Understanding the Cognitive Influences of Interpretability Features on How Users Scrutinize Machine-Predicted Categories

Jiaming Qu
University of North Carolina at Chapel Hill
Chapel Hill, North Carolina, USA
jiaming@unc.edu

Jaime Arguello
University of North Carolina at Chapel Hill
Chapel Hill, North Carolina, USA
jarguello@unc.edu

Yue Wang
University of North Carolina at Chapel Hill
Chapel Hill, North Carolina, USA
wangyue@unc.edu

ABSTRACT

The goal of interpretable machine learning (ML) is to design tools and visualizations to help users scrutinize a system's predictions. Prior studies have mostly employed *quantitative* methods to investigate the effects of specific tools/visualizations on outcomes related to objective performance—a human's ability to correctly agree or disagree with the system—and subjective perceptions of the system. Few studies have employed *qualitative* methods to investigate *how* and *why* specific tools/visualizations influence performance, perceptions, and behaviors. We report on a lab study ($N = 30$) that investigated the influences of two interpretability features: confidence values and sentence highlighting. Participants judged whether medical articles belong to a predicted medical topic and were exposed to two interface conditions—one with and one without interpretability features. We investigate the effects of our interpretability features on participants' performance and perceptions. Additionally, we report on a qualitative analysis of participants' responses during an exit interview. Specifically, we report on *how* our interpretability features impacted different cognitive activities that participants engaged with during the task—reading, learning, and decision making. We also describe ways in which the interpretability features introduced challenges and sometimes led participants to make mistakes. Insights gained from our results point to future directions for interpretable ML research.

CCS CONCEPTS

• **Human-centered computing** → **User studies**; **Empirical studies in interaction design**; • **Applied computing** → **Annotation**.

KEYWORDS

Interpretable Machine Learning, User Study, Cognitive Activities

ACM Reference Format:
Jiaming Qu, Jaime Arguello, and Yue Wang. 2023. Understanding the Cognitive Influences of Interpretability Features on How Users Scrutinize Machine-Predicted Categories. In *ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '23), March 19–23, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3576840.3578315

CHIIR '23, March 19–23, 2023, Austin, TX, USA

ACM ISBN 979-8-4007-0035-4/23/03…$15.00
https://doi.org/10.1145/3576840.3578315

1 INTRODUCTION

Information access systems increasingly apply machine learning (ML) to automatically classify documents. This process can help end users find relevant content more effectively and efficiently by filtering large volumes of search results based on relevant categories. For example, systems such as AMiner [30] and Dimensions [9] allow users to filter publications and authors using automatically assigned research topics. Similarly, the U.S. National Library of Medicine (NLM) has recently started using machine learning to automatically assign Medical Subject Heading (MeSH) terms to scientific articles that can be searched through PubMed [20].

One of the primary goals of interpretable ML is to build systems that are scrutable by end users. In recent years, researchers have developed different tools and visualizations to help people scrutinize ML-based predictions in domains such as topic categorization [22], sentiment analysis [1], deception detection [14], content moderation [4], and disease diagnosis [15]. In these studies, researchers have investigated whether specific tools and visualizations can improve outcomes related to *objective performance* (e.g., a human's ability to correctly agree or disagree with the system), as well as *subjective perceptions* of the system (e.g., trust) and the experience of scrutinizing its predictions (e.g., satisfaction).

Studies in interpretable ML have mostly taken a *quantitative* approach. Quantitative approaches mainly focus on statistical patterns aggregated across study participants and leverage data originating from participants' decisions to agree or disagree, questionnaire responses, and behaviors logged by the system. Quantitative approaches provide few insights about *how* or *why* specific tools and visualizations impacted performance, perceptions, and behaviors. Additionally, they provide few insights into the challenges introduced by specific tools and visualizations or the reasons they might fail to help people scrutinize a system's predictions.

In this paper, we report on a lab study ($N = 30$) that employed both quantitative and qualitative analyses (i.e., a mixed methods approach). During the study, participants were shown a series of biomedical articles (title + abstract) that were each *automatically* assigned to a specific medical *topic* (e.g., mortality) within the context of a medical *subject* (e.g., COVID-19). For each article, participants were asked to agree or disagree with the system. Each participant was exposed to two interface conditions (i.e., a within-subjects design). In the Baseline condition, participants had to base their decisions using only the article's title and abstract. Conversely, in the Conf+Sent condition, participants were provided with two interpretability features: confidence value and sentence highlighting. The confidence value feature displayed the system's confidence in its prediction of the target topic. The sentence highlighting feature

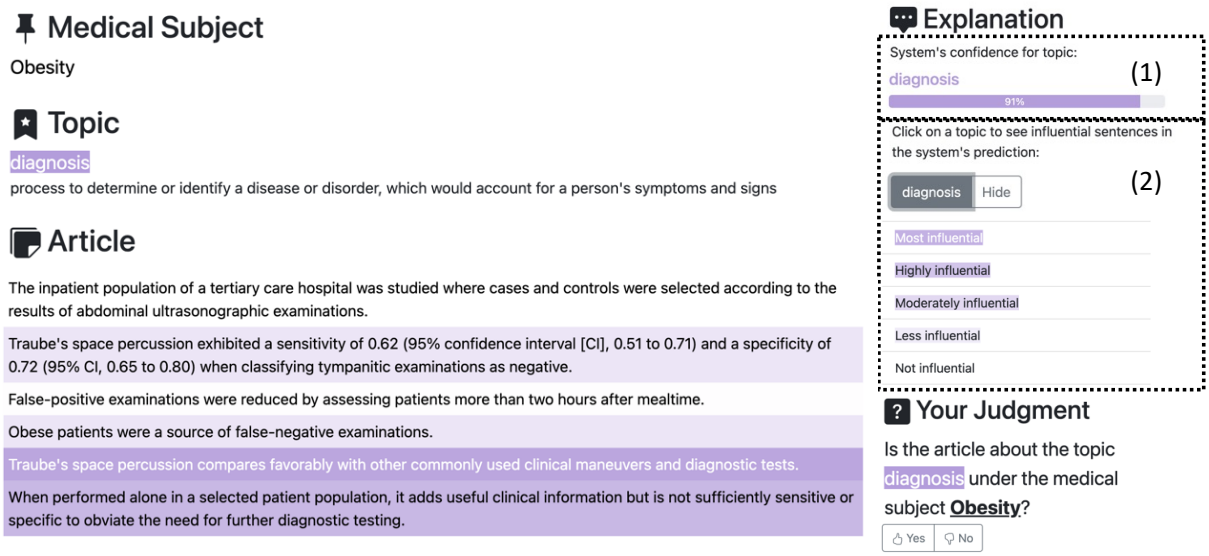

Figure 1: A screenshot of the Conf+Sent interface condition, which included two interpretability features: (1) confidence values and (2) sentence highlighting. The Baseline condition excluded both interpretability features.

allowed participants to determine which sentences were the most influential in predicting the target topic. As shown in Figure 1, the highlighting intensity of each sentence conveyed which sentences were more or less influential. After each interface condition, participants completed a questionnaire that asked about their perceptions of the system and the task. Finally, at the end of the study session, participants completed an exit interview in which they were asked about their general strategies and experiences in both interface conditions. Additionally, participants were shown articles for which they made correct and incorrect decisions and were asked about their thought processes during those specific cases. Participants' comments during the exit interview were analyzed using qualitative techniques to gain insights about the influences of our two interpretability features on participants' approach to the task and the challenges they faced.

Our study investigated three research questions (RQ1-RQ3).

- **RQ1:** What are the effects of our interpretability features on participants' ability to correctly agree or disagree with the system? To address RQ1, we compared participants' decisions to agree or disagree against ground-truth labels.
- **RQ2:** What are the effects of our interpretability features on participants' perceptions of the system and the task? To address RQ2, we analyzed participants' responses to questionnaires completed after each interface condition.
- **RQ3:** What are the effects of our interpretability features on participants' approaches to the task and the challenges they faced? With respect to challenges faced, we also considered possible reasons for why participants made mistakes.

In terms of RQ1 and RQ2, while our interpretability features did *not* improve participants' performance (RQ1), they did improve participants' perceptions of the system and their experiences scrutinizing the system's predictions (RQ2). Additionally, in terms of RQ3, our qualitative analysis of participants' comments during the exit interview revealed several important trends.

First and foremost, while deciding to agree or disagree with the system, participants engaged in three distinct (albeit interrelated) cognitive activities: (1) reading, (2) learning, and (3) decision making. In terms of reading, participants spent time reading the given article, scanning the text for evidence, and examining the context of topically relevant terms. In terms of learning, participants spent time learning about the meaning of specific terms in the text (e.g., What is plasma?), as well as the definition and scope of the target topic (e.g., What is the scope of immunology?). Finally, in terms of decision making, participant spent time weighing the evidence in order to decide whether to agree or disagree with the system.

It is not surprising that participants engaged in these three cognitive activities. However, these three activities provided us with a useful framework for analyzing the influences of our interpretability features. Our RQ3 results found that our interpretability features influenced the way participants approached all three activities. Additionally, our RQ3 results revealed challenges introduced by our interpretability features with respect to all three activities.

2 RELATED WORK

The goal of interpretable machine learning (ML) is to explain or present ML models in *understandable* ways to support decision making, model debugging, scientific discovery, and/or auditing [6, 19]. Text classification is a frequently studied task in this line of research. Prior work has considered both *global* and *local* explanations. Global approaches explain a model's overall behavior across the entire feature space and local approaches explain a model's decision for a specific instance. In our study, participants scrutinized a system's decision to assign a document to a category. In the Conf+Sent condition, participants interacted with two *local* interpretability features. For text classification, local explanations often highlight parts of the document (e.g., words or sentences) that strongly influenced the classifier to predict the target category [16, 18, 23]. Other forms of local explanations include example-based [12], rule-based [24], and counterfactual approaches [34].

Much research has used *quantitative* techniques to evaluate the effectiveness of interpretability tools/visualizations in supporting humans with specific tasks. Some of these evaluation tasks mimic the types of tasks that ML developers engage in (i.e., model debugging before deployment). Examples include: (1) discovering *spurious* features that may not generalize beyond the training data [23]; (2) comparing classifiers and selecting the best one [23, 27]; and (3) evaluating a classifier by guessing its prediction based on its local explanations [5, 8, 24]. Other evaluation tasks (such as ours) mimic the types of tasks that "end users" of ML systems engage in [4, 14, 21, 22, 32]. Such tasks are often referred to as *machine-assisted classification tasks*—a human must decide whether a document belongs to a category given a system's prediction and interpretability features. Previous studies have focused on tasks such as classifying the toxicity of social media comments [4], the authenticity of hotel reviews [13, 14], and the sentiment of book reviews [1]. Our study focused on the academic domain. We simulate a scenario in which a searcher is interested in a topic, the search system (e.g., PubMed or Dimensions) returns an article automatically classified as belonging to the topic, and the searcher must judge whether the article truly belongs to the topic.

Studies have found that interpretability features do not *consistently* help users make more accurate classification decisions. When the classifier is *less* accurate than unaided users, interpretability features can be helpful in some tasks [1] but not others [4, 32]. Conversely, when the classifier is *more* accurate than unaided users, interpretability features can help users improve their accuracy to a level that is higher than their own unaided accuracy but still lower than the classifier's accuracy [13, 14].

Researchers have proposed several explanations for why interpretability features do not consistently help users. First, it is challenging for users to make sense of machine interpretations because machines and humans do not reason in the same manner. For example, linear models sum the weights of individual words, while users aim to comprehend the text as a whole [28]. Second, since ML models represent words and categories as nominal symbols, they cannot intrinsically help users understand the meaning of unfamiliar terms or the target categories (e.g., the distinction between tree species in predicting forest coverage [31]). Third, ML models identify words that are strongly correlated with a category. However, humans look for logical reasons or hypotheses to explain a word's predictive power. For example, in fake review detection, Lai et al. [13] found "Chicago" to be predictive of fake reviews. Humans might need additional support in understanding (and trusting) this phenomenon—perhaps the city of Chicago has many fake reviews or perhaps genuine reviews tend to mention local neighborhoods versus the entire city. Fourth, machine interpretations can appear overly persuasive, leading to automation bias. This trend has been observed in tasks where machines outperform humans [14] and tasks where humans outperform machines [1]. Finally, machine explanations can bias a user's attention. For example, highlighting parts of a document can influence users to only read the highlighted parts and overlook the unhighlighted parts [4].

In our study, while our interpretability features improved participants' perceptions, they did not improve their performance. Importantly, our qualitative analysis of participants' responses during the exit interview provides insights about ways in which our interpretability features introduced challenges and failed to support important cognitive processes inherent in such tasks. In this respect, our results point to additional reasons for why interpretability features may not consistently help users (other than the ones above) and suggest directions for future work.

3 METHODS

To investigate RQ1-RQ3, we conducted a laboratory study with 30 participants (19 female and 11 male). Participants were recruited using an opt-in mailing list of graduate students at our university and by posting flyers around our campus. Participants' ages ranged from 21 to 49 (Mean = 26.77, S.D. = 6.51). During the study, participants were asked to judge whether a biomedical research article belongs to a machine-predicted medical topic. Given the complexity of the task, participants were required to have completed (or be pursuing) a graduate degree in a STEM-related field. Participants had educational backgrounds in information and library science (18), biomedical science (8), computer and data science (2), and city planning (2). Not every participant had an educational background in biomedical science. Instead, we wanted our participant pool to also include domain novices who might struggle with technical jargon and the meaning/scope of specific medical topics. Prior research estimates that 4.5% of web searches are health-related [7]. Thus, we wanted to include participants who might search for biomedical articles as part of their work and for personal reasons [33]. The study was approved by our university's Institutional Review Board.

Our goal was to investigate the influences of two interpretability features (i.e., confidence value and sentence highlighting) on the way people *scrutinize* and *judge* the accuracy of predictions made by an ML-based system. Participants were exposed to two interface conditions (i.e., a within-subjects design). The Baseline condition excluded both interpretability features and the Conf+Sent condition included both interpretability features (Section 3.1).

For each interface condition, participants were shown a sequence of 12 biomedical articles (only the title + abstract). Each article was associated with a medical subject and a topic. Participants were instructed that *medical subjects* represent domains of medical research (e.g., COVID-19) and that *topics* represent more specific and nuanced aspects of medical research (e.g., mortality) within the context of the medical subject. Participants were instructed that medical subjects were assigned manually by experts and that topics were predicted by a system using machine learning technology. Hence, the medical subject is *always* correct, but the medical topic may be correct or incorrect. For each article, participants were asked to: (1) judge whether the article *truly* belongs to the assigned topic (within the context of the medical subject) and (2) indicate their agreement or disagreement with the system's prediction of the medical topic. Within each sequence of 12 articles, 6 articles were true positive cases (i.e., the correct decision was to agree with the system) and 6 articles were false positive cases (i.e., the correct decision was to disagree with the system). However, participants were *not* aware of this distribution of true and false positive cases.

The study protocol proceeded as follows. After signing a consent form, participants completed a short demographics questionnaire. Next, participants watched a video describing the general purpose and protocol of the study. As previously mentioned, each participant was exposed to both interface conditions. Each treatment

involved the same sequence of steps. First, participants watched a short video introducing the next interface condition. Then, after completing a practice task, participants judged a sequence of 12 biomedical articles. After each sequence, participants completed a post-task questionnaire about their perceptions of the interface and the task (Section 3.4). Finally, participants completed an exit interview that asked about their strategies and experiences with both interfaces (Section 3.5). The order of interface conditions was balanced—15 participants were exposed to the Baseline condition first and 15 participants were exposed to the Conf+Sent condition first. Participants received US$30 for participating in the study. Our study materials are available online.

3.1 Interface Conditions

Each participant was exposed to two interface conditions. Figure 1 shows the interface in the Conf+Sent condition, which included both interpretability features (confidence value and sentence highlighting). The interface in the Baseline condition looked the same, but excluded both interpretability features.

Baseline Condition: In the Baseline condition, participants made judgements based only on the article title and abstract. Again, each article was associated with a medical subject (always correct) and a topic (possibly incorrect). The medical subject was displayed at the top of the page and the medical topic was displayed below the medical subject and was highlighted using a unique color. The interface also provided a short definition of the medical topic. The article's title and abstract were shown in the region below. As shown in Figure 1, participants were prompted with the question: "Is the article about topic X under medical subject Y?" Participants indicated their agreement by selecting "yes" or "no".

Conf+Sent Condition: As shown in Figure 1, in the Conf+Sent condition, participants had access to both interpretability features. The confidence value feature displayed the system's confidence in the article being assigned to the corresponding topic (e.g., diagnosis in Figure 1). Confidence values were represented as colored bars ranging from 0% to 100%, with the exact value displayed inside. As explained in Section 3.2, confidence values corresponded to the prediction confidence output by a logistic regression classifier.

Compared to the confidence value feature, the sentence highlighting feature was more interactive. As illustrated in Figure 1, clicking on the topic button (e.g., "diagnosis" in Figure 1) highlighted the most "influential" sentences in classifying the article as belonging to the topic. As displayed in the interface, participants were instructed to interpret the color intensity of each sentence as its level of influence. Clicking the "hide" button removed all sentence highlighting. The sentence highlighting feature was implemented by training a *document-level* logistic regression classifier (i.e., trained on titles + abstracts) and using it to make *sentence-level* predictions. Given a specific topic and article, the color intensity of each sentence s was determined according to

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

where $\mathcal{P}_{\text{raw}}(t|s)$ denotes the probability that sentence s belongs to topic t according to the document-level classifier for t. By default, a logistic regression classifier 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 $\mathcal{P}_{\text{raw}}(t|s) < 0.5$, meaning that the document-level classifier is more confident that s *does not* belong to t than vice-versa. Normalized confidence values were binned into five levels to be consistent with the color key in Figure 1.

Before each interface condition, participants watched a video introducing the next interface. In the video for the Conf+Sent condition, participants were instructed that confidence values close to 100% indicate that the system is highly confident that the article belongs to the corresponding topic and that values close to 50% indicate that the system is highly *unsure*. Participants were also instructed that the sentence highlighting feature allowed them to "see which sentences the system used to base its prediction". Participants were instructed that "brightly colored sentences contain the *most* evidence, lightly colored sentences contain *some* evidence, and unhighlighted sentences contain *no* evidence."

3.2 Study Design

Dataset: The biomedical articles used in our study originated from a PubMed dump from July 2021. The dataset contained about 29 million articles, each associated with a title, abstract, and Medical Subject Headings (MeSH) metadata. Of the available metadata, we focused on MeSH *headings* and *subheadings* [17]. In this study, we referred to these as *medical subjects* and *medical topics*, respectively.

For the study, we first selected six medical subjects that we believed would be relatively easy for participants to understand: COVID-19, depression, diabetes, hypertension, obesity, and sleep. These six medical subjects were associated with 76 unique medical topics. For the study, we selected 12 topics that: (1) were frequent enough to train a classifier (i.e., had enough positive examples); (2) were neither too general nor abstract; and (3) did not have similar meaning to other selected topics. We selected genetics, metabolism, therapy, pathology, drug effects, surgery, hematology, immunology, diagnosis, mortality, epidemiology, and complications.

Classification: To generate topic predictions, confidence values, and to implement the sentence highlighting feature, we trained 12 logistic regression classifiers (one per medial topic). All classifiers used the same unigram representation, which excluded stopwords and rare terms. To train each classifier, we gathered a balanced training set of 150K positive and 150K negative cases. On a held-out set of 200K randomly sampled articles, 10 (out of 12) classifiers achieved F1 scores between 0.40 and 0.70. Two classifiers, "complications" and "mortality", had F1 scores of about 0.30.

Experimental Design: Our experimental design is depicted in Figure 2. To conduct the study, we used 360 distinct articles. Each article was associated with a medical subject (e.g., COVID-19) and a *predicted* medical topic (e.g., mortality). These 360 articles were organized in 15 batches of 24 articles. Each batch of 24 articles was divided into two sequences of 12 articles. Each sequence of 12 articles met the following criteria: (1) two articles per medical subject; (2) one article per medical topic; and (3) one true positive and one false positive case per medical subject. Our study involved 30 participants. Every two participants completed the same batch of 24 articles (i.e., two sequences). The order of articles was kept consistent, but both participants experienced the interface conditions in different order—one participant ("user 1" in Figure 2) experienced

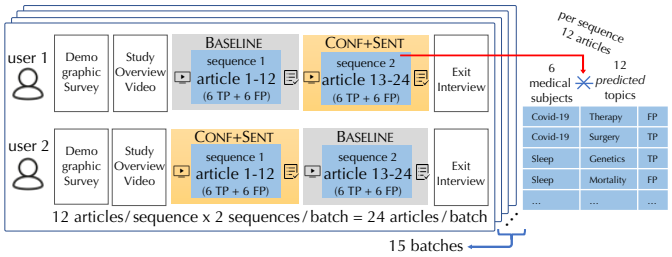

Figure 2: Experimental Design.

the BASELINE condition first and the other participant ("user 2" in Figure 2) experienced the CONF+SENT condition first.

This design ensured the following desired criteria. First, participants were exposed to a wide range of articles, medical subjects, and medical topics. Second, each sequence contained an equal number of true positive and false positive cases. This information was not communicated to participants but it enabled us to gauge participants' performance by comparing their accuracy to 50%—the expected accuracy from either always (dis)agreeing with the system or randomly guessing with equal probability. Finally, it balanced the exposure of interface conditions and their ordering across articles. In real-life scenarios, the distribution of true positives and false positives may not be balanced (e.g., a good system will likely present more true positives than false positives). Here we choose a balanced distribution to ensure that always (dis)agreeing with the system is not more accurate than random guessing and therefore a user accuracy higher than 50% can only be attributed to the user's correct decisions and not always (dis)agreeing with the system.

3.3 Performance Metrics (RQ1)

In RQ1, we investigate the effects of the interface condition on the extent to which participants made correct agree or disagree decisions. We measured performance using five metrics.

(1) Accuracy: the percentage of correct agree/disagree decisions. Each sequence of 12 articles included 6 true positive and 6 false positive cases. Thus, we expected accuracy values to be $> 50\%$.

(2) Precision: the percentage of "agree" decisions involving a true positive case. Precision captures whether participants *rejected* false positive cases when they agreed with the system.

(3) Recall: the percentage of true positive cases for which participants agreed with the system. Recall captures whether participants agreed with the system for *all* true positive cases.

(4) Yes Rate: the percentage of times participants agreed with the system. Yes-rate does not measure performance. Rather, it provides insights into participants' tendencies to agree with the system across interface conditions.

(5) Completion Time: The average time (in seconds) it took participants to agree or disagree with the system. Similar to yes-rate, completion time does not measure performance. However, it tells us how long it took participants to make decisions, which may be related to participants' level of engagement or effort.

3.4 Post-task Questionnaire (RQ2)

To address RQ2, participants completed a two-part questionnaire after judging each sequence of 12 articles in a specific interface condition. Participants responded to agreement statements using a 7-point scale ranging from "strongly disagree" to "strongly agree".

The first part of the questionnaire included 7 items that asked about participants' perceptions of: (1) satisfaction with their performance (1 item), (2) difficulty (1 item), (3) confidence (1 item), and (4) understandability of the system's predictions (4 items). The four understandability items had high internal consistency (Cronbach's $\alpha = 0.91$). Thus, we averaged responses to all four items to form one understandability measure.

The second part of the questionnaire asked about system usability (i.e., ease of use). To this end, we used the System Usability Scale (SUS) [3], which includes 10 items. Participants' responses had high internal consistency (Cronbach's $\alpha = 0.72$). Thus, we averaged responses to all 10 items to form one system usability measure.

3.5 Exit Interview & Qualitative Analysis (RQ3)

To investigate RQ3, after the post-task questionnaire for the *last* interface condition, participants completed an exit interview with the study moderator. All questions and answers were stated verbally and recorded. Participants were asked two groups of questions—one group of questions about the BASELINE condition and one group of questions about the CONF+SENT condition. These groups of questions were asked in the same order as the order in which the participant was exposed to the interface conditions.

BASELINE Condition Questions: First, participants were asked about their general strategies for deciding to agree or disagree with the system. They were asked follow-up questions about the types of evidence used to base their decisions and how they used each type of evidence. Then, participants were shown four articles along with their decisions: (1) one true positive for which they *correctly agreed*, (2) one true positive for which they *incorrectly disagreed*, (3) one false positive for which they *correctly disagreed*, and (4) one false positive for which the *incorrectly agreed*. For each article, participants were asked about their rationale to agree or disagree with the system without knowing the ground truth. If no article was associated with a specific case, then that case was skipped.

CONF+SENT Condition Questions: For the CONF+SENT condition, participants were asked *all of the same questions* asked for the BASELINE condition (described above). However, when asked about their general strategies, participants were asked a few additional questions about their use of both interpretability features. For each interpretability feature, participants were asked: (1) whether they engaged with the interpretability feature; (2) their motivation of engaging with the feature (i.e., what they were hoping to achieve); (3) their use of the feature for deciding to agree or disagree with the system (i.e., how the feature helped); and (4) any challenges faced while engaging with the feature.

To address RQ3, participants' responses during the exit interview were analyzed using qualitative techniques. First, all authors on the paper independently analyzed responses from 3 (out of 30) participants. Next, the authors met several times to define a set of thematic categories. Ultimately, we decided on four themes: (1) general strategies and heuristics; (2) usage of the confidence value and sentence highlighting features (i.e., motivation of engaging with the feature and its benefits); (3) challenges faced while engaging with the confidence value and sentence highlighting features; and (4) reasons for making incorrect decisions. Finally, one of the authors analyzed the data from all 30 participants using an inductive coding approach. That is, within each theme, new codes and

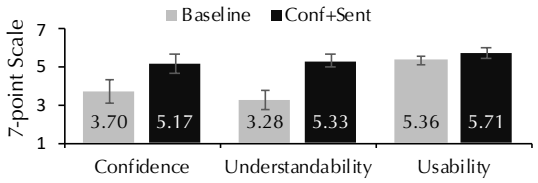

Figure 3: Significant effects of the interface condition on participants' perceptions (means and 95% confidence intervals).

definitions were developed as new and interesting phenomena were encountered. The codebook is in the online appendix.

4 QUANTITATIVE RESULTS: RQ1 & RQ2

In RQ1 and RQ2, we investigate the effects of the interface condition on participants' performance and perceptions, respectively. To test for statistical significance, we used multilevel modeling. Given the within-subjects nature of our data, we included the participant as a random factor (i.e., random y-intercept).

RQ1 Results: In RQ1, we investigate the effects of the interface condition on participants' performance when deciding to agree or disagree with the system. In all cases, we analyzed performance at the *sequence level*. For example, we first computed accuracy values for each sequence of 12 articles and then compared accuracy values across interface conditions.

The interface condition did *not* have a significant effect for any measure of performance. In the Conf+Sent condition, participants achieved *slightly* better performance in terms of accuracy (0.631 vs. 0.600), precision (0.639 vs. 0.597), and recall (0.717 vs. 0.683). Additionally, participants had very similar yes rates (0.586 vs. 0.583). In Section 6, we elaborate on possible reasons we did not observe significant differences in performance between interface conditions. Other studies also have found that interpretability features do not necessarily yield improvements in performance [4, 31, 32]. For example, [4] evaluated different ways of highlighting relevant portions of the text with respect to the predicted category. Results found that participants often made mistakes because they completely ignored unhighlighted portions of the text that contained positive evidence.

In terms of the average completion time, participants took significantly longer to make decisions in the Conf+Sent versus Baseline condition (119.99 vs. 105.13 seconds, $p < .05$). Based on our RQ3 results (Section 5), we see several possible explanations for this trend. First, participants reported engaging with our interpretability features to understand how they worked and decide whether they should be trusted. Second, participants reported using our interpretability features as a "sanity check". In such cases, participants first made their own decision and then engaged with our interpretability features to validate their hypothesis. Third, participants also reported that seeing low confidence values and few highlighted sentences made them spend more time scrutinizing the article before making a decision.

RQ2 Results: In RQ2, we investigate the effects of the interface condition on participants' perceptions of: (1) satisfaction, (2) difficulty, (3) confidence, and (4) understandability of the system's predictions, and (5) system usability. The interface condition had a significant effect on three measures. As shown in Figure 3, in the Conf+Sent condition, participants reported significantly higher levels of confidence ($p < .001$), understandability ($p < .001$), and system usability ($p < .05$).

5 QUALITATIVE RESULTS: RQ3

In RQ3, we aim to understand participants' behaviors in both interface conditions. As described in Section 3.5, we performed a qualitative analysis of participants' responses during the exit interview. We first report on the cognitive activities that participants engaged in during the task. Then for each of the two interface conditions, we describe participants' strategies and challenges faced.

5.1 Cognitive Activities

Based on participants' comments during the exit interview, we identified three major cognitive activities that participants engaged in during the task.

Reading comprehension was recognized as the first and foremost cognitive activity that participants engaged in. During the exit interview, participants were asked about *how* they read in both interface conditions, which helped us understand the influences of both interpretability features on participants' reading behaviors.

Learning was identified as the next cognitive activity. Not surprisingly, while reading, some participants had difficulty understanding aspects of the article or the given topic. Participants often took steps to overcome these knowledge gaps (e.g., by searching for term definitions) before making a final decision.

Decision making was the final cognitive activity participants engaged in. Ultimately, participants had to decide whether to agree or disagree with the system. As we will see, participants adopted different strategies and heuristics when deciding to agree/disagree.

5.2 Approaches in the Baseline Condition

Reading Comprehension: During the exit interview, participants commented on the *strategies* they took to read articles and the types of *evidence* they sought to inform their decisions.

Participants commented on two reading strategies: intensive reading ($N = 18$) and speed reading ($N = 14$). During intensive reading, participants carefully read the article to understand it thoroughly. During speed reading, participants merely skimmed the article to save time. Importantly, these two approaches to reading were not mutually exclusive. For example, P10 said: *"I skim first, but if I cannot understand the article, I will read each line to make sure I fully understand."*

Participants commented on seeking three types of evidence. All participants ($N = 30$) mentioned keywords spotting, which included the exact topic of the article, synonyms of the topic, terms provided in the topic definition (displayed on the interface), and semantically relevant terms based on the participant's prior knowledge. Participants ($N = 10$) also mentioned seeking a deep comprehension of the article. Participants mentioned that, in some cases, the simple presence of specific keywords was insufficient to determine whether the article belong to the topic. This was often true for highly nuanced topics. Finally, some participants ($N = 12$) looked for parts of the abstract describing the main objective of the research. Oftentimes, participants sought this information by *"check[ing] the first and last few lines of the abstract"*.

Learning: Participants performed learning when they did not fully understand terms in the abstract or the definition and scope of the target topic.

Many participants ($N = 19$) mentioned that the topic definition provided in the interface clarified the topic and helped them think about keywords to look for in the abstract.

Additionally, participants were permitted to search for definitions and background information about unfamiliar terms in the abstract. Several participants ($N = 7$) mentioned doing this (mostly using Google). Interestingly, one participant mentioned focusing on *"[unfamiliar] terms that appeared frequently in the abstract"*. This suggests that participants used specific heuristics (e.g., frequency) to make inferences about the importance of unfamiliar terms.

Decision Making: participants commented on agreeing with the system when they found *positive* evidence in the article (e.g., topically-relevant keywords) that matched their understanding of the given topic.

In general, participants agreed with the system more often than they should have. In the Baseline condition, they agreed with the system 58.6% of the time even though half the articles in each sequence were false positives. Participants gave the system the "benefit of the doubt" and assumed the system was more accurate than it was. One explanation is that participants had trouble finding *convincing* evidence to disagree (i.e., negative evidence). Participants stated that highly confident disagreements only happened when they found *"solid evidence to prove [the article] did not mention the topic or it was about something else."* Participants also mentioned confidently disagreeing only when they *understood* the system's mistake. For example, P11 said: *"The article was about surgery within COVID-19, but I did not make a confident decision [to disagree] until I found it was about educating [surgeons] online during the COVID-19 pandemic."* In this case, the participant noticed that the article did not discuss surgery as a *treatment* for COVID-19.

5.3 Approaches in the Conf+Sent Condition

In the Conf+Sent condition, participants were free to engage with or ignore our interpretability features. For example, some participants ($N = 15$) preferred reading articles with the sentence highlighting turned "on" and others ($N = 18$) with the sentences highlighting turned "off". Participants exhibited some common behaviors in both conditions. For example, participants who preferred reading articles with the sentence highlighting turned "off" exhibited similar reading behaviors as those found in the Baseline condition. In this section, we focus on *new* behaviors that were influenced by our interpretability features and were *not* observed in the Baseline condition.

Planning How to Read: Some participants mentioned using our interpretability features to form an initial hypothesis about the article's association with the topic before reading. Participants commented on basing these initial hypotheses on the confidence values ($N = 21$) as well as the number and intensity of highlighted sentences ($N = 4$). These initial hypotheses influenced three aspects of the task. First, both features influenced participants' expected difficulty in making a judgement ($N = 21$). Seeing a high confidence value and/or many (intensely) highlighted sentences influenced participants to expect the judgement to require less time and effort. Second, in some cases ($N = 8$), interpretability features influenced how participants decided to read the article (e.g., intensive reading versus skimming + keyword searching). Third, some participants ($N = 7$) commented that the confidence value feature influenced their decision to use the sentence highlighting feature. Two participants (P9 and P17) mentioned that *low* confidence values influenced them to *avoid* the sentence highlighting feature because it was not reliable: *"the system itself is not sure and does not deserve using."* This can be viewed as undesirable or counterproductive behavior. Even if the article is a boundary case, a highlighted sentence may still accurately represent the most relevant part of that article.

Deciding What to Read: In the Conf+Sent condition, our interpretability features influenced how participants read articles and sought evidence to base their decisions. Participants ($N = 16$) commented that the sentence highlighting feature influenced how they read articles and sought evidence. Based on participants' comments, the sentence highlighting feature influenced these behaviors in three ways: (1) order (e.g., reading highlighted sentences first); (2) effort (e.g., spending more time on highlighted sentences); and (3) focus (e.g., reading only highlighted sentences).

Participants appreciated that the sentence highlighting feature helped direct their attention to specific parts of the abstract. Some participants ($N = 11$) commented that this helped them complete the task more *quickly* and others ($N = 5$) more *accurately*.

Deciding What to Learn: In some cases, participants were largely unfamiliar with the given topic and needed to learn about the domain. In such cases, participants ($N = 16$) commented on using the sentence highlighting feature to guide their learning. We observed two different cases of this behavior.

First, in some cases, participants used the highlighted sentences to learn about topic-relevant keywords. For example, P12 said: *"I learned from highly influential sentences that the word plasma is a good indicator of the topic hematology, as it appeared very frequently."* Second, participants commented on using the highlighted sentences to determine which unfamiliar terms they needed to learn about. Seeing an unfamiliar term in a highlighted sentence prompted participants to gather background information about that term. For example, P22 said: *"Sometimes, although a [medical term] has been highlighted in a sentence, I still cannot understand what it's talking about so I need to search."*

Concurring on a Decision: Our interpretability features also played a role in helping participants validate decisions they had already made. Participants commented on using the confidence value feature ($N = 18$) and sentence highlighting feature ($N = 23$) to double-check their own decisions.

Participants commented that when they had decided to *agree* with the system, seeing a *high* confidence value validated their judgement and raised their confidence to agree. In some cases ($N = 3$), when participants had decided to *disagree* with the system, seeing a *low* confidence value (i.e., close to 50%) did not necessarily validate their judgement. In these cases, participants may have misinterpreted a 50% confidence value as "somewhat confident" rather than "very unsure" (i.e., just above the 50% threshold). We return to this point when in Section 5.5. Participants also used the highlighted sentences to verify whether the most and least influential sentences matched their expectations.

Some participants preferred one feature over the other for the purpose of validation. For example, P3 said: *"The confidence values*

are clear and more straightforward." Conversely, P19 said: *"The confidence values are only numbers, but I can find key evidence inside the highlighted sentences which makes me feel more secure."*

Debating on a Decision: Participants commented on using the confidence values ($N = 8$) and highlighted sentences ($N = 25$) to *scrutinize* the system's predictions. For example, P11 said: *"I will be more careful if I want to choose 'Yes' against low values or 'No' against high values."* Similarly, P30 stated: *"I confirm the system is wrong when I find the most [...] influential sentences do not make sense to me."* The sentence highlighting feature also helped participants understand *why* the system made a mistake. For example, in one case, a highlighted sentence contained topic-relevant keywords but referred to future research (i.e., not the focus of the article).

Delegating a Decision: Finally, in some cases, participants commented on being extremely unsure about their decision. This was mostly due to their unfamiliarity with the given topic or the complexity of the abstract. In such cases, rather than guessing randomly, participants commented on relying on the confidence value feature ($N = 2$) and sentence highlighting feature ($N = 1$). P6 noted that this only happened *"when the system [was] convincing, by showing a high confidence value and lots of highlighted sentences."*

5.4 Challenges in the Baseline Condition

As previously mentioned, during the exit interview of each condition, participants were shown two cases in which they made an incorrect decision and were asked to comment on their thought processes and rationales. Based on participants' comments, we identify two major sources of error that caused mistakes: (1) misunderstandings about the article and (2) misunderstandings about the target topic. Participants made mistakes when they were unfamiliar with terms in the article. In some cases, they searched for definitions and background information about unfamiliar terms to better *"understand their meaning"*. In other cases, they simply ignored unfamiliar terms to *"save time and do the task fast[er]"*. Participants also made mistakes when they were unfamiliar with the topic's definition or scope (i.e., inclusion and exclusion criteria). For example, the topic of *mortality* is defined as the "measure of the number of deaths in a population from a given cause in a set time period". One of our false positive articles discussed interviews with suicide survivors. Some participants incorrectly considered the article to be about mortality simply because it mentioned death.

In addition, participants made mistakes when they saw topically-relevant keywords without considering the context. For example, topically-relevant keywords appeared in sentences describing *future research directions*, which was not the main focus of the research. This type of mistake was made when participants incorrectly agreed with a *false positive* prediction (i.e., the bag-of-words classifier *also* missed the important contextual information surrounding the topically relevant keywords).

5.5 Challenges in the Conf+Sent Condition

Even with the help of our interpretability features, participants could still misunderstand the article or target topic and make incorrect decisions. Additionally, our interpretability features sometimes misled participants. This mostly happened when participants were unsure about their decision. For some false positive cases, participants commented on agreeing with the system when they *"saw a high confidence value and [topically-relevant] keywords among highlighted sentences."* In such cases, participants (like the system) failed to verify whether the topically-relevant keywords appeared in contexts describing the *main focus* of the research. For some true positive cases, participants commented on disagreeing with the system based on the interpretability features. For example, P14 and P29 commented on disagreeing with the system simply because they saw low confidence values close to 60%.

We noticed the following five major challenges faced by participants in the Conf+Sent condition.

First, many participants ($N = 18$) struggled with borderline cases with close to 50% confidence and few highlighted sentences, and they claimed to *"get almost no assistance from the system"*.

Second, some participants' comments revealed an incomplete mental model of the system. Several participants ($N = 12$) believed that a sentence's color intensity indicated its *number* of topically-relevant keywords. This is an incomplete mental model—a brightly colored (i.e., highly influential) sentence might *also* contain only a few (or a single) *highly* relevant keyword(s).

Third, in some cases ($N = 10$), mistakes made by our interpretability features resulted in loss of trust. P5 stated: *"the highlighted part contains (several) keywords but does not make sense."* When participants found errors in the highlighted sentences, they felt *"unsure and insecure"*. For example, P27 said: *"I was confused and stopped using the system after I found it made obvious mistakes."*

Fourth, some participants ($N = 8$) commented on struggling to interpret confidence values. For example, P23 said: *"I can tell the system is more persuasive when the confidence value is over 95% compared to 60%, but I do not know how big the difference is if there was only [a] 1% gap."* This result suggests that participants may have interpreted extreme confidence values in the same way, but mid-range values (e.g., 60%, 65%, 70%) very differently.

Finally, some participants ($N = 5$) needed time to establish trust and gauge how an interpretability feature can be useful. P16 said: *"I will first spend some time checking how they [interpretability features] work so that I can decide whether I should trust [them] or not."*

6 DISCUSSION AND IMPLICATIONS

In this section, we summarize our results, compare them to results from prior work, and discuss their implications.

6.1 RQ1: Effects of Performance

Our two interpretability features did *not* significantly improve participants' performance in correctly (dis)agreeing with the system. In terms of accuracy, participants had an average accuracy of 63% in the Conf+Sent condition and 60% in the Baseline condition.

Our RQ1 results are an important contribution to interpretable ML research. Prior work has evaluated interpretability features using a wide range of methods. Some tools have been evaluated anecdotally by discussing the insights provided by example explanations [29, 35]. Similar to our work, other evaluations have invited study participants to perform different types of tasks with the interface [5, 11, 23, 24, 27]. For example, tools have been evaluated based on their ability to help study participants: (1) correctly

identify the best classifier between two options [23, 27]; (2) correctly identify problematic features that should be dropped [23]; and (3) correctly guess the model's (hidden) prediction [5, 8, 24, 27]. These evaluation methods have one thing in common—participants played the role of a "system developer" who wants to understand a model's behavior. In our study, our interpretability features were evaluated based on participants' ability to correctly (dis)agree with a system's prediction. In this respect, participants played the role of an "end user" who is leveraging a system's predictions to perform a task (e.g., find content that is relevant to a topic). Interpretability features might improve performance for some tasks but not others. For example, they might improve a person's ability to guess what a system might do in a specific scenario but not improve a person's ability to determine if a prediction is (in)correct. Our RQ1 results suggest that our evaluation task is worth considering in future studies, perhaps in conjunction with other evaluation tasks.

6.2 RQ2: Effects of Perceptions

Our two interpretability features significantly improved participants' perceptions of the task and the system. Specifically, in the Conf+Sent condition, participants reported greater perceptions of confidence, understandability, and usability.

In some studies, system explanations have *not* increased people's confidence in their decisions. For example, Panigutti et al. [21] conducted a study in which clinicians diagnosed patients using information from previous hospital visits. Participants were assisted by an AI system that provided its own diagnosis (with and without explanations). Here, system explanations did not increase clinicians' confidence in their own diagnoses. The extent to which interpretability features increase confidence may be *conditioned* on the seriousness of the decision task. Diagnosing a patient is a more serious task than deciding whether an article belongs to a topic. Future work is needed to understand whether the seriousness of the decision task *moderates* the influence of interpretability features on people's self-confidence.

6.3 RQ3: Effects on the Scrutiny Process

Our qualitative analysis of participants' comments during the exit interview found that they engaged in three main activities: reading, learning, and decision making. In retrospect, this is not surprising—these are sensible activities involved in judging whether a document belongs to a topic. Interestingly, however, our RQ3 results found that our interpretability features influenced how participants engaged in *all three processes*.

In terms of reading, our interpretability features influenced participants' decisions about: (1) how to read an article (e.g., skimming high confidence cases and closely reading low confidence cases) and (2) which parts of the article to read (e.g., highlighted sentences).

In terms of learning, our interpretability features influenced how participants learned about the domain. An ML model can be viewed as a function f that maps a document $x \in X$ to a category $y \in Y$. One might argue that the goal of interpretable ML is to help humans understand f (i.e., opening the "black box") [2]. Our RQ3 results suggest that interpretable ML systems must *also* support people in learning about x and y. In terms of learning about x, our participants leveraged the sentence highlighting feature to: (1) decide which unknown terms to learn about and (2) determine whether topically relevant terms appeared in meaningful contexts within the abstract. For example, participants disagreed with the system when topically relevant terms appeared in mentions of "future research directions" (i.e., not the focus of the article). In terms of learning about y, participants reported using the highlighted sentences to learn about topically relevant keywords that they searched in subsequent cases for the same topic. Indeed, to effectively scrutinize a system's predictions, users must fully understand the target category's definition and its scope (i.e., inclusion/exclusion criteria). Schoeffer et al. [25] also found that interpretability features helped participants learn about the target category.

Finally, our interpretability features influenced participants' decision-making process in several ways. First, participants reported agreeing with the system when they noticed topically relevant keywords in the highlighted sentences and disagreeing with the system otherwise. Second, some participants reported making their own decisions and using our interpretability features as a "sanity check" (i.e., to verify their own judgement). Finally, in cases where they were highly unsure, participants reported relying on our interpretability features to make a guess (e.g., agreeing with the system for high confidence cases with many highlighted sentences).

Importantly, our RQ3 results also point to ways in which our interpretability features introduced challenges or failed to alleviate challenges inherent in the task participants were asked to perform. We discuss these challenges in terms of: (1) issues related to participants' mental models of the system, (2) issues related to participants relying too heavily on our interpretability features, and (3) activities not supported by our interpretability features.

Mental Models: Our results point to several ways in which participants employed mental models that were problematic. First, some participants had difficulty interpreting our confidence value feature. For example, some participants interpreted a 50% confidence as being "somewhat confident" instead of "highly unsure". Second, some participants ignored the sentence highlighting feature in cases where the system had low confidence. This is like saying: "If the system is unsure, then the sentence highlighting feature is unreliable." Of course, this is not necessarily the case. Even in borderline cases, the highlighted sentences might still represent the most topically relevant portions of the abstract. Finally, some participants expected the highlighted sentences to contain *many* topically relevant keywords. However, those familiar with machine learning know that positive predictions can also include cases with only a few *highly* relevant keywords. Prior work has also observed that people often have mental models of ML systems that are inaccurate or incomplete [26, 28]. In future work, these issues could be potentially addressed through interface features that explain how an interpretability features should be used in specific scenarios (e.g., "While the system is unsure, these highlighted sentences still represent the most topically relevant portions of the abstract.").

Overreliance: In some cases, participants made mistakes because they relied too heavily in our interpretability features. In some cases, participants blindly agreed with high confidence cases and blindly disagreed with low confidence cases. Similarly, some participants made mistakes by completely ignoring unhighlighted sentences that contained important information. Future work should consider research on detecting when people are relying too heavily on specific features of the system.

Additionally, our results suggest that participants trusted the system more often than not, a trend also observed in previous work [1, 10, 14]. In our study, participants should have agreed with the system 50% of the time but agreed with the system about 60% of the time regardless of the interface condition. During the study, participants were unaware that half the cases were false positives. This was done intentionally to simulate a scenario in which end users do not know about a system's classification accuracy. Future research should systematically investigate the effect of knowing a system's accuracy on users' tendencies to agree with the system. A specific condition was explored in Lai et al. [14].

Unsupported Activities: Our RQ3 results also found several potential activities that could be supported by future tools. First, textual documents have implicit internal structures. For example, medical research abstracts often include mentions of: (1) the research questions investigated, (2) methods employed, (3) results and implications, and (4) future research directions. Participants sometimes leveraged this implicit structure when deciding to agree or disagree with the system. One could imagine future systems that make this implicit structure more explicit. For example, if the target topic is *drug effects*, users may benefit from knowing which portions of the abstract discuss *research results*. Alternatively, if the target topic is *therapy*, users may benefit from knowing which portions of the abstract discuss *methods employed*.

Second, participants reported struggling with topics that are highly nuanced and not associated with a clearly defined vocabulary (e.g., drug effects). Future research should investigate and support the strategies employed by users for such topics.

Third, participants often searched the web for information about unknown terms in the abstract and the target topic. Our interpretability features did little to support this learning process. Scrutable systems should make term definitions highly accessible, especially for technical terms that are predictive of a target category. Additionally, to support learning about a target topic, systems should enable users to see highly confident cases and borderline cases for that topic. That is, users may want to learn about a specific topic by comparing highly confident cases against borderline cases. A similar idea was explored in Kim et al. [11].

Finally, as might be expected, our results found that participants struggled with borderline cases. Highly confident cases are easier because they contain positive evidence (e.g., highlighted sentences) that can be scrutinized. Borderline cases are more challenging because they lack positive evidence. This suggests an important direction for future work. One could imagine systems that display positive evidence that is *missing* from the instance. Importantly, this missing evidence should be relevant to the instance and the target topic. Such systems might help users understand *why* a case is borderline: "This article is a borderline case because it is missing words X, Y, and Z."

7 CONCLUSION

We reported on a lab study that investigated the influences of two interpretability features (i.e., confidence value and sentence highlighting) on participants' performance, perceptions, and behaviors. Our interpretability features did not improve participants' performance but did improve their certain perceptions of the system and their experience. Our qualitative analysis of participants' responses during an exit interview found that participants engaged in three cognitive activities during the task (i.e., reading, learning, and decision making). Our interpretability features influenced *all three* activities. We uncovered ways in which our interpretability features introduced challenges and led participants to make mistakes. Finally, we discussed ways in which our interpretability features failed to support participants with important processes. For example, participants often needed help with understanding aspects of the abstract (e.g., unknown terms) and the meaning/scope of the predicted topic. We discussed ways in which future tools might better support these processes.

REFERENCES

[1] 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.

[2] David A Broniatowski. 2021. Psychological Foundations of Explainability and Interpretability in Artificial Intelligence. (2021).

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

[4] 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.

[5] Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. 2018. Learning to Explain: An Information-Theoretic Perspective on Model Interpretation. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research)*. PMLR, 883–892.

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

[7] Gunther Eysenbach and Ch Kohler. 2003. What is the prevalence of health-related searches on the World Wide Web? Qualitative and quantitative analysis of search engine queries on the internet. In *AMIA annual symposium proceedings*, Vol. 2003. American Medical Informatics Association, 225.

[8] Peter Hase and Mohit Bansal. 2020. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5540–5552.

[9] Daniel W Hook, Simon J Porter, and Christian Herzog. 2018. Dimensions: building context for search and evaluation. *Frontiers in Research Metrics and Analytics* 3 (2018), 23.

[10] 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.

[11] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. 2016. Examples are not enough, learn to criticize! Criticism for Interpretability. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

[12] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*. PMLR, 1885–1894.

[13] Vivian Lai, Han Liu, and Chenhao Tan. 2020. " Why is' Chicago'deceptive?" Towards Building Model-Driven Tutorials for Humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[14] 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.

[15] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1675–1684.

[16] 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.

[17] Henry J Lowe and G Octo Barnett. 1994. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *JAMA* 271, 14 (1994), 1103–1108.

[18] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[19] W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences* 116, 44 (2019), 22071–22080.

[20] National Library of Medicine. 2021. MEDLINE 2022 Initiative: Transition to Automated Indexing. https://www.nlm.nih.gov/pubs/techbull/nd21/nd21_medline_2022.html. Accessed: 2022-01.

[21] Cecilia Panigutti, Andrea Beretta, Fosca Giannotti, and Dino Pedreschi. 2022. Understanding the Impact of Explanations on Advice-Taking: A User Study for AI-Based Clinical Decision Support Systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 568, 9 pages.

[22] Jiaming Qu, Jaime Arguello, and Yue Wang. 2021. A Study of Explainability Features to Scrutinize Faceted Filtering Results. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 1498–1507.

[23] 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.

[24] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[25] Jakob Schoeffer, Niklas Kuehl, and Yvette Machowski. 2022. "There Is Not Enough Information": On the Effects of Explanations on Perceptions of Informational Fairness and Trustworthiness in Automated Decision-Making *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1616–1628.

[26] Hendrik Schuff, Alon Jacovi, Heike Adel, Yoav Goldberg, and Ngoc Thang Vu. 2022. Human Interpretation of Saliency-Based Explanation Over Text. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 611–636.

[27] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 618–626.

[28] Aaron Springer and Steve Whittaker. 2020. Progressive disclosure: When, why, and how do users want algorithmic transparency information? *ACM Transactions on Interactive Intelligent Systems (TiiS)* 10, 4 (2020), 1–32.

[29] S Wachter, B Mittelstadt, and C Russell. 2018. Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harvard Journal of Law and Technology* 31, 2 (2018), 841–887.

[30] Huaiyu Wan, Yutao Zhang, Jing Zhang, and Jie Tang. 2019. Aminer: Search and mining of academic social networks. *Data Intelligence* 1, 1 (2019), 58–76.

[31] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces*. 318–328.

[32] Hilde JP Weerts, Werner van Ipenburg, and Mykola Pechenizkiy. 2019. A human-grounded evaluation of SHAP for alert processing. *arXiv preprint arXiv:1907.03324* (2019).

[33] Peace Ossom Williamson and Christian IJ Minter. 2019. Exploring PubMed as a reliable resource for scholarly communications services. *Journal of the Medical Library Association: JMLA* 107, 1 (2019), 16.

[34] Linyi Yang, Eoin M Kenny, Tin Lok James Ng, Yi Yang, Barry Smyth, and Ruihai Dong. 2020. Generating plausible counterfactual explanations for deep transformers in financial text classification. *arXiv preprint arXiv:2010.12512* (2020).

[35] Muhammad Rehman Zafar and Naimul Mefraz Khan. 2019. DLIME: A Deterministic Local Interpretable Model-Agnostic Explanations Approach for Computer-Aided Diagnosis Systems. In *ACM SIGKDD Workshop on Explainable AI/ML (XAI) for Accountability, Fairness, and Transparency*. 6.