A Study of Explainability Features to Scrutinize Faceted Filtering Results

Jiaming Qu, Jaime Arguello, Yue Wang
School of Information and Library Science, University of North Carolina at Chapel Hill, USA
jiaming@live.unc.edu,{jarguello,wangyue}@unc.edu

ABSTRACT

Faceted search systems enable users to filter results by selecting values along different dimensions or *facets*. Traditionally, facets have corresponded to properties of information items that are part of the document metadata. Recently, faceted search systems have begun to use machine learning to automatically associate documents with facet-values that are more subjective and abstract. Examples include search systems that support topic-based filtering of research articles, concept-based filtering of medical documents, and tag-based filtering of images. While machine learning can be used to infer facet-values when the collection is too large for manual annotation, machine-learned classifiers make mistakes. In such cases, it is desirable to have a scrutable system that explains *why* a filtered result is relevant to a facet-value. Such explanations are missing from current systems. In this paper, we investigate how explainability features can help users interpret results filtered using machine-learned facets. We consider two explainability features: (1) showing prediction confidence values and (2) highlighting *rationale* sentences that played an influential role in predicting a facet-value. We report on a crowdsourced study involving 200 participants. Participants were asked to scrutinize movie plot summaries predicted to satisfy multiple genres and indicate their agreement or disagreement with the system. Participants were exposed to four interface conditions. We found that both explainability features had a positive impact on participants' perceptions and performance. While both features helped, the sentence-highlighting feature played a more instrumental role in enabling participants to reject false positive cases. We discuss implications for designing tools to help users scrutinize automatically assigned facet-values.

CCS CONCEPTS

• **Computing methodologies → Supervised learning**; • **Information systems → Search interfaces**.

KEYWORDS

Faceted Filtering; Explainable Machine Learning; User Study

ACM Reference Format:
Jiaming Qu, Jaime Arguello, Yue Wang. 2021. A Study of Explainability Features to Scrutinize Faceted Filtering Results. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21), November 1–5, 2021, Virtual Event, QLD, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3459637.3482409

1 INTRODUCTION

Faceted search systems enable users to explore large collections of documents in meaningful ways. It is widely adopted in e-commerce search engines and digital libraries where documents are associated with rich metadata. In domains where document collections are large and manual facet annotation is prohibitive, machine learning algorithms are increasingly used to assign facet-values to documents [18, 28, 36, 38, 41]. Figure 1 shows a few such examples. Microsoft Academic allows users to filter articles using machine-learned topics [38]; SciSight allows users to filter COVID-19 papers using machine-extracted concepts [18]; Google Images allows users to sort images by machine-predicted tags.

Machine learning algorithms are able to scale facet annotation to large and growing collections [28, 41] and even support user-defined ad-hoc facets [24]. However, these algorithms are not perfect and their predicted facet-values can be erroneous or puzzling to end users. To illustrate, in Figure 1(a), the topic "*Interface (Java)*" was assigned to articles returned by the query "document classification interface". A user may naturally wonder: Is it because Java was used to create document classification interfaces, or because the system assigned the wrong label and the correct label is "*User Interface*"? Similar doubts may arise for the topic "*Word (computer architecture)*". Unfortunately, Microsoft Academic does not currently explain *why* topics are assigned to specific articles.

Conventional facet-values (e.g., author, publication year, price) are usually self-explanatory. However, as illustrated in Figure 1, machine-learned facet-values (e.g., research topics, medical concepts, image themes) can be subjective and abstract. In such cases, faceted navigation interfaces should provide visualizations and tools to help users understand *why* specific facet-values have been assigned to a document. Such explanations may help users: (1) locate supporting evidence for why a facet-value has been assigned to a document, (2) reject a filtered search result if the supporting evidence is weak, (3) understand the meaning of a facet-value from the system's perspective, and (4) gain insights into how the system makes inferences and why it makes occasional mistakes. System-provided explanations can help *demystify* machine-learned facet-values and help users scrutinize decisions made by the system.

In this paper, we explore two different strategies for explaining facet-values automatically assigned to documents:

Confidence Values: One straightforward strategy for explaining predicted categories (including facet-values) is to display confidence values. Well-calibrated confidence values communicate a system's (un)certainty in its decisions [13, 32]. Recent work has shown that displaying confidence values can help users calibrate

CIKM '21, November 1–5, 2021, Virtual Event, QLD, Australia

ACM ISBN 978-1-4503-8446-9/21/11…$15.00
https://doi.org/10.1145/3459637.3482409

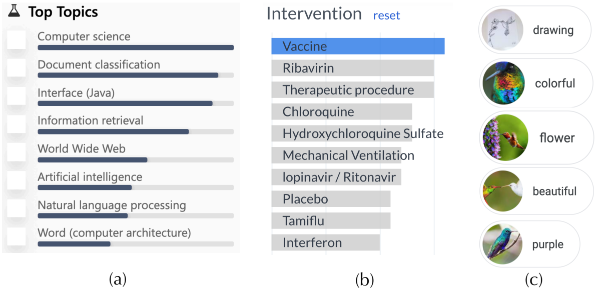


Figure 1: Examples of machine-learned facets in real systems. (a) Research topics for the query "document classification interface" in Microsoft Academic. (b) Intervention concepts in SciSight, a faceted search system for COVID-19 papers. (c) Image tags for the query "hummingbird" in Google Images. All accessed in May 2021.

their trust in the system [43]. Additionally, a low-confidence prediction may draw a user's attention as it may actually be incorrect.

Sentence Highlighting: A second strategy for explaining automatically assigned categories is to provide a *rationale* for why a category was assigned to a document. One approach is to visually highlight which parts of a document *influenced* the system to predict a specific category [25, 35]. In fact, search engines have long been "rationalizing" search results by highlighting query terms (or semantically related terms) in the summary snippet of each top result. Here we extend this idea to explain automatically inferred facet-values. We developed an interactive tool that allows users to select a predicted category and see which parts of a document (i.e., sentences) played an influential role in predicting the category.

We developed an experimental interface where classifier-predicted movie genres were used to filter movie plots and the above strategies were used to explain system predictions. We studied how these *explainability features* may assist users in judging the relevance of movie plots in a controlled user study on Mechanical Turk. The study investigated three research questions (RQs) related to users' perceptions, performance, and behaviors:

RQ1: How do both explainability features influence users' perceptions of satisfaction, confidence, difficulty, workload, and system understandability and usability?

RQ2: How do both explainability features influence users' objective performance in judging the relevance of a faceted filtering result when compared against ground truth labels?

RQ3: How do both explainability features influence users' behaviors in terms of the time required to make decisions and the level of engagement with different components of the interface?

By exploring these questions, we gain insights about designing systems that enable users to scrutinize predicted facet-values.

2 RELATED WORK

Faceted Filtering: Faceted search systems enable users to filter search results and navigate a collection in a structured manner [11, 15]. The success of faceted search systems hinges on the availability of high-quality metadata to generate facets. When these metadata naturally come with documents (e.g. product attributes in e-commerce), faceted filtering and navigation can be readily implemented. In many scenarios, however, facet-values need to be manually assigned to documents. For instance, MeSH terms are manually assigned to MEDLINE articles [33]. Prior research has aimed to automate this process by using unsupervised algorithms to cluster search results into topics [6, 22] or by using supervised learning algorithms to associate documents with facet-values from pre-defined schema [18, 28, 36, 38, 41].

While machine learning is a powerful tool to advance faceted search, machine-learned classifiers unavoidably make mistakes. Therefore, it is desirable for faceted search systems to be transparent and explain *why* specific facet-values have been assigned to a document. Such explanations have only been investigated in a few research prototypes. The PubTator and PubTator Central systems highlight automatically extracted concepts for different biomedical facets (e.g., diseases, genes, etc.) [39, 40]. The RobotReviewer system highlight sentences from biomedical articles that may indicate a biased clinical trial [29]. In our study, we focus on how similar types of explanations may affect users' subjective perceptions of the system and objective task performance, an emerging and underexplored question in this research area.

Interpretable Predictive Algorithms: Interpretable machine learning (ML) has recently attracted considerable attention [12, 31]. In our study scenario, searchers want to know why a document is relevant to a machine-predicted facet-value. Within the interpretable ML community, this is called *local explanation* for an individual instance. Two major approaches exist for local explanations. Feature-based approaches identify and highlight parts of the input (e.g., words, sentences, paragraphs, metapixels) as *rationales* for why a model predicted a certain label [9, 27, 35]. Example-based methods use similar examples with gold-standard labels to explain the predicted label for an instance [21]. For text classification tasks, feature-based methods have been more widely adopted [7, 23, 25, 43]. Previous research has also investigated how confidence values can help users calibrate their trust towards specific predictions [26, 43]. Our work considers different combinations of these strategies (i.e., displaying confidence values and highlighting sentences) to explain predictions made by text classification models.

Improving system explainability has also gained interest in the information retrieval community. Current search systems highlight query terms and semantically related terms in summary snippets as a form of explanation [30]. Different visualization techniques have been proposed to explain why a result is relevant to each query term [16, 17, 34]. Our work studies a scenario where the filtering criteria are abstract facet-values instead of query terms. In recommender systems research, prior work has developed approaches that leverage a user's history, preference profile, and social network to explain item recommendations [42]. Balog et al. [1] emphasized the distinction between justification (i.e., providing a plausible explanation) and transparency/scrutability (i.e., providing an honest account of how the system works). In this regard, our explainability features promotes system transparency and scrutability.

User-Centered Explainability Studies: The goal of explaining predictions is to support users in accomplishing their tasks. These tasks may include decision-making, sense-making, debugging, and auditing tasks [10, 20, 37]. In our work, system explanations are intended to collaborate with users to more accurately judge whether a result actually satisfies a set of criteria.

It is believed that when assisted by an intelligent system, human users should be able to achieve better decision outcomes than the system itself or an unaided user [19]. However, in practice, such collaborations can be a delicate balancing act [3, 4]. In particular, previous user studies have found that feature-based explanations of text classification decisions failed to help users outperform the system alone [7, 23]. Zhang et al. [43] argued that it is crucial to display, not only a system's rationale for a prediction, but also its confidence value. Confidence values can help users know when to trust and when to question a system's predictions. We share similar considerations in adopting our two explainability features. In contrast to prior work [7, 23], our results found that our explainability features helped participants make more accurate judgements and outperform both unaided users and the system itself.

3 METHODS

3.1 Study Overview

To investigate our three research questions, we conducted a crowdsourced study ($N = 200$) on Amazon Mechanical Turk (MTurk). Our main objective was to investigate how our two explainability features influence the way people *scrutinize* and *evaluate* faceted filtering results where facet-values are machine-learned. As a case study, we used the movie domain. Participants in the study were exposed to movie plot summaries automatically predicted to belong to two or three pre-selected movie genres (referred to as the *selected* genres).[1] For each plot summary, participants were asked to either *agree* or *disagree* with the system. Participants were instructed to *agree* with the system if they believed that the movie *correctly* belongs to all selected genres and *disagree* with the system if they believed that the movie does not belong to all selected genres (i.e., at least one of them is wrong). Movies were pre-classified into selected genres using only the movie's plot summary. Similarly, participants made agree/disagree decisions based only on the plot summary. The study investigated two explainability features: prediction confidence values and sentence highlighting. These features were investigated in isolation and in combination. Thus, participants were exposed to four interface conditions (Section 3.2).[2]

3.2 Interface Conditions

Participants in our study were exposed to four interface conditions. Figure 2 shows a screenshot of condition CONF+SENT, which included both explainability features: confidence values (CONF) and sentence highlighting (SENT).

Condition BASELINE: In the BASELINE condition, participants had to make agree/degree decisions based only on the plot summary. The BASELINE interface did not show prediction confidence values and did not include the sentence-highlighting feature. The pre-selected genres were highlighted on the left menu and listed at the top of the interface. The plot summary was shown in the middle region. Participants were prompted with the question: "Does the movie belong to ALL of the selected genres?" Participants indicated their agreement/disagreement with the system by selecting either "yes" or "no" and clicking the submit button.

Condition CONF: In the CONF condition, the interface looked the same as in the BASELINE condition. However, the interface also included prediction confidence values as illustrated in Figure 2. Each genre had its own color. Prediction confidence values were shown as colored bars ranging from 0% to 100%. The exact confidence values were embedded in each colored bar. Before each interface condition, participants watched a video introducing the features of the next interface. In the video for condition CONF, participants were told that confidence values ranged from 0% to 100%. Participants were instructed that values close to 100% indicate that the system is highly confident that the movie belongs to that genre, values close to 0% indicate that the system is highly confident that the movie *does not* belong to that genre, and values close to 50% indicate that the system is unsure. As explained in Section 3.3, confidence values corresponded to well-calibrated posterior probabilities predicted by logistic regression classifiers.

Condition SENT: In the SENT condition, the interface looked the same as the in BASELINE condition. However, the interface also included the sentence-highlighting feature as illustrated in Figure 2. The sentence-highlighting feature allowed participants to select a specific genre in order to see the sentences from the plot summary considered by the system to be "influential" in deciding that the movie belongs to that genre. Again, each genre was associated with its own color. Upon clicking on a genre, the interface highlighted sentences with different degrees of color intensity. As shown in Figure 2, a reference key was provided to remind participants that greater intensity means "more influential". Clicking the "hide" button removed all sentence highlighting from the plot summary. In the video for condition SENT, participants were told that this feature allowed them to "see which sentences the system used to base its predictions". By selecting a genre, "you can see which sentences influenced the system to decide that the movie belongs to that genre". Additionally, participants were instructed that "brightly colored sentences contributed the *most* evidence towards the genre, lightly colored sentences contributed *some* evidence, and sentences with no highlighting contributed *no* evidence".

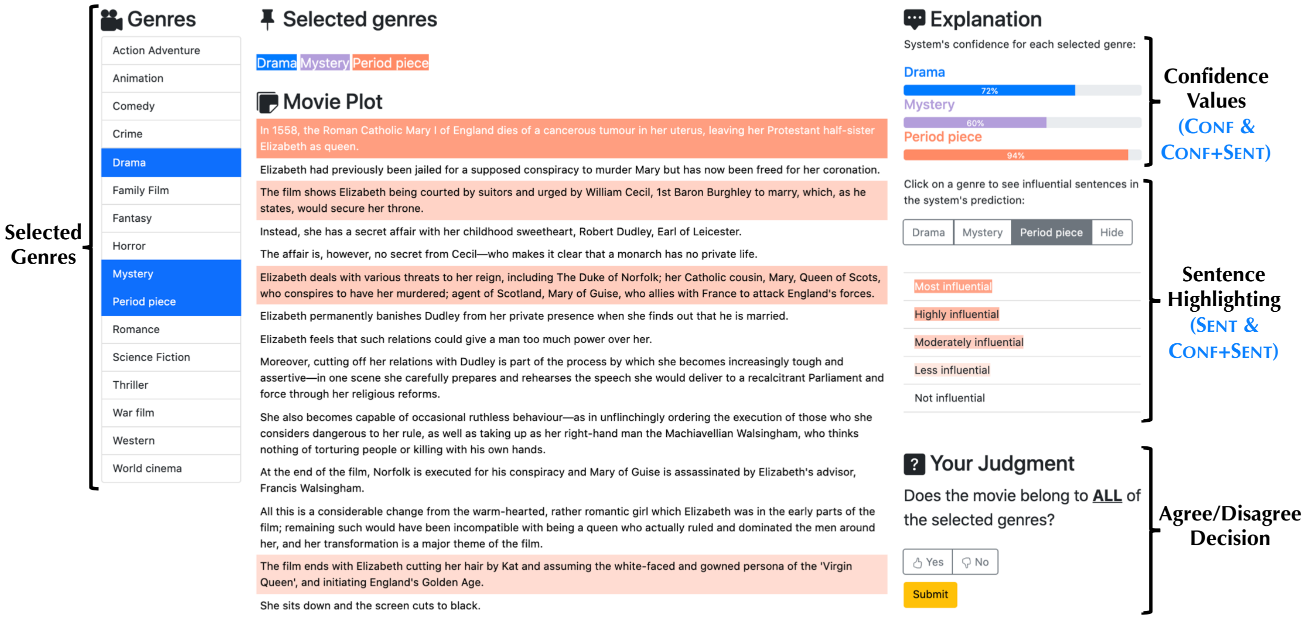
From a system perspective (not explained to participants), the sentence-highlighting feature was implemented as follows. First, as explained in Section 3.3, we trained independent logistic regression classifiers (one per genre) using whole plot summaries. Then, we used these document-level models to make sentence-level predictions. Given a specific genre and plot summary, the color intensity of each sentence s was determined in two steps. First, sentence-level prediction confidence values were normalized according to:

$$\mathcal{P}_{\text{norm}}(s|g) = \frac{\mathbf{max}\,(0, \mathcal{P}_{\text{raw}}(s|g) - 0.5)}{0.5}\,, \tag{1}$$

where $\mathcal{P}_{\text{raw}}(s|g)$ denotes the probability that sentence s belongs to genre g according to the document-level classifier for g. By default, a logistic regression model outputs a positive prediction if its confidence value is greater than 0.5. Equation (1) was designed to output normalized confidence values in the range [0,1]. Additionally, it was designed to output a value of 0 if $\mathcal{P}_{\text{raw}}(s|g) < 0.5$, meaning that the document-level classifier is more confident that s *does not* belong to g than vice-versa. Normalized confidence values were binned into five discrete levels to be consistent with the key illustrated in Figure 2 (0-.20 = "not influential", 0.21-0.40 = "less influential", etc.).

[1] We did not consider the scenario where only one genre was pre-selected, as it effectively overlaps with interpreting binary classification decisions, which has been studied in previous work [7, 23, 26].

[2] All our study materials are available online. This online appendix provides access to: (1) all videos that introduced participants to the study and each interface condition, (2) interactive examples of our interface conditions (Section 3.2), and (3) the full text of our post-task questionnaire (Section 3.6).



In 1558, the Roman Catholic Mary I of England dies of a cancerous tumour in her uterus, leaving her Protestant half-sister Elizabeth as queen.

Elizabeth had previously been jailed for a supposed conspiracy to murder Mary but has now been freed for her coronation.

The film shows Elizabeth being courted by suitors and urged by William Cecil, 1st Baron Burghley to marry, which, as he states, would secure her throne.

Instead, she has a secret affair with her childhood sweetheart, Robert Dudley, Earl of Leicester.

The affair is, however, no secret from Cecil—who makes it clear that a monarch has no private life.

Elizabeth deals with various threats to her reign, including The Duke of Norfolk; her Catholic cousin, Mary, Queen of Scots, who conspires to have her murdered; agent of Scotland, Mary of Guise, who allies with France to attack England's forces.

Elizabeth permanently banishes Dudley from her private presence when she finds out that he is married.

Elizabeth feels that such relations could give a man too much power over her.

Moreover, cutting off her relations with Dudley is part of the process by which she becomes increasingly tough and assertive—in one scene she carefully prepares and rehearses the speech she would deliver to a recalcitrant Parliament and force through her religious reforms.

She also becomes capable of occasional ruthless behaviour—as in unflinchingly ordering the execution of those who she considers dangerous to her rule, as well as taking up as her right-hand man the Machiavellian Walsingham, who thinks nothing of torturing people or killing with his own hands.

At the end of the film, Norfolk is executed for his conspiracy and Mary of Guise is assassinated by Elizabeth's advisor, Francis Walsingham.

All this is a considerable change from the warm-hearted, rather romantic girl which Elizabeth was in the early parts of the film; remaining such would have been incompatible with being a queen who actually ruled and dominated the men around her, and her transformation is a major theme of the film.

The film ends with Elizabeth cutting her hair by Kat and assuming the white-faced and gowned persona of the 'Virgin Queen', and initiating England's Golden Age.

She sits down and the screen cuts to black.

System's confidence for each selected genre:

Click on a genre to see influential sentences in the system's prediction:

Does the movie belong to **ALL** of the selected genres?

Figure 2: Screenshot of interface in condition Conf+Sent. Sentence highlighting shown for Period Piece genre. Confidence values were only shown for conditions Conf and Conf+Sent. The sentence-highlighting feature was only available for conditions Sent and Conf+Sent.

Condition Conf+Sent: The Conf+Sent condition included both explainability features. As illustrated in Figure 2, participants could see prediction confidence values and also had access to the sentence-highlighting feature. In the video for condition Conf+Sent, participants were introduced to the confidence value feature using the same wording as in condition Conf and the sentence-highlighting feature using the same wording as in condition Sent.

3.3 Faceted Filtering Data Preparation

Dataset Curation: The data used in our study originated from Bamman et al. [2]. The original dataset contains 42,303 movie plot summaries gathered from Wikipedia. Each movie plot summary is associated with one or more genres from a total set of 364 genres. The original dataset was developed (in part) to evaluate machine learning approaches for multiclass classification—assigning instances to one or more categories.

The original dataset was curated as follows. First, the original dataset included many rare genres without enough positive examples to train a reasonably good classifier. Therefore, we first identified the 30 most frequent genres in the dataset. These 30 genres included a mix of *topical* and *non-topical* genres. Therefore, in our second step, we manually selected 16 topical genres to consider in our study (e.g., Action/Adventure, Romance, Horror, etc.). We omitted 14 non-topical genres that we expected to be difficult for a machine-learned classifier or human to "detect" based solely on a movie's plot summary. We excluded non-topical genres such as Short Film, Indie Film, and Black & White. Finally, we noticed that plot summaries varied widely in length. Therefore, in our final step, we filtered plot summaries with less than 5 or more than 40 sentences. Our final curated dataset included 24,346 plot summaries associated with one or more genres from the set of 16 total genres.

Classification: To generate genre predictions and confidence values for conditions Conf and Conf+Sent, we used 16 logistic regression classifiers (one per genre). All classifiers used the *same* bag-of-words representation, which included all non-stopword unigrams with a minimum frequency of 1% of the total term frequency. Genre predictions and confidence values were generated using 10-fold cross-validation. Movies were assigned to all genres with a prediction confidence value greater than 0.5. We used the stopword list and sentence tokenizer in NLTK, and scikit-learn for training classifiers. Our classifiers achieved an F1 score between 0.51 and 0.75 for 12 (out of 16) genres. The four most difficult genres were World Cinema, Fantasy, Mystery, and Period Piece ($0.25 \leqslant \text{F1} \leqslant 0.44$).

3.4 Experimental Design

We designed the study to meet the following criteria: (1) expose participants to all interface conditions (a within-subjects design), (2) expose participants to a wide range of movie plot summaries and genres, and (3) vary the order in which participants experienced the interface conditions. The design is illustrated in Figure 3.

We designed our study as follows. First, we identified all movies automatically assigned to either two or three genres (11,228 out of 24,346 movies). From this set of movies, we formed 50 *batches* of 40 movies each (i.e., 2,000 unique movies in total). Within each batch, we formed 4 *sequences* of 10 movies each. Across all batches, each sequence of 10 movies had the following two constraints. First, each sequence of 10 movies included 5 *true positive* and 5 *false positive* cases. A true positive case means that all predicted genres are correct and a false positive case means that *at least one* predicted genre is incorrect based on the ground truth genres. Second, within each sequence of 10 movies, every genre was included (i.e., predicted) in at least one movie. Participants were instructed that the system could make mistakes (i.e., they should not always agree with the system). Participants were *not* told that each sequence of 10 movies had 50% true positive and 50% false positive cases.

To control for learning and fatigue effects, we varied the order in which participants were exposed to interface conditions (Figure 3).

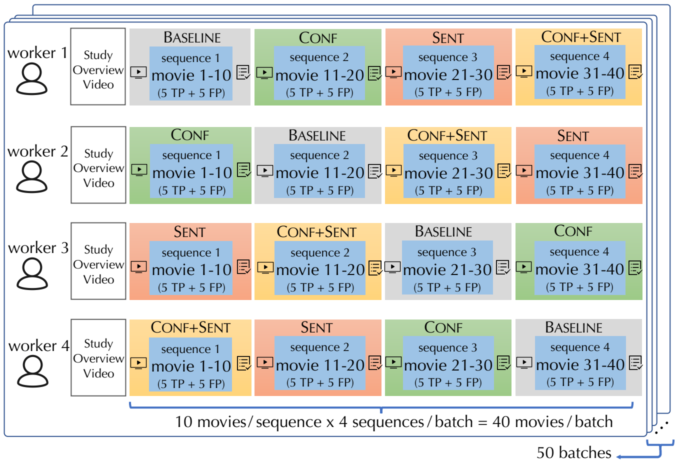


Figure 3: Experimental Design. Within each of our 50 batches, participants judged 4 sequences of movies (10 movies per sequence). The order of movies remained consistent, but we rotated the order of the interface condition according to a Latin square design.

Using a Latin square, four treatment conditions yields four orders. Each batch of movies (out of 50) was completed by *four* redundant MTurk workers. Within each batch, the order of movies remained consistent—all four redundant workers experienced the same order of sequences and the same order of movies within each sequence. However, each of the four redundant workers experienced the interface conditions in a different order. Ultimately, we gathered data from 200 *distinct* participants (i.e., 50 batches × 4 redundant MTurk workers per batch). MTurk workers were not allowed to complete the experiment more than once (even for a different batch).

3.5 Study Protocol and MTurk Quality Control

Study Protocol: As shown in Figure 3, each participant was exposed to a *batch* of 40 movies. Each batch consisted of 4 *sequences* of movies (10 movies per sequence). For each sequence of 10 movies, participants were asked to make agree/disagree decisions using a specific interface (i.e., a within-subjects design).

The study protocol proceeded as follows. After providing informed consent, participants watched a video describing the study. The video situated participants in the following scenario:

Imagine that you want to watch a movie associated with a specific combination of genres (e.g., comedy and romance). Using a recommendation system, you specify these genres and the system outputs a relevant movie. The recommendation system uses machine learning or artificial intelligence (AI) to make predictions. Additionally, the system makes predictions by considering only the movie's plot summary. The system (like many AI systems) sometimes makes mistakes. During this experiment, you will be exposed to four different interfaces that allow users to scrutinize and evaluate predictions made by an AI recommendation system. For each interface, you will be exposed to a sequence of 10 movies. Each movie will be associated with two or three pre-selected genres. You will be asked to either agree or disagree with the system. You should agree with the system if you believe that the movie belongs to all pre-selected genres based on the plot summary. Conversely, you should disagree with the system if you believe that the movie does not belong to all pre-selected genres. Additional videos will explain the features of each interface.

As shown in Figure 3, each interface condition involved the same sequence of three steps. First, participants watched a video introducing the features associated with the next interface condition. Second, participants were exposed to a sequence of 10 movies in which they were asked to agree/disagree with the system. After each sequence, participants completed a post-task questionnaire about their perceptions of the interface and their experience (Section 3.6). After completing all four sequences of 10 movies, each participant was paid US$ 8.00 for the Human Intelligence Task (HIT). Each participant was able to complete only one of our HITs.

MTurk Quality Control: To help recruit high-quality MTurk workers with English proficiency, we restricted our HITs to workers with at least 500 completed HITs, a 95% acceptance rate or higher, and workers within the U.S. Data collection proceeded in two rounds. During the first round, we published 200 HITs. After analyzing the data, we noticed that some participants completed the HIT too quickly. These participants were compensated. However, during the second round, we decided to republish HITs from 20 (out of 200) first-round workers who had a median judgement time of no greater than 12 seconds per movie. This resulted in data from 200 participants who we believe provided genuine responses.

3.6 Post-task Questionnaire (RQ1)

In RQ1, we investigate the effects of the interface condition on participants' perceptions of the interface and their experience. To address RQ1, participants completed a three-part questionnaire after judging each sequence of 10 movies using a specific interface.

The first part of the post-task questionnaire included 9 items that asked about participants' perceptions of: (1) satisfaction with their performance (1 item), (2) difficulty (1 item), (3) confidence when agreeing/disagreeing with the system (3 items), and (4) understandability of the system's predictions (4 items). Participants responded to agreement statements using a 7-point scale ranging from "strongly disagree" to "strongly agree". The three confidence items had low internal consistency (Cronbach's α=0.59). Therefore, we analyzed responses to these items individually. Conversely, the four understandability items had high internal consistency (Cronbach's α=0.93). Therefore, we averaged responses to these four items to form one understandability measure.

The second part of the post-task questionnaire asked about system usability. To this end, we used the System Usability Scale (SUS) [5]. The SUS includes 10 items about system usability (i.e., ease of use). Again, participants responded to agreement statements using a 7-point scale ranging from "strongly disagree" to "strongly agree". The SUS includes five positive items (i.e., higher values indicate higher usability) and five negative items (i.e., higher values indicate lower usability). The five negative items were reverse-coded. Responses from participants had high internal consistency (Cronbach's α=0.82). Therefore, we averaged responses to all 10 items to form one system usability measure.

The third part of the post-task questionnaire asked about workload. To this end, we used the NASA-TLX questionnaire [14], which includes six items about workload: (1) mental demand, (2) physical demand, (3) temporal demand, (4) task outcome, (5) effort, and (6) frustration. With one exception, participants responded to statements using a 7-point scale ranging from "very low" to "very high". For task outcome, the 7-point scale ranged from "perfect" to "failure". Responses from participants had high internal constancy (Cronbach's α=0.82). Therefore, we averaged responses to these 6 items to form one workload measure.

3.7 Performance Metrics (RQ2)

In RQ2, we investigate the effects of the interface condition on the extent to which participants made correct agree/disagree decisions—agreed with the system for true positive cases and disagreed with the system for false positive cases. We measured participants' performance from different perspectives.

Accuracy: This measure considers the percentage of times participants made correct agree/disagree decisions. As previously mentioned, each sequence 10 movies included 5 true positive and 5 false positive cases (Figure 3). Accuracy values are in the range [0,1]. However, because we included an equal number of true positive and false positive cases per sequence, we expected accuracy values to be greater than 0.5, the *expected* accuracy for participants who either always agreed, always disagreed, or made random decisions with equal probability.

Normalized Accuracy: Unnormalized accuracy values do not account for the fact that the task may have been inherently difficult (i.e., perfect accuracy is unlikely). Additionally, unnormalized accuracy values ignore the fact that some movies in our dataset may have been more difficult than others. Based on our experimental design (Figure 3), each sequence of 10 movies was judged by four redundant participants, each in a different interface condition. To compute normalized accuracy, we used min-max scaling. That is, each participant's accuracy for a given sequence was normalized using the minimum and maximum accuracy values from all four redundant workers who made judgements for the same sequence.

Precision: This measure considers precision with respect to *true positive* movies. In other words, this measure considers the percentage of "agree" decisions that involved a true positive (vs. false positive) movie. Precision measures the extent to which participants *rejected* false positive movies when they agreed with the system.

Recall: This measure considers recall with respect to *true positive* movies. In other words, this measure considers the percentage of true positive movies for which participants agreed with the system. Recall measures the extent to which participants agreed for *all* true positive movies.

Yes Rate: This measure considers the percentage of times participants agreed with the system. Yes-rate does not measure performance per se (i.e., it does not consider the ground truth genre labels). However, it provides insights about participants' tendencies to agree with the system across interface conditions.

3.8 Behaviors (RQ3)

In RQ3, we investigate the effects of the interface condition on participants' behaviors using the interface. We considered three behavioral measures.

Completion Time (seconds): This measure considers the amount of time participants took to make agree/disagree decisions.

Mouse on the Plot Summary (percentage): Prior work has shown a strong correlation between mouse and visual gaze position [8]. This measures considers the percentage of time participants positioned their mouse pointer over a movie's plot summary.

Sentence Highlighting (percentage) In conditions Sent and Conf+Sent, participants could select genres to see which sentences the system considered to be more or less "influential". This measure considers the percentage of time participants had the sentence-highlighting feature turned "on" (for any genre).

4 RESULTS

In this section, we present results for RQ1-RQ3. In all cases, the analysis was done at the sequence level (10 movies using a specific interface). For example, when comparing accuracy values across interface conditions (RQ2), we first computed accuracy values for each sequence of movies, and then we compared *average* accuracy values across interface conditions. In all cases, to test for statistically significant differences across interface conditions, we used *repeated measures* ANOVAs and Bonferroni-corrected *paired t*-tests to compare between all interface condition pairs. In Section 5, we present a few follow-up experiments to gain additional insights about the influences of the interface condition in specific scenarios.

4.1 RQ1: Effects on Perceptions

In RQ1, we investigate the effects of the interface condition on participants' perceptions of satisfaction, difficulty, confidence when agreeing or disagreeing with the system, confidence increase, understandability, system usability, and workload.

Figure 4(a) shows these measures and indicates all significant pairwise comparisons. The interface condition had a significant main effect on all measures: (1) satisfaction ($F(3,597) = 11.21$, $p < .001$); (2) difficulty ($F(3,597) = 13.19$, $p < .001$); (3) confidence when agreeing with the system ($F(3,597) = 10.72$, $p < .001$); (4) confidence when disagreeing with the system ($F(3,597) = 7.69$, $p < .001$); (5) confidence increase ($F(3,597) = 67.76$, $p < .001$); (6) understandability ($F(3,597) = 130.21$, $p < .001$); (7) usability ($F(3,597) = 25.80$, $p < .001$); and (8) workload ($F(3,597) = 24.37$, $p < .001$).

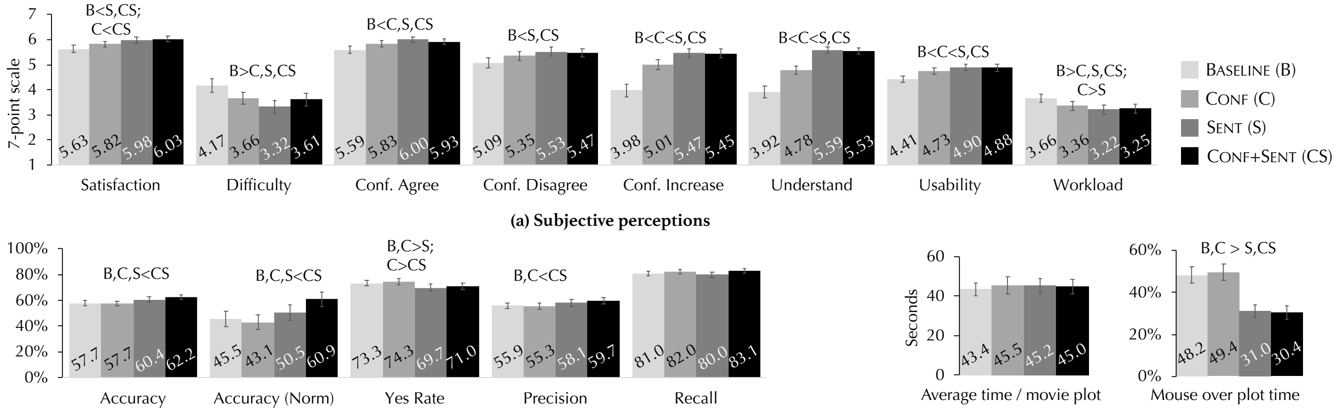
Our RQ1 results found four important trends. Our four interface conditions varied according to the inclusion/exclusion of two explainability features: confidence values and sentence highlighting. The first two trends suggest that having either explainability feature available (Conf or Sent) is better than having none (Baseline).

First, participants reported better perceptions when the interface displayed confidence values. Compared to Baseline, participants in Conf reported lower levels of difficulty and workload, as well as greater levels of confidence, understandability, and usability.

Second, participants reported better perceptions when the interface included the sentence-highlighting feature. Compared to condition Baseline, participants in condition Sent reported lower levels of difficulty and workload, as well as greater levels of satisfaction, confidence, understandability, and usability.

Third, if only one feature is available, sentence-highlighting is better than confidence values. This trend can be seen by comparing condition Sent versus Conf. Compared to Conf, participants in Sent reported lower levels of workload, as well as greater levels of understandability, usability, and confidence increase.

Finally, in terms of perceptions, having both explainability features available seems like the best option. This trend can be seen by comparing condition Conf+Sent versus Conf and Sent. Compared to condition Conf, participants in condition Conf+Sent reported greater levels of satisfaction, understandability, usability and confidence increase. Additionally, across all measures, participants reported similar perceptions between condition Conf+Sent and Sent (i.e., no significant differences). Later in Section 4.2, we show that objective performance was slightly higher in condition Conf+Sent than Sent.



(a) Subjective perceptions

(b) Objective performance

(c) Behavioral Measures

Figure 4: Main effects of interface on (a) perceptions, (b) performance metrics, and (c) behavioral measures.

4.2 RQ2: Effects on Performance

In RQ2, we investigate the effects of the interface condition on participants' performance when deciding to agree or disagree with the system. As described in Section 3.7, we measured performance from five perspectives: (1) accuracy, (2) normalized accuracy, (3) percentage of times the participant agreed with the system (i.e., yes rate), and (4-5) precision and recall with respect to the 5 true positive movies in each sequence of 10 movies.

Figure 4(b) shows these measures and indicates all significant pairwise comparisons. The interface condition had a significant main effect on four measures: (1) accuracy ($F(3{,}597) = 6.01$, $p < .001$); (2) normalized accuracy ($F(3{,}597) = 8.18$, $p < .001$); (3) yes rate ($F(3{,}597) = 5.78$, $p < .005$); and (4) precision ($F(3{,}597) = 5.18$, $p < .005$). The interface condition had no significant effect on recall.

Our RQ2 results found three important trends. First, participants had a strong bias towards agreeing versus disagreeing with the system. Across all interface conditions, participants agreed with the system about 70-75% of the time (i.e., yes rate). As previously mentioned, regardless of the interface condition, each sequence of 10 movies included 5 movies where the correct choice was to agree with the system (i.e., true positives for all selected genres) and 5 movies where the correct choice was to disagree with the system (i.e., false positives for at least one selected genre). Overall, while participants should have agreed/disagreed with the system about 50% of the time, they gave the system the "benefit of the doubt".

Second, participants achieved the best performance with the CONF+SENT interface in terms of accuracy, normalized accuracy, and precision. Our results suggest that the CONF+SENT interface helped participants identify false positives. To illustrate, compared to the BASELINE condition, participants using CONF+SENT had lower yes-rates (agreed less) and higher precision (i.e., had a greater percentage of times they correctly disagreed with the system).

Finally, participants achieved the best performance in condition CONF+SENT, which included both explainability features. An important question is: Did both explainability features contribute equally to performance? Our results suggest that *most* (albeit not all) of the performance improvement associated with condition CONF+SENT can be attributed to the sentence-highlighting feature. This trend can be seen by comparing condition CONF+SENT against conditions CONF and SENT. Including the sentence-highlighting feature yielded a significant performance improvement from condition CONF to condition CONF+SENT in terms of three metrics: accuracy, normalized accuracy, and precision. Conversely, including the confidence value feature yielded a significance performance improvement from condition SENT to condition CONF+SENT in terms of only one metric: normalized accuracy. Thus, while both explainability features added value, it appears that the sentence-highlighting feature added more value than the confidence value feature.

4.3 RQ3: Effects on Behaviors

In RQ3, we investigate the effects of the interface condition on participants' behaviors. To this end, we considered three behavioral measures: (1) the average completion time (in seconds), (2) the percentage of time the mouse hovered over the movie's plot, and (3) the percentage of time that sentences were highlighted on the interface (only for conditions SENT and CONF+SENT). Figure 4(c) shows these measures and indicates all significant pairwise comparisons.

Our results found three main trends. First, as shown in Figure 4(c), the interface condition did not have a significant effect on the time participants took to agree/disagree with the system. Participants took about 45 seconds per movie across all interface conditions.

Second, as shown in Figure 4(c), the interface condition had a significant effect on the percentage of time that participants hovered their mouse over the movie's plot summary ($F(3{,}597) = 87.41$, $p < .001$). This value was much greater for the BASELINE and CONF conditions as compared to the SENT and CONF+SENT conditions. One possible explanation, partially supported by the next trend, is that participants in conditions SENT and CONF+SENT made heavy use of the sentence-highlighting feature. In other words, participants used their mouse to click genres to highlight, which drove their mouse away from the region of the interface containing the plot summary. We elaborate on this trend in Section 5.

Finally, participants in conditions SENT and CONF+SENT made heavy use of the sentence-highlighting feature. Across both conditions, the sentence-highlighting feature was turned "on" 61% and 62% of the time. Based on a paired t-test, this difference was not significant. In Section 5, we investigate whether the confidence values in condition CONF+SENT influenced participants to highlight sentences for certain genres more than others.

5 DISCUSSION

In this section, we summarize the insights revealed by our results, compare them to results from prior work, and report on additional analyses regarding RQ2 and RQ3.

5.1 Summary of Results

Our results suggest that both explainability features (i.e., confidence values and sentence highlighting) had positive effects on participants' perceptions of the interface and their experience (RQ1), as well as their objective performance in making correct agree/disagree decisions (RQ2). Both explainability features had positive effects in isolation (i.e., conditions Conf and Sent versus Baseline) and in combination (i.e., condition Conf+Sent versus Conf and Sent).

Our RQ1-RQ3 results suggest four trends worth noting.

First, both explainability features were designed to "nudge" participants to consider *additional* information—per-genre prediction confidence values and influential sentences in the plot. Both explainability features improved perceptions and performance. An important question is: Did they also require more effort? Our results suggest that this is not the case. In terms of perceptions, participants reported significantly *lower* levels of difficulty and workload in conditions Conf, Sent, and Conf+Sent versus Baseline (Figure 4(a)). This result suggests that **both explainability features made it *easier* (and not *harder*) for participants to scrutinize the system**. In line with this trend, participants spent roughly equal time per movie across all interface conditions (Figure 4(c)).

Second, our results suggest that **between both explainability features, the sentence-highlighting feature was slightly more influential**. Based on our RQ3 results, participants made heavy use of the sentence-highlighting feature. In terms of perceptions (Figure 4(a)), condition Sent had either comparable or better outcomes than condition Conf and comparable outcomes as condition Conf+Sent. Similarly, in terms of performance (Figure 4(b)), condition Sent had comparable outcomes to condition Conf+Sent across all metrics *except* for normalized accuracy. In terms of normalized accuracy, condition Conf+Sent outperformed all other conditions. Note that the sentence-highlighting feature alone did not significantly improve performance metrics in RQ2 (Sent vs. Baseline). This trend resonates with results from prior studies, which found that feature-based explanations did not improve the classification accuracy of participants [7, 43]. However, in our results, when confidence values were also provided (Conf+Sent vs. Sent), the sentence-highlighting feature yielded significant improvements in performance. One possible explanation is that the confidence values helped participants *contextualize* how the highlighted sentences should be interpreted. Highlighted sentences for high-confidence genres should be used for sanity-check verification and highlighted sentences for low-confidence genres should invite more questioning.

Third, our results found that participants had a strong bias in favor of agreeing with the system (Figure 4(b)). Participants should have agreed with the system about 50% of the time, but agreed with the system about 70-75% of the time across interface conditions. Our RQ2 results suggest that **our explainability features mostly improved participants' performance by helping them *reject* false positive cases**. Again, in this respect, the sentence-highlighting feature played a critical role. This trend can be observed by comparing Conf+Sent versus Conf. Condition Conf+Sent had a lower yes-rate and higher precision. In other words, by seeing influential sentences, participants were better able to detect false positive mistakes. We believe that the sentence-highlighting feature allowed participants to understand *why* the system made a mistake. For example, for one false positive movie, the system highlighted the following sentence as being "most influential" to the genre Crime: "Leonard Hoffman is an L.A. insurance agent with a problem on his hands." In this case, the word "agent" (taken out of context) probably influenced the system to incorrectly predict Crime. This trend is also consistent with prior work, which found that showing feature-based explanations encouraged users to be more critical and agree less with the system [7, 43].

5.2 Additional RQ2 Analysis

Our RQ2 results found that the interface condition had a significant effect on the extent to which participants correctly agreed/disagreed with the system. An important follow-up question is: Were our explainability features (i.e., confidence values and sentence highlighting) more helpful in some situations than others? More specifically, **did the explainability features help participants make more accurate decisions for challenging movies**?

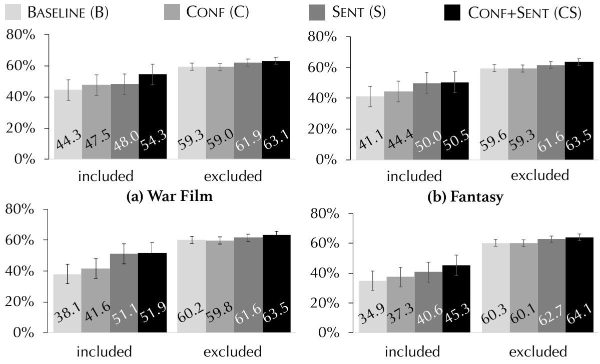
We explore this question by considering whether our explainability features helped participants make more accurate decisions for movies with a specific genre in the selected set of genres as filtering criteria. To perform this analysis, we used multilevel modeling. Specifically, we fit a multilevel *logistic* regression model to predict whether the participant made a correct decision (i.e., binary outcome) based on the interface condition and whether the movie had a specific genre in the selected set. In the model, we included terms to capture a possible interaction effect between the interface condition and the movie having a specific genre. Participant ID was included as a random factor (i.e., random y-intercept).

Table 1 shows the accuracy of participants' decisions for movies that excluded versus included each specific genre in the selected set. Column '% diff.' shows the percent increase in accuracy for movies that included the corresponding genre. The top genres had a significantly positive main effect on accuracy—participants had higher accuracy for movies that included the genre. The bottom genres had a significantly negative main effect—participants had lower accuracy for movies that included the genre. In other words, the bottom genres were the most difficult for participants. There are several possible explanations for why these genres were difficult. First, some of these genres may be closely related to others (e.g., Fantasy and Science Fiction). Second, some of these genres may have more diverse vocabulary (e.g., Mystery). Finally, perhaps many participants were unfamiliar with some of these genres (e.g., Period Piece). The most difficult genres for participants also happen to be the most difficult genres for classifiers (Section 3.3).

The interface condition had a significant interaction effect for only one genre: Period Piece ($p < .05$). Figure 5 shows the interaction effect between the interface condition and each of the four most difficult genres, including Period Piece. While the interaction effect was only significant for Period Piece, all of the most difficult genres had the same trend—the interface condition had a *stronger* impact on the accuracy of participants' decisions for movies that included (versus excluded) the genre in the selected set. In other words, **our explainability features were more helpful in challenging situations**.

Table 1: Differences in accuracy values (%) for movies that included versus excluded each genre in the selected set. '***', '**', and '*' denote significant differences at $p < .001$, $p < .01$, and $p < .05$ level.

	included	excluded	% diff.
Animation	71.3	57.7	24%***
Family Film	68.6	57.8	19%***
Comedy	65.7	56.6	16%***
Romance	64.3	58.3	10%***
World Cinema	62.4	58.9	6%*
Crime	61.4	59.2	4%
Drama	60.1	59.1	2%
Action/Adventure	59.8	59.4	1%
Thriller	56.3	60.2	-6%**
Science Fiction	54.1	60.1	-10%***
Western	51.4	60.4	-15%***
Horror	49.5	61.0	-19%***
War Film	48.5	60.8	-20%***
Fantasy	46.5	61.0	-24%***
Period Piece	45.7	61.3	-25%***
Mystery	39.5	61.8	-36%***



(a) War Film

(b) Fantasy

(c) Period Piece

(d) Mystery

Figure 5: Accuracy values across interface conditions for movies included versus excluded the genre of War Film, Fantasy, Period Piece, and Mystery in the selected set.

This trend was more pronounced (and significant) for Period Piece. A period piece is a movie set in the past, where the historical context plays an important role in the plot. One possible explanation is that the sentence-highlighting feature helped participants understand the definition of this nuanced genre and make more accurate agree/disagree decisions. It should be noted that even the classifier for Period Piece had low performance (F1 = 0.25).

5.3 Additional RQ3 Analysis

Our RQ3 results found that participants made heavy use of the sentence-highlighting feature in Sent and Conf+Sent. As mentioned in Section 4.3, the sentence-highlighting feature was turned "on" 61-62% of the time across both conditions. In this section, we present a follow-up analysis on how our explainability features influenced participants' behaviors.

Our RQ3 results found no significant differences on the extent to which participants used the sentence-highlighting feature in conditions Sent and Conf+Sent. In spite of this, we wanted to know whether the confidence values included in condition Conf+Sent (and not Sent) influenced participants to scrutinize some genres more than others. In other words: **Did the confidence values influence participants to scrutinize genres more *deliberately* and less *randomly*?**

To address this question, we performed the following analysis. First, for each movie judgement in conditions Sent and Conf+Sent, we computed a probability distribution over selected genres. This distribution describes the percentage of time each selected genre had sentence-highlighting turned "on" (out of the total time sentence-highlighting was turned "on" for any genre). Then, we computed the average entropy of this probability distribution in conditions Sent and Conf+Sent. In this respect, lower entropy values indicate that participants highlighted sentences for some genres more than others (i.e., the focus on genres was more skewed and less uniform).

In condition Sent, we observed an average entropy of 0.894. Conversely, in condition Conf+Sent, we observed an average entropy of 0.834 (a lower value). Again, using multilevel modeling, the difference in entropy values was statistically significant ($p < .001$). This result suggests that **participants used the sentence-highlighting feature to scrutinize genres more *selectively* in Conf+Sent and more *randomly* or *uniformly* in Sent**. We interpret this as evidence that the confidence values displayed in condition Conf+Sent influenced participants to scrutinize certain genres more than others.

6 CONCLUSION

In this paper, we conducted an empirical study to investigate how two explainability features (i.e. confidence values and sentence highlighting) may assist users to scrutinize and evaluate faceted filtering results, where facet-values are automatically assigned and can have mistakes. We found that both explainability features made the task easier as perceived by participants. Additionally, both features enabled participants to achieve better objective performance at no extra cost in time.

Implications: Our results imply that it is useful for faceted search systems that automatically assign facet-values to documents to provide tools (e.g., confidence values and rationale highlighting) for searchers to scrutinize why a document has been assigned a specific facet-value. The benefit of such tools can be substantial, especially in tasks where information needs are nuanced (e.g., professional search) and a large number of false positives need to be sifted through (e.g., the screening phase of a literature search). A searcher may leverage confidence values to selectively scrutinize specific facet-values, and then use the rational-highlighting feature to assess if the system made a prediction using valid evidence.

Future directions: First, our results show that in the best case, participants still accepted about 40% false positives. Future work should focus on developing scrutability features that help users counteract this automation bias and effectively reject false positive predictions. Second, our results suggest that neither of the two explainability features helped improve recall. Future work should investigate common traits of such false negative cases and develop scrutability features to save these cases from being rejected. Finally, the current work was limited to a synthetic use-case on a result presentation interface without interactive search functionality. A natural next step is to evaluate the effects of our explainability features in tasks where users themselves create faceted search and filtering criteria or perform faceted relevance judgments.

REFERENCES

[1] Krisztian Balog, Filip Radlinski, and Shushan Arakelyan. 2019. Transparent, scrutable and explainable user models for personalized recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 265–274.

[2] David Bamman, Brendan O'Connor, and Noah A Smith. 2013. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 352–361.

[3] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. 2019. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2429–2437.

[4] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[5] John Brooke. 2013. SUS: a retrospective. *Journal of usability studies* 8, 2 (2013), 29–40.

[6] Claudio Carpineto, Stanislaw Osiński, Giovanni Romano, and Dawid Weiss. 2009. A survey of web clustering engines. *ACM Computing Surveys (CSUR)* 41, 3 (2009), 1–38.

[7] Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2020. Feature-Based Explanations Don't Help People Detect Misclassifications of Online Toxicity. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 95–106.

[8] Mon Chu Chen, John R. Anderson, and Myeong Ho Sohn. 2001. What Can a Mouse Cursor Tell Us More? Correlation of Eye/Mouse Movements on Web Browsing. In *CHI '01 Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, NY, USA, 281–282.

[9] Ian Covert, Scott Lundberg, and Su-In Lee. 2020. Feature Removal Is a Unifying Principle for Model Explanation Methods. *arXiv preprint arXiv:2011.03623* (2020).

[10] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[11] Jody Condit Fagan. 2010. Usability studies of faceted browsing: A literature review. *Information Technology and Libraries* 29, 2 (2010), 58–66.

[12] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 80–89.

[13] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*. PMLR, 1321–1330.

[14] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*. Advances in Psychology, Vol. 52. North-Holland, 139–183.

[15] Marti Hearst. 2009. Integrating Navigation with Search. In *Search User Interfaces*. Cambridge University Press, Chapter 8.

[16] Marti A Hearst. 1995. Tilebars: Visualization of term distribution information in full text information access. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 59–66.

[17] Orland Hoeber, Daniel Schroeder, and Michael Brooks. 2009. Real-world user evaluations of a visual and interactive Web search interface. In *2009 13th International Conference Information Visualisation*. IEEE, 119–126.

[18] Tom Hope, Jason Portenoy, Kishore Vasan, Jonathan Borchardt, Eric Horvitz, Daniel Weld, Marti Hearst, and Jevin West. 2020. SciSight: Combining faceted navigation and research group detection for COVID-19 exploratory scientific search. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 135–143. https://doi.org/10.18653/v1/2020.emnlp-demos.18

[19] Ece Kamar. 2016. Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence.. In *IJCAI*. 4070–4073.

[20] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[21] Been Kim, Cynthia Rudin, and Julie A Shah. 2014. The bayesian case model: A generative approach for case-based reasoning and prototype classification. In *Advances in neural information processing systems*. 1952–1960.

[22] Weize Kong and James Allan. 2014. Extending faceted search to the general web. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. 839–848.

[23] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 29–38.

[24] Benjamin CG Lee and Daniel S Weld. 2020. Newspaper Navigator: Open Faceted Search for 1.5 Million Images. In *Adjunct Publication of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 120–122.

[25] Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing Neural Predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 107–117.

[26] Piyawat Lertvittayakumjorn and Francesca Toni. 2019. Human-grounded Evaluations of Explanation Methods for Text Classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 5198–5208.

[27] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems* 30 (2017), 4765–4774.

[28] Yuqing Mao and Zhiyong Lu. 2017. MeSH Now: automatic MeSH indexing at PubMed scale via learning to rank. *Journal of biomedical semantics* 8, 1 (2017), 1–9.

[29] Iain J Marshall, Joël Kuiper, and Byron C Wallace. 2016. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. *Journal of the American Medical Informatics Association* 23, 1 (2016), 193–201.

[30] Siyu Mi and Jiepu Jiang. 2019. Understanding the Interpretability of Search Result Summaries. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 989–992.

[31] Christoph Molnar. 2020. *Interpretable machine learning*.

[32] Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*. 625–632.

[33] National Library of Medicine. 2017. Frequently Asked Questions about Indexing for MEDLINE. https://www.nlm.nih.gov/bsd/indexfaq.html. Accessed: 2021-05.

[34] Jerome Ramos and Carsten Eickhoff. 2020. Search Result Explanations Improve Efficiency and Trust. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1597–1600.

[35] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " 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*. 1135–1144.

[36] Axel J Soto, Piotr Przybyła, and Sophia Ananiadou. 2019. Thalia: semantic search engine for biomedical abstracts. *Bioinformatics* 35, 10 (2019), 1799–1801.

[37] Harini Suresh, Steven R Gomez, Kevin K Nam, and Arvind Satyanarayan. 2021. Beyond Expertise and Roles: A Framework to Characterize the Stakeholders of Interpretable Machine Learning and their Needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[38] Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. 2020. Microsoft academic graph: When experts are not enough. *Quantitative Science Studies* 1, 1 (2020), 396–413.

[39] Chih-Hsuan Wei, Alexis Allot, Robert Leaman, and Zhiyong Lu. 2019. PubTator central: automated concept annotation for biomedical full text articles. *Nucleic acids research* 47, W1 (2019), W587–W593.

[40] Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. 2013. PubTator: a web-based text mining tool for assisting biocuration. *Nucleic acids research* 41, W1 (2013), W518–W522.

[41] Honghan Wu, Giulia Toti, Katherine I Morley, Zina M Ibrahim, Amos Folarin, Richard Jackson, Ismail Kartoglu, Asha Agrawal, Clive Stringer, Darren Gale, et al. 2018. SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. *Journal of the American Medical Informatics Association* 25, 5 (2018), 530–537.

[42] Yongfeng Zhang, Xu Chen, et al. 2020. Explainable Recommendation: A Survey and New Perspectives. *Foundations and Trends® in Information Retrieval* 14, 1 (2020), 1–101.

[43] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.