features matching the user, other users' views on the recommendation, reasons based on the user's social network and textual or visual explanations. Establishing the goal of an explanation within its context is necessary to best choose its form and how to create it. In Afchar et al. (2022) the author discusses the process of evaluating how best to explain music recommendations made to its users. The scope of the explanation is split into local and global: whether to provide explanations for the algorithm as a whole or for each user (input) recommendation (output) pair. They also describe the interpretability of the explanation as being either intrinsic or post-hoc: the former being where the explanation coincides with the recommendation model and the latter where the explanation is best estimated separately from the model. One key implication of this in the context of recommendations, is that compared to intrinsic interpretability, post-hoc methods "disentangle model design from explanation design, allowing to consider XAI systems in a later stage, or to apply them to already working models" Afchar et al. (2022). Within the context of citizen science, in Ben Zaken et al. (2022), the post-hoc and local explanation approach was taken with a focus on a few of the aforementioned explanation types. The paper found that including such explanations did increase user satisfaction however there was no statistical difference in the rate of engagement. Therefore, it is hard to say for this case the effect that explanations had on users' motivation nor on long term sustained participation as this was not recorded.

In the example given above, the algorithm used for generating justifications was purposely designed to be more interpretable. In contrast to this, Zhang and Chen (2020) presents other models that use deep learning to return explanations. By consequence of this, it highlights that when the explanations are generated post-hoc this leads to a situation where "the recommendation and explanation models models are still black boxes". Despite this, such explanation models have been shown to be effective with regards to certain explainability goals. Musto et al. (2021) creates natural language explanations for movie and book recommendations based on their reviews and uses neural language models in the text generation phase. A user study also demonstrated that this algorithm was preferred over their baseline in every measured aspect, such as engagement. Another point worth noting is that long explanations performed better, in almost all the recorded metrics, than short ones. The study also introduced another explanation algorithm that generated a context aware text justification based on a questionnaire completed by the user prior to restaurant recommendations being made. A minimum of one contextual answer had to be given such as family-meal or breakfast. This data was used to choose a template for the final explanation in the hopes of making it more effective. When more than one contextual aspect was provided the majority of participants preferred this explanation over one that did not take context into account. Interestingly, when only one piece of contextual information was provided most user were indifferent between the explanation types, perhaps suggesting that personalisation in recommendation explanations is only more effective when the user has multiple conflicting requirements. Again within the context of music recommendation, a generated playlist's title is a type of explanation for what it contains. In van Bree (2023), the

author evaluates the effects of the wording in the titles on the choice of playlist by users. Specifically, they used personal pronouns to make the the titles appear more personal. They recorded a significant preference for playlists with a personalised description over those without it. However, they do not evaluate whether this leads to higher engagement overall and nor, in this case, are the titles truly personalised for the user's characteristics. Chatti et al. (2022) brings together two different aspects from the previously mentioned examples: how to choose the extent of a text explanation with regards to the user's characteristics. They sought to explore the relationship between user characteristics and the favoured explanation complexity within the context of social media posts. This was done by categorising six different characteristics and seven goals for explanations. The study found and stated key guidance for effective recommendation explanation design based on its data: the explanation design should be led by the user's goal, share a variable amount of detail and auto-personalise the content it contains based on user characteristics.

2.3 Text Summarisation and Personalisation Using Natural Language Processing

2.3.1 Text Summarisation

Natural language processing techniques to summarise long or multiple texts are primarily categorised as either extractive or abstractive Kouris et al. (2021). Extractive being the method of identifying key words or sentences in the source text and then compiling them together to make a summary from those words or phrases. On the other hand, abstractive summarisation is closer to how a human would summarise text: involves processing the whole text so as to understand the information conveyed and thus generate new sentences to best condense the text while preserving details. Abstractive summarisation methods are themselves broadly classified into structure, semantic or neural based approaches Kouris et al. (2021). Structural approaches make use of pre-defined structures such as templates to build the final summary from abstracted features whereas the semantic approach will feed a semantic representation of the text into a language model. Neural based approaches need only use "an appropriate neural network model" and "are often based on seq2seq models of encoder-decoder architecture" Kouris et al. (2021).

2.3.2 Attention Sequence to Sequence Models

A sequence to sequence model (seq2seq) is a type of neural network architecture designed for tasks involving variable-length input sequences that are mapped into variable-length output sequences. It involves an encoder which may take a sentence (in the context of natural language processing) and subdivide it into words taken one at a time to produce a fixed size representation called a hidden state. Recurrent Neural Networks (RNNs) are often used as the encoder as they are able handle sequential and variable sized inputs. Each time the next piece of data is inputted, the RNN updates the hidden state based on the current input and the previous state. Thus the RNN allows for

the encoder to have a 'contextual' representation of the sentence at current input word. The RNN can also be implemented bidirectionally so both information from previous and following states are used in the encoding. The hidden state is then outputted and this is used as the decoder's input. The decoder then utilises the current hidden state and all previous decoded outputs to generate the next most likely output. Similarly to the encoder, RNNs are again used for the same stated properties. Training the model involves tuning the weights of the respective networks with the goal of minimising a loss function which quantifies the difference between the output sentences and the target sentences. This type of seq2seq model can however struggle with long input data (long sentences or documents) due to compressing all the previous encoder hidden states into one fixed sized hidden state for decoding Bahdanau et al. (2016). Therefore, an attention Seq2Seq model was proposed for abstractive text generation for machine translation. The encoding is done as before but for the entire input sentence, the hidden state passed to the decoder at a particular word contains (for a bidirectional RNN) all the past states and future states, they are however weighted with respect to their similarity to the state at the current word. Therefore, the decoder will be aware of context at a sentence level while paying particular attention to similar features with respect to the current hidden state when making predictions of the next word Bahdanau et al. (2016). This was then adapted for text summarisation in Nallapati et al. (2016). The authors employed various techniques specific to the needs of text summarisation such as limits on the size of output vocabulary and modeling for rare words. This resulted in significant improvements for retaining contextual information when summarising long texts with abstraction.

2.3.3 Summary Text Personalisation

In e-commerce several different papers document a process for using customer's historical data to summarise product reviews in a personalised way. In Li et al. (2019) they utilise an attention seq2seq model to present summaries of another user's review on hotels highlighting its features predicted to be most important to the reader and using their own writing style. A representation of a user was embedded into the encoding stage and the user's vocabulary was used at the decoding step. No user study was conducted but when looking at historical user data from Trip-advisor they measured the inclusion of key desired aspects, such as service and location, in their summarised reviews and compared this to their inclusion in gold standard summaries. They found that their model outperformed by a large margin the other state-of-the-art models on test. Another interesting approach is to write summaries using the user's predicted preferences for the structure and style of the text. In Chen et al. (2020) the author sought to increase click-through rates on emails by personalising the subject line. They use a travel website's organised tours data set. This contains subject-description pairs and user click history where the first click per session is taken to have the most enticing subject line and thus is the ground truth. The model is split into two parts. Firstly, given a current description and subject, find the closest matching description in the training corpus by word frequency and semantic distance and use its summary as the template. Secondly, using an attention seq2seq model summarise the description using an encoding of the user click history and chosen template. (They also add an additional

step to utilise user feedback on subject lines). They report that in their user study 61.4% of participants preferred the subject line generated by their model when compared to a leading template summarisation model. This suggests that personalised headers may be a good tool for improving engagement. They do not, however, compare this to subject lines that remain unchanged nor report how this affects the click-through rate.

2.4 (User Studies?)

2.5 (Persuasive Methods In Explainability?)

A dissertation usually contains several chapters.

Chapter 3

Conclusions

3.1 Final Reminder

The body of your dissertation, before the references and any appendices, *must* finish by page 40. The introduction, after preliminary material, should have started on page 1.

You may not change the dissertation format (e.g., reduce the font size, change the margins, or reduce the line spacing from the default single spacing). Be careful if you copy-paste packages into your document preamble from elsewhere. Some LaTeX packages, such as `fullpage` or `savetrees`, change the margins of your document. Do not include them!

Over-length or incorrectly-formatted dissertations will not be accepted and you would have to modify your dissertation and resubmit. You cannot assume we will check your submission before the final deadline and if it requires resubmission after the deadline to conform to the page and style requirements you will be subject to the usual late penalties based on your final submission time.

Bibliography

Afchar, D., Melchiorre, A., Schedl, M., Hennequin, R., Epure, E., and Moussallam, M. (2022). Explainability in music recommender systems. *AI Magazine*, 43(2):190–208.

Arimura, H. (1997). Learning acyclic first-order horn sentences from entailment. In *Proc. of the 8th Intl. Conf. on Algorithmic Learning Theory, ALT '97*, pages 432–445.

Bahdanau, D., Cho, K., and Bengio, Y. (2016). Neural machine translation by jointly learning to align and translate.

Ben Zaken, D., Segal, A., Cavalier, D., Shani, G., and Gal, K. (2022). Generating recommendations with post-hoc explanations for citizen science. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '22, page 69–78, New York, NY, USA. Association for Computing Machinery.

Ben Zaken, Gal, S. S. C. (2021). Intelligent recommendations for citizen science. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14693–14701.

Chang, C.-C. and Keisler, H. J. (1990). *Model Theory*. North-Holland, third edition.

Chatti, M. A., Guesmi, M., Vorgerd, L., Ngo, T., Joarder, S., Ain, Q. U., and Muslim, A. (2022). Is more always better? the effects of personal characteristics and level of detail on the perception of explanations in a recommender system. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '22, page 254–264, New York, NY, USA. Association for Computing Machinery.

Chen, Y.-H., Chen, P.-Y., Shuai, H.-H., and Peng, W.-C. (2020). Tempest: Soft template-based personalized edm subject generation through collaborative summarization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7538–7545.

Geoghegan, H., Dyke, A., Pateman, R., West, S., and Everett, G. (2016). Understanding motivations for citizen science. *Final report on behalf of UKEOF, University of Reading, Stockholm Environment Institute (University of York) and University of the West of England.*

Goncalves, J., Ferreira, D., Hosio, S., Liu, Y., Rogstadius, J., Kukka, H., and Kostakos, V. (2013). Crowdsourcing on the spot: Altruistic use of public displays, feasibility, performance, and behaviours. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '13, page 753–762. Association for Computing Machinery.

Gonçalves, J., Kostakos, V., Karapanos, E., Barreto, M., Camacho, T., Tomasic, A., and Zimmerman, J. (2014). Citizen motivation on the go: The role of psychological empowerment. *Interacting with Computers*, 26(3):196–207.

Kamar, E., Horvitz, E., Bowyer, A., and Miller, G. (2016). Intervention strategies for increasing engagement in crowdsourcing: Platform, predictions, and experiments.

Kouris, P., Alexandridis, G., and Stafylopatis, A. (2021). Abstractive Text Summarization: Enhancing Sequence-to-Sequence Models Using Word Sense Disambiguation and Semantic Content Generalization. *Computational Linguistics*, 47(4):813–859.

Li, J., Li, H., and Zong, C. (2019). Towards personalized review summarization via user-aware sequence network. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Musto, C., de Gemmis, M., Lops, P., and Semeraro, G. (2021). Generating post hoc review-based natural language justifications for recommender systems. *User Modeling and User-Adapted Interaction*, 31:629–673.

Nallapati, R., Zhou, B., dos santos, C. N., Gulcehre, C., and Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence rnns and beyond.

Ngo, M. K., Altmann, C. S., and Klan, F. (2022). How the general public appraises contributory citizen science: Factors that affect participation. *Citizen science: Theory and Practice*.

Rotman, D., Hammock, J., Preece, J., Hansen, D., Boston, C., Bowser, A., and He, Y. (2014). Motivations affecting initial and long-term participation in citizen science projects in three countries. *IConference 2014 Proceedings*.

Rotman, D., Preece, J., Hammock, J., Procita, K., Hansen, D., Parr, C., Lewis, D., and Jacobs, D. (2012). Dynamic changes in motivation in collaborative citizen-science projects. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, page 217–226, New York, NY, USA. Association for Computing Machinery.

Sultan, A., Segal, A., Shani, G., and Gal, Y. K. (2022). Addressing popularity bias in citizen science. In *Proceedings of the 2022 ACM Conference on Information Technology for Social Good*, GoodIT '22, page 17–23, New York, NY, USA. Association for Computing Machinery.

van Bree, L. (2023). Framing theory on music streaming platforms: How vocabulary influences the user experience.

West, S., Dyke, A., and Pateman, R. (2021). Variations in the motivations of environmental citizen scientists. *Citizen Science: Theory and Practice*.

Zhang, J. and Curley, S. P. (2018). Exploring explanation effects on consumers' trust in online recommender agents. *International Journal of Human–Computer Interaction*, 34(5):421–432.

Zhang, Y. and Chen, X. (2020). Explainable recommendation: A survey and new perspectives. *Foundations and Trends® in Information Retrieval*, 14(1):1–101.

Appendix A

First appendix

A.1 First section

Any appendices, including any required ethics information, should be included after the references.

Markers do not have to consider appendices. Make sure that your contributions are made clear in the main body of the dissertation (within the page limit).

Appendix B

Participants' information sheet

If you had human participants, include key information that they were given in an appendix, and point to it from the ethics declaration.

Appendix C

Participants' consent form

If you had human participants, include information about how consent was gathered in an appendix, and point to it from the ethics declaration. This information is often a copy of a consent form.