Towards Trustworthiness in the context of Explainable Search

Sayantan Polley
Otto von Guericke University
Magdeburg, Germany
sayantan.polley@ovgu.de

Rashmi Raju Koparde
Otto von Guericke University
Magdeburg, Germany
rashmi.koparde@st.ovgu.de

Akshaya Bindu Gowri
Otto von Guericke University
Magdeburg, Germany
akshaya.bindu@ovgu.de

Maneendra Perera
Otto von Guericke University
Magdeburg, Germany
hetti.perera@st.ovgu.de

Andreas Nürnberger
Otto von Guericke University
Magdeburg, Germany
andreas.nuernberger@ovgu.de

ABSTRACT

Explainable AI (XAI) is currently a vibrant research topic. However, the absence of ground truth explanations makes it difficult to evaluate XAI systems such as Explainable Search. We present an Explainable Search system with a focus on evaluating the XAI aspect of Trustworthiness along with the retrieval performance. We present SIMFIC 2.0 (Similarity in Fiction), an enhanced version of a recent [1] explainable search system. The system retrieves books similar to a selected book in a query-by-example setting. The motivation is to explain the notion of similarity in fiction books. We extract hand-crafted interpretable features for fiction books and provide global explanations by fitting a linear regression and local explanations based on similarity measures. The Trustworthiness facet is evaluated using user studies, while the ranking performance is compared by analysis of user clicks. Eye tracking is used to investigate user attention to the explanation elements when interacting with the interface. Initial experiments show statistically significant results on the Trustworthiness of the system, paving way for interesting research directions that are being investigated.

CCS CONCEPTS

• **Computing methodologies → Artificial intelligence**; • **Information systems → Evaluation of retrieval results**.

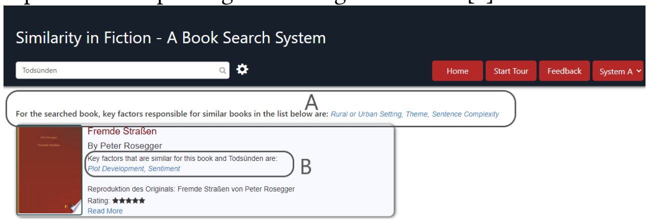
KEYWORDS

XAI; Explainability; XIR; user trust; user study; eye-tracker

ACM Reference Format:
Sayantan Polley, Rashmi Raju Koparde, Akshaya Bindu Gowri, Maneendra Perera, and Andreas Nürnberger. 2021. Towards Trustworthiness in the context of Explainable Search. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3404835.3462799

1 INTRODUCTION

Explainable Artificial Intelligence (XAI) systems attempt to “uncover” the hidden logic behind the decisions made by an AI system. The idea is to provide transparency to the end-user and gain user trust. Often this is a regulatory requirement such as the EU GDPR “Right to an explanation”. XAI can have different perspectives depending on the scenario - classification or retrieval. In a classification setting, the focus is often on the development of add-on methods like LIME [2], LRP [3] to explain a classification decision. While in the IR setting, the focus is on explaining the relative and global rankings ([1, 4, 5]). In the context of searching books, the goal is to explain the notion of similarity in the text to end-users. The similarity between books may depend on subjective aspects such as writing style, emotions, and related aspects from the Humanities perspective. As it often happens in IR systems, there are challenges in obtaining ground truth relevance. The problem is aggravated in the XAI setting when we also do not have ground truth explanations. Hence we attempt to focus on the evaluation aspect while improving an existing XAI search [1].



For the searched book, key factors responsible for similar books in the list below are: *Rural or Urban Setting, Theme, Sentence Complexity*

Key factors that are similar for this book and Todsünden are:
Plot Development, Sentiment

Reproduktion des Originals: Fremde Straßen von Peter Rosegger

Figure 1: SIMFIC 2.0 User Interface. Part A shows the Global Explanation. Part B shows a Local Explanation.

As a new approach in finding fiction books, SIMFIC 1.0 [1] introduced an explainable book search system. Every book (selected from 19th-century fiction books from Project Gutenberg[1]) was represented by a compact (twenty two features), hand-crafted interpretable feature space. The similarity between the user selected query book and all other books was computed by comparing the book feature vectors. It provided a global level explanation for the top-10 books on the home page based on feature selection using classifiers. The retrieval scenario is framed into a binary classification setting where the top ten relevant books belong to class one while all other books of the corpus belong to class two. Global explanation (see Part A in Fig. 1) displays the key features that make the top ten books different from other books. SIMFIC 1.0 gave promising results when the rankings were compared to a retrieval model based on bag-of-words feature representation. But it lacks some aspects such as providing local explanations for each retrieved book and did not evaluate the explanations on specific XAI aspects.

[1] https://www.gutenberg.org/

SIGIR ’21, July 11–15, 2021, Virtual Event, Canada.
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8037-9/21/07.
https://doi.org/10.1145/3404835.3462799

The objective of SIMFIC 2.0 is to address these shortfalls with a focus on XAI evaluation. Contributions of this work are:

- *Generate Local Explanations*: We develop methods to generate local explanations (see Part B in Fig. 1) exploring similarity measures. Besides, we explore a new global explanation method by posing the problem as a linear regression setting rather than a binary classification (described above).
- *Generalization on other Languages*: Investigate the generalization of the overall approach on German books.
- *XAI Evaluation*: We evaluate the specific aspects of XAI evaluation, such as Trustworthiness in a user study, by adapting the existing definitions from the XAI community.
- *Ranking Evaluation*: We compare retrieval performance by analyzing user click-through data.

We find encouraging results showing the effectiveness of our approach. This paves way for investigating the usage of our search system[2] for book selling platforms and extending the idea of comparing compact features in domains such as law books.

2 RELATED WORK

Chiang et al. [6] researched on finding books of specific genre based on the title and cover of the book. Alharthi et al. [7] explored linguistic features present in books for recommending relevant books using feature selection of different categories . There are a plethora of publications on explainable search focused on neural rankers [4, 5] along with the use of causality in search [8]. However, works on the evaluation of explanations (XAI) are limited. Mohseni et al. [9] performed in-depth research and categorized different aspects for different types of explanations. In another evaluation study, Shlomo et al. [10], consider the trust facet with three dimensions: presentation, explanation, and priority. Based on the ranking of these factors, they check the difference noticed in the user's habits. From the search user interface perspective, researchers [11] use eye tracker software for tracking the eye movements of the user to investigate specific facets on the interface with a focus on user age groups.

3 SYSTEM APPROACH

The idea behind our approach starts with the notion of similarity in text. Documents can be similar in various aspects such as semantic, syntactic, or a combination. A fiction book is a relatively long piece of document. Hence we divide each book into smaller sections called "chunks" (see Fig. 2, [1]), and features are extracted per chunk of a book and averaged. SIMFIC features are stored and used in retrieval by comparing using similarity measures for ranking. A web application is built with Angular as front-end, Java Spring boot application as the back-end, MongoDB to capture user clicks anonymously. Our data set consisted of 1200 English books and 470 German books.

3.1 Features

The literary features are selected based on domain knowledge [12] from the context of 19th Century fiction books. We make simplifying assumptions (in line with digital humanities community) to extract these features. For example, the writing style is characterized by the ratio of the usage of punctuation, median sentence length, conjunctions, and others. Table 1 shows the twenty two (f0-f21) features generated in SIMFIC 1.0 along with the new features (f22-f46) introduced in SIMFIC 2.0. The last three feature types in table 1, namely (1) Content-based Genre, (2) Main character and Dialog Interaction, (3) Plot development - constitute the new features proposed in SIMFIC 2.0.

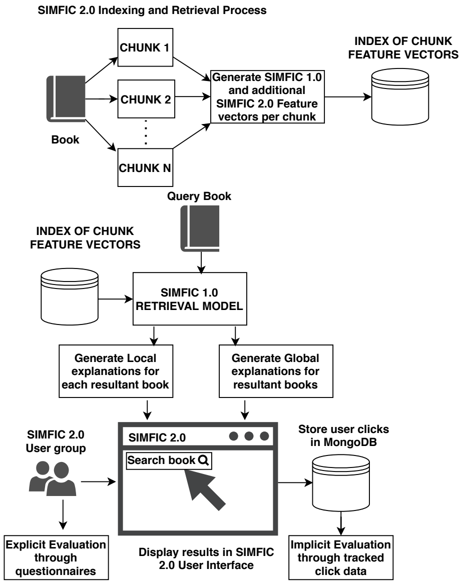

SIMFIC 2.0 Indexing and Retrieval Process
CHUNK 1
CHUNK 2
CHUNK N
Book
Generate SIMFIC 1.0 and additional SIMFIC 2.0 Feature vectors per chunk
INDEX OF CHUNK FEATURE VECTORS
Query Book
INDEX OF CHUNK FEATURE VECTORS
SIMFIC 1.0 RETRIEVAL MODEL
Generate Local explanations for each resultant book
Generate Global explanations for resultant books
SIMFIC 2.0 User group
SIMFIC 2.0
Search book
Store user clicks in MongoDB
Explicit Evaluation through questionnaires
Display results in SIMFIC 2.0 User Interface
Implicit Evaluation through tracked click data

Figure 2: Indexing and Retrieval process in SIMFIC 2.0

Feature Types	Corresponding Features
Writing style	Paragraph Count(f0), Female Pronoun(f1), Male Pronoun(f2),Personal Pronoun(f3), Possesive Pronoun(f4), Preposition(f5), Colon(f9), Semi colon(f10), Hyphen(f11), Interjection(f12), Sentence Length(f14), Punctuation subordinating Conjunction(f13)
Sentence Complexity	Co-ordinating Conjunction(f6), Comma(f7), Period(f8), Punctuation subordinating Conjunction(f13), Sentence Length(f14)
Female oriented	Female Pronoun(f1)
Male oriented	Male Pronoun(f2)
Rural or Urban Setting	Quotes(f15), Number of Characters(f20)
Sentiment	Negative(f16), Positive(f17), Neutral(f18)
Ease of Readability	Flesch Reading Score(f19)
Plot Complexity	Number of Characters(f20)
Lexical Richness	Type Token Ratiof(f21)
Content-based Genre	Theme(f22-f32)
Dialog and Character	Conversation ratio(f33), Number of Speakers(f34), Presence of a main character(f35)
Plot Development	Plot Development(f36-f46)

Table 1: List of Features

[2] https://simfic-falcon.herokuapp.com/

3.1.1 New Features. One of the important facets of a fiction book is its genre which can be detective, romance, horror, etc. We use Doc2Vec to generate a compact vector and take PCA projection to capture "content based genre". Vala et al. [13] find that even though the question, how many characters appear in a novel looks simpler, identifying the exact number of characters is an open question in literacy analysis. Most of the current approaches of identifying them depend on Named-entity recognition (NER) and co reference resolutions [13]. We have used a combination of tags *CorefChainAnnotation* and *NamedEntityTagAnnotation* provided by StanfordCoreNLP library to identify characters. Various works in fiction differ in how the plot advances throughout the book. The readers are more interested in receiving book recommendations that have similar plots at the start and end of the book. We focus on capturing the plot at the start and end of the book through Topic modeling (Latent Dirichlet Allocation - LDA). The words of a topic at the start and end of the book are compared with the ground truth words based on UMass measure [14]. UMass score is a topic coherence technique that is used here to find the similarity between the ground truth words and the topic words. Good topics have a higher cumulative similarity score with the ground truth words.

3.2 Retrieval

Each chunk j for a book i is represented by a feature vector ($book_{i,j}$). The basic idea is to calculate the L^2 norm between query book chunks (C_q) and chunks from books in the corpus (C_i) and convert it to a similarity score. We accumulate these values from chunk to book level, penalize the aggregate by the number of chunks (C_q+C_i) and finally create a ranked list using this measure (refer Eq. 1).

$$sim(book_q, book_i) = \frac{\sum_{j=1}^{C_i} \sum_{k=1}^{C_q} \frac{1}{1+L^2(book_{i,j}, book_{q,k})}}{C_q + C_i} \tag{1}$$

Figure 2 shows the Indexing and Retrieval process in SIMFIC 2.0.

3.3 Explanations

We provide global and local explanations to the search results as follows.

3.3.1 Global Explanations. When a user enters a query book, similar books based on the handcrafted features are retrieved. These top results are the basis for providing global explanations. As the first step, we build a "training data set" with the input variables and the target variable. The input variables are the feature vector values for the books in the search results. We then average these feature values over all the chunks of a book. The target variable is the similarity value (Eq 1) of the query book and a particular resultant book. We argue that a linear model is simple and explainable (like the popular XAI method LIME [2]). Hence we use the "training data set" to learn a linear regression model with 5-fold cross-validation. Features with the highest weights are considered important in retrieving the results. We pose the selected features as global explanations in the interface (see Part A of Fig. 1).

3.3.2 Local Explanations. Local explanations provide a summary for every search result book. It explains why a particular book emerged in search results by measuring the similarity between the query and the feature vector. The feature vector of a book is the average of feature vectors of chunks. Then similarly, a single query vector is obtained. Finally, we compare the vectors to measure similarity scores. To obtain this similarity score, we used the Canberra distance measure [15].

$$d(r_i, q_i) = \frac{|r_i - q_i|}{|r_i| + |q_i|}$$

Here (r_i, q_i) gives the distance between the resultant vector and its query vector for i^{th} feature. The top three features with the lowest scores are displayed as the local explanations for a book (see Part B of Fig. 1). The choice for hyper-parameters in the feature extraction process, explanation generations, and the choice of similarity measures were empirically estimated or decided. This was done on a sample of fifty relatively "known popular books", where we have prior knowledge on similarity based on consensus.

4 RESULTS AND DISCUSSION

In the user study, we evaluate our proposed system against two baseline approaches: a standard TF-IDF based bag-of-words model, a pseudo-random model. We present three systems in Latin block design to avoid potential bias due to usage order. The user study is carried out in an offline fashion under supervision in a controlled lab environment with the TobiiPro T60 eye-tracker. Twenty four users participated in the user study. Twenty three participants belonged to the age group 18-32, and 1 belonged to the age group 32-52. On average, 91.7% of participants use search systems daily. On average, English proficiency is 82.4%, while for German, it is 41.6%, we had only two native German language speakers. There are two search tasks, the first search task is to search with one book from the default set of "known popular books" where we have consensus on similar books in the humanities community. The second search task is to search for any book of the user's choice from our book collection. We ask a set of questions on these two search tasks.

Our focus is to evaluate the systems based on Trustworthiness. A trustworthy system should give fair and reliable results along with its explanations. To make our work comparable we adapt the existing XAI definitions ([9, 16]) in the community, in table 2. As these are subjective, we attempt to make Trustworthiness explicit to the user, by examples such as, "for Titanic movie as a query, a trustworthy search system should return romantic movies, movies by James Cameron (Director), Leonardo DiCaprio, Kate Winslet (Actors) ".

4.1 Filtration of User Responses

Filtration of responses in user studies is an important step to eliminate potential outliers. We filter the responses based on the number of clicks tracked for each participant in MongoDB. We eliminate responses with less than five clicks for all systems combined. Six responses are removed based on user clicks and 2 responses due to a small recording time (<10 minutes).

4.2 Compare Rankings by user clicks

We examine the user clicks by adapting a method [22] that yields the same results as evaluation with traditional relevance judgments under soft assumptions. We group the user clicks of each participant and count the number of clicks tracked for the three systems. We define that the relevance of a system is proportional to the number

XAI Facet	Measured by
Trustworthiness of search results for each search task	Given the definition of trustworthiness, on what scale the results are trustworthy for this task?[9, 16, 17] - Likert Scale (1-5)
Trustworthiness of explanations for the whole system	Given the definition of trustworthiness, on what scale do you trust the key global and local factors (For ex. Writing Style, Theme) responsible for similar books in the list for this system [10] - Likert Scale (1-5)
Trustworthiness of overall system	Given the definition of trustworthiness, On what scale would you trust this system to search books in the future [18] - Likert Scale (1-5)
Understandability of Explanations	1. Was the tooltip explanations for key factors easily understandable?[19] - Likert Scale (1-5) 2. If not, why? [20]- Short answer question
Soundness and Completeness of Explanations for the whole system	1. Do you think that the global factors provided are accurate on why the books are selected? [21] - Likert Scale (1-5) 2. Do you think global factors had all of the information on why it selected the books? [21] - Likert Scale (1-5) 3. Do you think that the local factors provided in each book are accurate on why it selected the book? [21] - Likert Scale (1-5) 4. Do you think local factors had all of the information on why it selected the book? [21] - Likert Scale (1-5)

Table 2: Evaluation of XAI facets in User study

of user clicks.

$$Rel_{X_i} > Rel_{Y_i} \leftrightarrow ClickCount_{X_i} > ClickCount_{Y_i}$$

Here, Rel_{X_i} denotes the Relevance and $ClickCount_{X_i}$ denotes the Number of clicks of i$^{\text{th}}$ participant for System X. A total of 325 clicks (SIMFIC 2.0 = 137, bag-of-words = 81, random = 107) are captured. When compared SIMFIC 2.0 and bag-of-words systems by paired t-test, we observed that it is statistically significant that the relevance of SIMFIC 2.0 is greater than the bag-of-words system (p-value = 0.03722, α = 0.05).

4.3 Evaluation of XAI by questionnaires

4.3.1 Trustworthiness of search results. Under this facet, we asked the participants to perform two search tasks - firstly search on a set of “known popular books” and another search task on any book of user’s choice. We observed that for the first task, the bag-of-words baseline (μ = 3.85, σ = 1.05) is slightly better than the SIMFIC 2.0 system (μ = 3.71, σ = 0.69) and the random baseline (μ = 3.1, σ = 1.18) on average. However, the Wilcoxon-Pratt Signed-Rank statistical Test (WPS) showed that the results are insignificant (p-value = 0.8125, α = 0.05). For the second task, the SIMFIC 2.0 model (μ = 3.85, σ = 0.63) is better than the bag-of-words baseline (μ = 3.5, σ = 1.18) and the random baseline (μ = 2.9, σ = 0.88).

4.3.2 Trustworthiness of explanations. As SIMFIC 2.0 is the only system that contains explanations, we evaluate its Trustworthiness with Likert scores. On average, the participants felt global factors (μ = 3.92, σ = 0.70) to be more trustworthy than local factors (μ = 3.71, σ = 0.69)

4.3.3 Trustworthiness of overall system. Here we observe that the SIMFIC 2.0 system (μ = 3.92, σ = 0.70) is better than the bag-of-words baseline (μ = 3.28, σ = 1.16) and random (μ = 3.21, σ = 1.2). The results are statistically significant according to the WPS test (For SIMFIC 2.0 and bag-of-words system p-value = 0.03906, α = 0.05 and for SIMFIC 2.0 and Random system p-value = 0.04688, α = 0.05). The above results provide empirical evidence indicating that SIMFIC 2.0 provides explanations for search results which are trustworthy. However, no conclusion could be drawn for German books due to less number of participants.

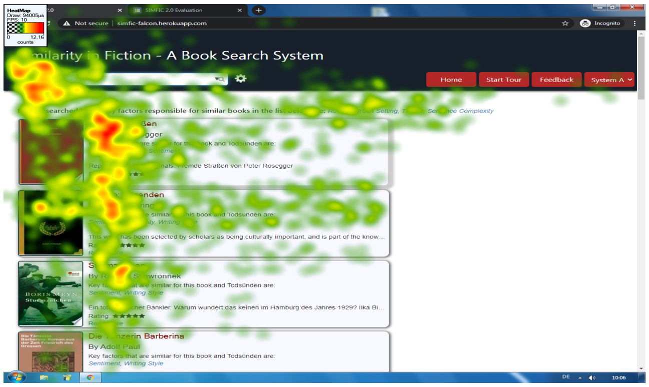

HeatMap
Draw: 94005µs
FPS: 10
0
12,16
counts
SIMFIC 2.0 Evaluation
Not secure | simfic-falcon.herokuapp.com
Incognito
milarity Fiction - A Book Search System
Home
Start Tour
Feedback
System A
searched factors responsible for similar books in the list
Setting,
Sentence Complexity
ßen
gger
for this book and Todsünden are:
Sentiment
Fremde Straßen von Peter Rosegger
enden
similar his book and Todsünden are:
Writing
This has been selected by scholars as being culturally important, and is part of the know...
By Skowronnek
at are similar for this book and Todsünden are:
Sentiment, Writing Style
cher Bankier. Warum wundert das keinen im Hamburg des Jahres 1929? Ilka Bi...
Rating: ★★★★★
BORIS MEYN
Sturmzeichen
Die zerin Barberina
By Adolf Paul
Key factors that are similar for this book and Todsünden are:
Sentiment, Writing Style
Die Tänzerin Barberina: Roman aus der Zeit Friedrich des Grossen
DE
10:06

Figure 3: Aggregated heat-map from Eye Tracker

4.4 Evaluation through eye tracker

We wanted to explore: Is there a scanning pattern followed by users and to what extent does a user pay attention to global and local explanations? The heat-map (refer Fig. 3) highlights the regions with varying fixations. A fixation is a spot that has the user’s focus. Here, the red color indicates a high fixation rate, while the green color indicates low fixation rates. We can observe the typical F-shaped pattern [23] describing the user’s gaze positions and scanning pattern in the interface. We notice that the regions with textual global and local explanations contain a fair amount of user’s fixation. These indicate that explanations play an important role in user perception.

5 CONCLUSION

We present an explainable search system for fiction books in a query-by-example setting. We make use of a compact and interpretable feature space. The system offers a global explanation by fitting a linear model for a set of top retrieved items along with the local explanation for each book based on similarity measures and features. We run XAI focused evaluation with the goal of estimating the Trustworthiness of the generated explanations by an in-lab user study. The ranking performance is compared via user clicks. Eye-trackers were used to explore areas of interest in the interface, from the user perspective. Empirical evidence with statistical significance indicates that the explanations are trustworthy. However, the results are based on subjective opinions. The system is currently being experimented with domain-specific data sets in different languages.

REFERENCES

[1] S. Polley, S. Ghosh, M. Thiel, M. Kotzyba, and A. Nürnberger, "Simfic: An explainable book search companion," in *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*. IEEE, 2020, pp. 1–6.

[2] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1135–1144. [Online]. Available: https://doi.org/10.1145/2939672.2939778

[3] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, "Layer-wise relevance propagation: an overview," *Explainable AI: interpreting, explaining and visualizing deep learning*, pp. 193–209, 2019.

[4] J. Singh and A. Anand, "Exs: Explainable search using local model agnostic interpretability," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 2019, pp. 770–773.

[5] Z. T. Fernando, J. Singh, and A. Anand, "A study on the interpretability of neural retrieval models using deepshap," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR'19. Association for Computing Machinery, 2019, p. 1005–1008.

[6] H. Chiang, Y. Ge, and C. Wu, "Classification of book genres by cover and title," 2015.

[7] H. Alharthi and D. Inkpen, "Study of Linguistic Features Incorporated in a Literary Book Recommender System," 2019. [Online]. Available: https://doi.org/10.1145/3297280.3297382

[8] M. Melucci, "Can structural equation models interpret search systems?" in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR'19. Association for Computing Machinery, 2019.

[9] S. Mohseni, N. Zarei, and E. D. Ragan, "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems," *ACM Trans. Interact. Intell. Syst. 1, 1, Article*, vol. 1, p. 46, 2020.

[10] S. Berkovsky, R. Taib, and D. Conway, "How to Recommend? User Trust Factors in Movie Recommender Systems." [Online]. Available: http://dx.doi.org/10.1145/3025171.3025209

[11] T. Gossen and J. Höbel, "A Comparative Study about Children's and Adults' Perception of Targeted Web Search Engines." [Online]. Available: http://dx.doi.org/10.1145/2556288.2557031

[12] M. L. Jockers, *Macroanalysis: Digital methods and literary history*. University of Illinois Press, 2013.

[13] H. Vala, D. Jurgens, A. Piper, and D. Ruths, "Mr. bennet, his coachman, and the archbishop walk into a bar but only one of them gets recognized: On the difficulty of detecting characters in literary texts," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2015, pp. 769–774.

[14] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011.

[15] G. Lance and W. T. Williams, "Computer programs for hierarchical polythetic classification ("similarity analyses")," *Comput. J*, vol. 9, pp. 60–64, 1966.

[16] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable AI: challenges and prospects," *CoRR*, 2018.

[17] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions," *CoRR*, 2017.

[18] J. L. Herlocker, J. A. Konstan, and J. Riedl, "Explaining collaborative filtering recommendations," in *Proceedings of the ACM Conference on Computer Supported Cooperative Work*. Association for Computing Machinery (ACM), 2000, pp. 241–250.

[19] R. R. Hoffman, M. Johnson, J. M. Bradshaw, and A. Underbrink, "Trust in automation," *IEEE Intelligent Systems*, vol. 28, no. 1, pp. 84–88, 2013.

[20] J. Vermeulen, G. Vanderhulst, K. Luyten, and K. Coninx, "PervasiveCrystal: Asking and answering why and why not questions about pervasive computing applications," in *Proceedings - 2010 6th International Conference on Intelligent Environments, IE 2010*. IEEE, jul 2010, pp. 271–276.

[21] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, and W.-K. Wong, "Too Much, Too Little, or Just Right? Ways Explanations Impact End Users' Mental Models," pp. 3–10, 2013.

[22] T. Joachims, "Evaluating retrieval performance using clickthrough data," in *Text Mining*, J. Franke, G. Nakhaeizadeh, and I. Renz, Eds. Physica Verlag, 2003.

[23] T. Gossen, J. Höbel, and A. Nürnberger, "Usability and perception of young users and adults on targeted web search engines," in *Proceedings of the 5th Information Interaction in Context Symposium*. Association for Computing Machinery, 2014.