Why is "Problems" Predictive of Positive Sentiment? A Case Study of Explaining Unintuitive Features in Sentiment Classification

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

Explainable AI (XAI) algorithms aim to help users understand how a machine learning model makes predictions. To this end, many approaches explain which input features are most predictive of a target label. However, such explanations can still be puzzling to users (e.g., in product reviews, the word "problems" is predictive of *positive* sentiment). If left unexplained, puzzling explanations can have negative impacts. Explaining unintuitive associations between an input feature and a target label is an underexplored area in XAI research. We take an initial effort in this direction using unintuitive associations learned by sentiment classifiers as a case study. We propose approaches for (1) automatically detecting associations that can appear unintuitive to users and (2) generating explanations to help users understand why an unintuitive feature is predictive. Results from a crowdsourced study (N = 300) found that our proposed approaches can effectively detect and explain predictive but unintuitive features in sentiment classification.

CCS CONCEPTS

• **Human-centered computing** → **User studies**; **Empirical studies in interaction design**; • **Computing methodologies** → **Machine learning**.

KEYWORDS

Interpretable Machine Learning, User Study, Unintuitive Features

ACM Reference Format:
Jiaming Qu, Jaime Arguello, and Yue Wang. 2024. Why is "Problems" Predictive of Positive Sentiment? A Case Study of Explaining Unintuitive Features in Sentiment Classification. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3630106.3658547

1 INTRODUCTION

Research on explainable artificial intelligence (XAI) has investigated a variety of approaches to explaining the complex behavior of a machine learning model. One simple and straightforward approach is to show which parts of an input (i.e., which features) are most influential. Such explanations are referred to as *feature importance* explanations. Different algorithms and visualizations have been developed to generate feature importance explanations [2, 31, 34, 45, 50]. Such explanations have been empirically shown to improve a user's performance in a variety of tasks, such as AI-assisted decision-making [2, 31].

Despite promising results, studies on feature importance explanations have not tackled an important problem—syntactically simple explanations (e.g., "feature x plays an important role in predicting category y") can still be puzzling or counterintuitive. For example, prior XAI research has found that having asthma lowers the risk of death among pneumonia patients [13]; that the word "Chicago" is a strong indicator of a Chicago hotel review being fake [30]; and that words like "host" and "posting" have a stronger association with Atheism than Christianity in a topical classification task [44, 45]. In these cases, features that are deemed "important" by a model may not immediately make sense to humans. Most XAI approaches that focus on feature importance do not further explain *why* a feature is important.

We use the term **unintuitive features** to describe this phenomenon. Unintuitive features are predictive from a model's perspective but are at odds with human intuition and common sense. An important question is: What makes a feature unintuitive? There are at least two possibilities. First, a predictive feature may be unintuitive because of anomalies in the training data, especially when training data is sparse and the feature is predictive due to overfitting. Second, a predictive feature may be unintuitive because it represents an *underlying* phenomenon that is not obvious to a human by simply seeing an explanation such as "feature x is predictive of category y". For example, within the context of automotive product reviews, the word "fit" (a seemingly positive word) is predictive of negative sentiment. At first, this seems paradoxical. It even seems that the classifier learned an incorrect association. However, this is not the case. The word "fit" predicts negative sentiment because people tend to use it when the product did not "fit". Conversely, people do not use "fit" in positive reviews because a product "fitting" is a minimum requirement unworthy of mentioning in a positive review. In our research, we focus on the second category and not the first. That is, we focus on features that are: (1) predictive from a model's perspective, (2) generalizable to test data, and (3) unintuitive to a human.

Prior work has mostly focused on algorithms that can translate a complex model's predictive behavior into syntactically simple forms [34, 45, 46, 49, 50], which is a necessary but not sufficient condition for algorithm-generated explanations to make sense to

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658547

humans. Most studies that have evaluated feature importance explanations have assumed that syntactically simple explanations are self-explanatory. Lai et al. [30] touched upon this issue to some extent (e.g., using manually curated rules to further explain why words like "Chicago" are predictive of a Chicago hotel review being fake). However, they did not provide an algorithmic solution for generating further explanations.

Explaining unintuitive features is an important problem in XAI research. Prior studies have found that unintuitive explanations can make users lose trust in a machine learning model [10, 14, 38, 42]. Additionally, if left unexplained, users may hypothesize wrong reasons for why an unintuitive feature is predictive. Schuff et al. [47] randomly highlighted words within reviews as being predictive of a sentiment. Results found that participants made up their own incorrect explanations for why those words were predictive. Thus, explaining unintuitive features may improve users' trust in a model and help them learn about the predictive task.

In this paper, we take initial steps toward addressing the issue of unintuitive features in XAI. As a case study, we focus on a sentiment classification task—predicting whether an Amazon product review is positive or negative. It is a task that can be performed by ordinary crowdworkers and machine learning models can perform reasonably well on, and therefore unintuitive features are not due to overfitting. We identified words that are predictive of a specific sentiment but likely to be perceived as unintuitive or paradoxical to a human. For example, the word "problems" (a seemingly negative word) is predictive of *positive* sentiment and the word "fit" (a seemingly positive word) is predictive of *negative* sentiment. We report on a crowdsourced user study (N = 300) that evaluated different tools designed to explain the predictiveness of an unintuitive feature. Participants were assigned to one of six interface conditions (a between-subjects design). Interface conditions varied based on the tools available to participants. The study investigated three research questions, which considered the effects of the interface condition on different types of dependent variables:

- **RQ1:** How does the interface condition affect participants' *understanding* of an unintuitive feature's predictiveness?
- **RQ2:** How does the interface condition affect participants' *perceptions* of the provided tools and their experiences?
- **RQ3:** How does the interface condition affect participants' *behaviors* during different tasks?

The study proceeded in two phases. Phase 1 of the study validated our assumption that the level of (un)intuitiveness of a predictive feature in sentiment classification can be computationally estimated. During Phase 1, participants were shown a batch of predictive features (i.e., words) and asked to judge which sentiment they expected the word to convey: positive, negative, or "not sure". These judgments were found to strongly correlate with those made by a large language model. Phase 2 of the study investigated the above three research questions. During Phase 2, participants completed four trials. During each trial, participants were shown a predictive but unintuitive word and asked to complete different judgments and tasks using the tools available in their assigned interface condition. We explored three different tools: (1) a visualization of the sentiment label distribution among training instances containing the word, (2) training examples of either sentiment containing the word, and (3) contextual patterns minded from training examples of either sentiment containing the word.

Our results found that participants had the best outcomes when provided with a combination of tools (data distribution + examples or contextual patterns). When provided with *only* the data distribution tool, participants were able to correctly judge the sentiment of the unintuitive features. However, they did not perceive the tool to be understandable, helpful, nor trustworthy. Seeing concrete examples and contextual patterns helped participants explain *why* an unintuitive feature is predictive. Through the case study, our paper contributes practical tools and design implications for explaining predictive yet unintuitive features, an important but underexplored area in XAI research.

2 RELATED WORK

Our research builds upon three areas of prior work: (1) technical approaches for explaining a model's predictions, particularly feature importance approaches, (2) prior studies where machine-generated explanations were found to be unintuitive or puzzling, and (3) empirical studies that evaluated XAI systems through user experiments.

Feature Importance Explanations: Explainable artificial intelligence (XAI) research has explored a wide range of approaches to help people understand a machine learning model's predictions. For example, to explain a prediction for a specific instance, approaches can highlight which features of the instance are the most indicative of the predicted label [34, 45, 46, 50], which training instances are the most influential in teaching the model to predict the label for this instance [27, 40, 53], and which training instances are most similar to the target instance and have the same ground truth label as the prediction [9, 52]. Among these approaches, feature importance (or feature attribution) explanations are highly popular. Such explanations highlight which parts of the input (e.g., words, sentences, superpixels) are most indicative of the predicted label [34, 45, 46, 50].

In XAI research, explanations can be categorized as either global or local. Global explanations provide insights about the overall behavior of the model, while local explanations elucidate a model's predictions on individual instances [17]. This distinction also applies to feature importance explanations. Global feature importance explanations show a feature's overall impact in the model's prediction logic, which is typically learned from the entire training data. These explanations highlight which features have the most influence on the model's predictions across a wide range of inputs [16]. Some machine learning models have mechanisms that can be leveraged for showing global feature importance, such as the coefficients from a support vector machine model [30, 31] or logistic regression model [7, 12]. In our study, we trained logistic regression classifiers and used regression coefficients to identify predictive (i.e., important) features. Compared to global explanations, local feature importance explanations are more intricate because they explain the importance of features for a specific instance. These explanations demonstrate why a model made a certain prediction for a given input. Prior research has developed different algorithms to compute feature importance locally, e.g., LIME [45], Integrated

Gradients [50], and SHAP [34]. Despite their differences, global and local feature importance methods share the same syntax: "feature x plays an important role in predicting label y for a specific instance or any label $y \in Y$ across instances" and assume that users can understand a model's prediction based on certain important features. However, studies have found that this assumption is not always true. That is, users are sometimes confused after learning which features play an important role. We discuss such studies below.

The Phenomenon of Unintuitive Explanations: Machine-generated explanations are intended to help users understand a model's behavior. However, such explanations are not guaranteed to always make sense to users. The phenomenon of unintuitive explanations is not rare in previous empirical studies. For example, Qu et al. [42] conducted a study in which participants scrutinized a document's machine-predicted categories by inspecting the most influential sentences, highlighted by the system. Participants commented on ignoring such explanations when they could not understand why the sentences were important. Unclear explanations can also have undesirable consequences for domain experts. In a study where pathologists completed a diagnostic task assisted by example-based explanations, participants exhibited confusion and self-doubt when they did not understand or agree with the explanations [9].

Prior studies have also observed unintuitive feature importance explanations. For example, studies have reported unexpected regression coefficients in analyses related to econometrics [18, 26], psychology [25], and education [35]. More recently, XAI research has reported predictive but unintuitive features in the medical domain (e.g., patients with pneumonia who have a history of asthma have a lower risk of death) [13] and text analysis domain (e.g., the word "Chicago" is a strong predictor of a review being fake) [30]. One solution to addressing unintuitive features is to consider feature interactions instead of single features [4, 22, 51]. However, feature interaction explanations can only resolve simple unintuitive cases such as negation because they focus on interactions between *pairs* of features [4, 22]. Text analysis often involves predicting complex phenomena (e.g., topic, sentiment, and intent). Such phenomena are abstract and can be influenced by patterns that go beyond pairwise interactions between words. This inspires us to develop tools to explain unintuitive text features in more complex situations by mining semantic patterns from the training data.

Empirical Studies in XAI: Evaluating XAI systems with human end-users gives direct evidence on the effectiveness of explanations in a concrete task scenario [17, 55]. To this end, numerous empirical studies have been conducted to investigate human-XAI interaction using a variety of tasks across different domains such as sentiment analysis [2], topic categorization [42], deceptive review detection [30], disease diagnosis [10], and toxicity detection [12]. To gauge the effects of XAI systems on human end-users, previous research conducted both quantitative and qualitative evaluations. Quantitative evaluations often measured end-users' (1) performance (e.g., decision accuracy [2, 10, 12, 30, 42]), (2) perceptions (e.g., confidence [10, 15], understanding [9, 42] and trust [10, 30]), and (3) behaviors (e.g., time spent on task [10, 12, 42]). Besides quantitative evaluations, prior studies have also used qualitative techniques to gain deeper insights into human-XAI interaction, such as conducting exit interviews with participants [10, 15, 42] or using a think-aloud protocol [6, 14, 38]. In our study, we collected both quantitative and qualitative data to investigate whether our developed tools helped participants understand the predictiveness of an unintuitive feature.

3 METHODS

3.1 Study Overview

To investigate RQ1-RQ3, we conducted a crowdsourced study using the Prolific platform. The study involved 300 participants (M = 116, F = 183, $Unreported$ = 1). Participants' ages ranged from 18 to 63 ($Mean$ = 35.41, $S.D.$ = 11.92). We restricted our study to English-speaking Prolific workers from USA, UK, and Canada who had completed at least 100 tasks with an acceptance rate ≥ 95% and had experience in online shopping and review writing. The study involved two phases: Phase 1 and Phase 2.

During Phase 1, participants were shown a list of 10 words. For each word, participants were asked to indicate which sentiment they expected the word to convey: "positive", "negative" or "not sure". Phase 1 used the same interface for all participants. Our goal for Phase 1 was to investigate whether an LLM-based zero-shot classifier can automatically estimate the (un)intuitiveness of a word that is predictive of positive or negative sentiment in product reviews. Each word was redundantly classified by five participants. This enabled us to compare the level of disagreement among participants with the level of (un)intuitiveness estimated using the LLM-based zero-shot classifier.

During Phase 2, participants completed four trials. During each trial, participants were shown a predictive yet unintuitive feature and were asked to complete several tasks and answer several questions (e.g., judge whether the feature is predictive of positive or negative sentiment). While Phase 1 used the same interface for all participants, Phase 2 involved an interface manipulation. Participants were assigned to 1 of 6 interface conditions (50 participants per condition). Interface conditions varied based on the combination of different tools that we designed to explain the predictiveness of an unintuitive feature (Section 3.5). After completing all four trials, participants completed a questionnaire that asked about their perceptions of the system and their experiences (Section 3.7). Participants were given US$ 6.00 for participation. The study was approved by our Institutional Review Board (IRB).

3.2 Identifying and Explaining Unintuitive Features in Sentiment Classification

Dataset and Models: The data used in our study originated from Ni et al. [37], which consists of Amazon product reviews of various product categories. We selected five product categories: Automotive, Electronics, Pet Supplies, Home and Kitchen, and Sports and Outdoors. In the original dataset, each review had a 5-star rating. We used reviews with 1 star as negative and 5 stars as positive. For each product category, we trained a logistic regression classifier using a balanced dataset of 200,000 reviews. All classifiers used a unigram TF-IDF representation with stopwords removed [3, 39]. Each classifier was tested on a balanced dataset of 10,000 reviews from the same product category. All classifiers achieved an F1 score ≥ 0.90.

Identifying Predictive Features: Unigram feature importance explanations are widely used in prior XAI research [28, 30, 36]. In this study, we identified predictive features in the above logistic regression classifiers. For each product category, we selected the 200 words with the highest coefficients as the most predictive of positive sentiment (denoted as $\mathcal{S}^+$) and the 200 words with the lowest coefficients as the most predictive of negative sentiment (denoted as $\mathcal{S}^-$). Words shown to participants during Phase 1 and Phase 2 were sampled from these sets.

Estimating the (Un)intuitiveness of a Predictive Feature: Predictive features can have different levels of (un)intuitiveness. For example, it is obvious why "great" is predictive of positive sentiment and "terrible" is predictive of negative sentiment. However, it is unclear why "problems" (a seemingly negative word) is predictive of *positive* sentiment and "fit" (a seemingly positive word) is predictive of *negative* sentiment. One approach to estimating the (un)intuitiveness of a word-sentiment relation would be through human assessment. However, this requires significant manual effort. Instead, we leveraged a large language model (LLM) to estimate whether a word-sentiment relation might be perceived as (un)intuitive to humans.

The basic idea is to use an LLM to approximate a human's perception that a word w conveys a sentiment y. We used an LLM called BART [33]. BART can predict the probability that one piece of text logically entails another. In this respect, it can be used as a classifier when one piece of text is the input and the other is the textual description of a candidate label (e.g., positive or negative sentiment). In this setup, the LLM is used as a "zero-shot classifier" because it does not require training data [54]. To estimate the (un)intuitiveness of a word-sentiment relation, we used the prompt "`In Amazon reviews of [CATEGORY] products, word` w `is` y" and asked the zero-shot classifier to predict the probabilities of class labels y = ["positive", "negative"] for word w.[1] We use $P_z(y|w)$ to denote the probability that word w conveys sentiment y according to the zero-shot classifier. A large value of $P_z(y = pos|w)$ suggests that w is intuitively predictive of positive sentiment. A large value of $P_z(y = neg|w) = 1 - P_z(y = pos|w)$ suggests that w is intuitively predictive of negative sentiment. Values close to 0.5 suggest that w is not intuitively associated with either positive or negative sentiment. By adopting this approach, we assume that the zero-shot classifier approximates a human's perception that word w conveys sentiment y. Phase 1 of our user study validated this assumption (Section 4.1).

Explaining Unintuitive Features: We developed three different tools to explain the relation between a word w and a sentiment y. Our tools explained a word's association with *both* sentiment labels (i.e., positive and negative). During Phase 2, participants were asked to scrutinize both associations and choose the one that made the most sense to them.

The first tool, **Distribution**, was designed to show posterior probabilities $P_D(y = pos|w)$ and $P_D(y = neg|w)$, inspired by prior work [28]. These probabilities were estimated based on the proportion of positive and negative reviews containing word w in the training data D. We used a pie chart to visualize these probabilities.

[1]We used the implementation in the HuggingFace library: https://huggingface.co/tasks/zero-shot-classification.

The second tool, **Example**, was designed to show training examples where w and y co-occur, another common approach in prior work [30]. We sampled 25 positive and 25 negative reviews containing w from the training data.

The third tool, **Pattern**, was designed to show *contextual patterns* where w and y co-occur. Contextual patterns are common phrases of variable lengths that appear in the training data. For example, the word "minutes" has positive contextual patterns such as "5 minutes to install" and negative contextual patterns such as "broke within minutes". To show contextual patterns for word w and sentiment y, we developed a *contextual pattern mining* algorithm, a novel contribution of this work. Given word w and sentiment y, the algorithm identifies common and diverse patterns (i.e., phrases) p in the training data that satisfy two conditions: (1) pattern p contains word w and (2) $P_z(y|p)$ is large (i.e., phase p clearly predicts sentiment y according to the zero-shot classifier). We describe the algorithm below.

Given word w and sentiment y, the first step is to find candidate contextual patterns that include w and are predictive of y. To this end, we first iterate over all training instances associated with sentiment y that contain word w. For each instance, we consider phrases of increasing length by adding words to the left and right of w. We limit ourselves to phrases no longer than five words to the left and right of w. For each newly-generated phrase, we use the zero-shot classifier to estimate $P_z(y|p)$, the probability that the phrase p is predictive of sentiment y. The shortest phrase p that gives $P_z(y|p) > 0.8$ (if any) is then considered a *candidate* contextual pattern for w and y. The next step is to select a small set of candidate contextual patterns to show for w and y. Inspired by the maximal marginal relevance (MMR) algorithm [11], we select the most frequent pattern first. Then, we iteratively select patterns that are both frequent and semantically dissimilar to previously selected patterns. To measure semantic similarity, we used cosine similarity between phrase embeddings computed by a transformer-based encoder [43]. This selection process ensured that participants were exposed to contextual patterns that are both predictive of y and diverse. Table 1 shows example positive and negative contextual patterns mined for words that were estimated as unintuitive by the zero-shot classifier.

3.3 Experimental Design

In this section, we describe how words were sampled for Phase 1 and Phase 2. Our goal of Phase 1 was to validate the use of a zero-shot classifier to estimate the (un)intuitiveness of a feature that is predictive of positive or negative sentiment. During Phase 1, we wanted to expose participants to predictive words with different levels of (un)intuitiveness. For each product category, we used the zero-shot classifier to estimate $P_z(y = pos|w)$ for every word in sets $\mathcal{S}^+$ and $\mathcal{S}^-$ (i.e., predictive of positive and negative according to the logistic regression model). We only estimated $P_z(y = pos|w)$ since $P_z(y = neg|w) = 1 - P_z(y = pos|w)$. Then, we sampled 120 words associated with different levels of (un)intuitiveness using stratified sampling. These 120 words were organized into 12 batches of 10 words each. Each batch was redundantly judged by five participants. These redundant judgements were used to estimate $P_u(y|w)$, the probability that word w predicts sentiment y according to human

Table 1: Examples of contextual patterns mined for unintuitive words. Words like "problems" and "minutes" are estimated to be unintuitively positive, while words like "star" and "money" are estimated to be unintuitively negative.

Word	Positive contextual patterns	Negative contextual patterns
problems	no more problems, without any problems	problems with, problems since
minutes	5 minutes to install, installs in minutes	broke within minutes, didn't last 5 minutes
star	star rating, 5 star, star product	one star, half star, negative star
money	worth the money, good value for money	waste of money, want my money back

users. As described in Section 4.1, a significant correlation between $P_u(y|w)$ and $P_z(y|w)$ validated our use of a zero-shot classifier to estimate the (un)intuitiveness of a predictive feature.

During Phase 2, our goal was to investigate how different tools can help people understand the predictiveness of an unintuitive feature. For each product category, we sampled 40 words from sets $\mathcal{S}^+$ and $\mathcal{S}^-$ as follows. First, we excluded words sampled for Phase 1. Then, we sampled words that met the following criteria. First, we included words with $P_z(y = pos|w) < 0.2$ from set $\mathcal{S}^+$. These are words that are *paradoxically* predictive of positive (e.g., the word "problems" being predictive of positive). Second, we included words with $P_z(y = pos|w) > 0.8$ from set $\mathcal{S}^-$. These are words that are *paradoxically* predictive of negative (e.g., the word "fit" being predictive of negative). Third, we included words with $0.2 \leq P_z(y = pos|w) \leq 0.8$ from sets $\mathcal{S}^+$ and $\mathcal{S}^-$. These are words that are unintuitive regardless of which sentiment they are predictive of. Finally, during Phase 2, participants were asked whether an AI system should consider the word as strong predictive evidence of sentiment. Therefore, we only included words that resulted in a statistically significant drop in performance if omitted from a logistic regression model. For each product category, each sample of 40 words was organized into 10 batches of 4 words each (2 from $\mathcal{S}^+$ and 2 from $\mathcal{S}^-$). Each batch was completed by six participants, each in a different interface condition. Sixty participants were assigned to each product category.

3.4 Study Protocol

The study protocol proceeded as follows. First, participants watched an overview video of the study. Then, participants completed Phase 1 of the study as follows. After watching an instructional video, participants were presented with a batch of 10 words and were instructed to judge the sentiment conveyed by each word. Participants were presented with the following prompt: "Think about Amazon reviews for [Category] products. Which sentiment is each word more likely to convey?" Participants were instructed to select the sentiment of each word based solely on their intuition. Participants were presented with the options of "positive, "negative", and "not sure".

Next, participants proceeded to Phase 2 of the study. During Phase 2, participants completed four trials. During each trial, participants were exposed to an unintuitive feature and were asked to complete a series of judgments. While Phase 1 used the same interface for all participants, Phase 2 involved an interface manipulation. Participants were assigned to one of six interface conditions (a between-subjects design). Interface conditions varied based on the tools available to participants. For each word, participants were asked to complete a series of judgments. First, participants were asked to judge the sentiment of the word. Participants used a range slider to indicate their perceived sentiment from very negative to very positive. The range slider did not have a midpoint. Therefore, participants had to choose between positive or negative. However, they could choose values close to the midpoint if they were unsure. Second, participants were prompted to list scenarios in which the word might be used to convey the selected sentiment. Participants were instructed that "scenarios can be phrases, sentences, or explanations based on your personal understanding." Participants were provided with a textbox to list scenarios as a bulleted list. Third, participants were asked to rate their confidence in their responses to the first two tasks on a 5-point scale. Next, participants were asked whether the AI system should consider the word as *strong* evidence of the sentiment they selected. Participants were asked to respond "yes" or "no" using radio buttons. Given that all words sampled for Phase 2 were highly predictive, resulting in statistically significant drops in performance if omitted, the correct answer was always "yes". However, participants did not know this. Then, participants were asked to rate their confidence in their response to the above question on a 5-point scale. After all four Phase 2 trials, participants were presented with a side-by-side comparison between the AI system's judgment of each word (based on regression coefficients) and their own judgment. Finally, participants completed a post-task questionnaire about their perceptions of the interface and the task. Our study materials and system demos are available online.

3.5 Phase 2 Interface Conditions

In Phase 2, participants were randomly assigned to one of six interface conditions (i.e., a between-subjects design). Interface conditions varied based on the tools available to participants. Participants answered the same questions in all interface conditions. Figure 1 describes the layout of the interface (A) and our three tools (B-D).

Baseline: In this condition, no tools were provided. Participants made judgments based solely on their intuition.

Example (Figure 1-B): In this condition, we randomly sampled 25 positive and 25 negative reviews from the training data containing the word to be judged.

Pattern (Figure 1-C): In this condition, we used the *contextual pattern mining* algorithm (Section 3.2) to display positive and negative contextual patterns associated with the word to be judged. We displayed up to five contextual patterns per sentiment and provided three sampled reviews per pattern-sentiment pair.

Distribution (Figure 1-D): In this condition, we provided the sentiment label distribution of training instances containing the word to be judged.

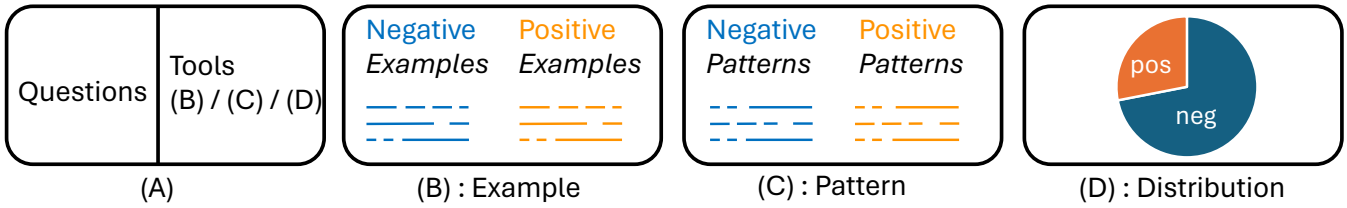

Figure 1: Phase 2 interface design. Questions and tools (if any) were displayed side-by-side (A). Visual representations of our tools are shown in subfigures B-D.

Example+Distribution: In this condition, participants had access to both lists of reviews and label distribution (i.e., both (B) and (D) in Figure 1). Compared to the Example condition, this condition also displayed the number of positive and negative reviews containing the word. Lists of reviews were shown by default, and participants could use radio buttons to switch between tools.

Pattern+Distribution: In this condition, participants had access to both contextual patterns and label distribution (i.e., both (C) and (D) in Figure 1). Compared to the Pattern condition, this condition also displayed the number of reviews associated with each contextual pattern. Contextual patterns were shown by default, and participants could use radio buttons to switch between tools.

3.6 Measures of Understanding (RQ1)

In RQ1, we investigated the effects of the interface condition on participants' understanding of a predictive but unintuitive feature. We measured participants' understanding from three perspectives.

Correctness of Sentiment Judgment: This binary measure considered whether a participant's judgment of a word's sentiment (i.e., positive vs. negative) aligned with the logistic regression classifier.

Correctness of Feature Consideration: This binary measure considered: (1) whether the participant's judgment of the word's sentiment (i.e., positive vs. negative) aligned with the logistic regression classifier and (2) whether the participant correctly indicated that the AI system should consider the word as strong evidence. Based on an ablation analysis, the correct answer for words sampled for Phase 2 was always "yes" (i.e., the AI system should consider the word as strong evidence). However, participants did not know this.

Correctness of Listed Scenarios: This measure considered the *proportion* of valid scenarios that participants listed to support their selected sentiment. To this end, we conducted a qualitative analysis of scenarios listed by participants. Each scenario was classified as "valid" if it met the following three criteria: (1) the scenario is relevant to the context of product reviews; (2) the scenario includes the word being judged; and (3) the scenario is relevant to the sentiment selected by the participant. Our qualitative analysis of scenarios involved developing a coding guide. After developing an initial coding guide, three of the authors annotated 30 lists of scenarios. Then, the authors discussed disagreements and refined the coding guide. Finally, to validate the coding guide, the same three authors annotated 30 new lists of scenarios. The Fleiss' Kappa agreement was $\kappa = 0.5816$, which is considered moderate agreement [32]. Finally, one author coded all remaining scenarios. In total, participants listed 3,447 scenarios.

3.7 Measures of Perceptions (RQ2)

In RQ2, we investigated the effects of the interface condition on participants' perceptions of the interface and their experiences. Our first two measures are referred to as **Confidence in Sentiment Judgment** and **Confidence in Feature Consideration**. The first measure corresponds to the participant's confidence in judging the sentiment of a word and listing scenarios in which the word might be used to convey the selected sentiment. The second measure corresponds to the participant's confidence in deciding whether the AI system should consider the word as strong predictive evidence. In both cases, participants rated their confidence on a 5-point scale ranging from (1) "not at all confident" to (5) "extremely confident".

After all four trials in Phase 2, participants completed a post-task questionnaire. Participants responded to agreement statements on a 7-point scale ranging from (1) "strongly disagree" to (7) "strong agree". The first part of the questionnaire asked four questions about: (1) **agreement** with the AI system's judgement of all four words, (2) **understanding** of the AI system's judgements and explanations, (3) **helpfulness** of the tools provided, and (4) **trust** in the AI system's effectiveness in predicting sentiment. The second part of the questionnaire asked about **system usability**. We used the 10-item System Usability Scale (SUS) [5]. Responses to all 10 items had high internal consistency (Cronbach's $\alpha = 0.91$). Therefore, we averaged responses to these 10 items to form one system usability measure. The third part of the questionnaire asked about **workload**. We used the 6-item NASA-TLX [20]. Responses to these items had low internal consistency (Cronbach's $\alpha = 0.45$). Therefore, we analyzed responses to these 6 items individually.

3.8 Measures of Behaviors (RQ3)

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

Intensity in Sentiment Judgment: This binary measure considered whether a participant made an extreme positive or negative judgment by choosing either the rightmost or leftmost positions on the range slider.

Time Interval (sentiment judgment): This measure considered the amount of time (in seconds) participants took to judge the sentiment of a word.

Time Interval (all questions): This measure considered the total amount of time (in seconds) participants took to answer all questions related to a word.

3.9 Statistical Analysis

RQ1-RQ3 considered the effects of the interface condition on different types of outcomes. To test for statistically significant effects, we fit linear regression models for real-valued measures and logistic regression models for binary measures. Additionally, some measures involved four values originating from the four Phase 2 trials. For such measures, we used multi-level modeling and added the participant ID as a random factor. In all models, we compared interface conditions against the Baseline condition to study the effects of providing *any* tools for explaining unintuitive features versus providing *none*. Non-Baseline conditions were included in each model as indicator variables.

4 RESULTS

4.1 Validating the Use of an LLM-based Zero-shot Classifier to Estimate Feature (Un)intuitiveness

In our study, we leveraged an LLM-based zero-shot classifier to estimate whether a predictive feature is perceived as (un)intuitive to a human. This approach assumes that the LLM can approximate a human's intuition about the relation between a word and a sentiment. To validate this assumption, we compared $P_z(y = pos|w)$ and $P_u(y = pos|w)$ across all 600 unique words judged during Phase 1. $P_z(y = pos|w)$ denotes the probability that word w predicts positive according to the zero-shot classifier. $P_u(y = pos|w)$ denotes the probability that word w predicts positive according to our participants. During Phase 1, each word w was judged by five redundant participants using the options of "positive", "negative", or "not sure". We estimated $P_u(y = pos|w)$ using the formula: $P_u(y = pos|w) = \frac{1 \cdot n_{pos} + 0.5 \cdot n_{ns} + 0 \cdot n_{neg}}{5}$. We use n_{pos}, n_{neg}, and n_{ns} to denote the number of participants who selected "positive", "negative" and "not sure" for word w, respectively. Essentially, this formula does a weighted aggregation of participants' judgments. Then, we computed the Pearson correlation coefficient (ρ) between $P_z(y = pos|w)$ and $P_u(y = pos|w)$. The result showed a significant correlation ($\rho = 0.9125, p < .001$). This high and significant correlation suggests that an LLM-based zero-shot classifier can estimate the extent to which a word-sentiment relation will be perceived as (un)intuitive to a human. More broadly, it suggests the possibility of using an LLM as a "surrogate average user" to automatically detect situations where XAI outputs are not self-explanatory and further explanations are warranted.

4.2 RQ1: Understanding

In RQ1, we investigated the effects of different interface conditions on participants' understanding of the predictiveness of an unintuitive feature. Figure 2 shows our RQ1 results. Our results found three main trends.

First, in the Baseline condition, participants achieved 50% accuracy in terms of "correctness of sentiment judgement" and 25% accuracy in terms of "correctness of feature consideration". In both cases, performance was not better than random guessing. Without our tools, only half of participants judged a word's sentiment correctly. Of these, only half correctly indicated that the model should consider the feature as strong predictive evidence. This result confirms that features selected for Phase 2 indeed appeared unintuitive to participants. As a result, in the Baseline condition, their judgments approximated random guessing.

Second, all interface conditions, except the Example condition, had significant effects on participants making more correct sentiment judgments. All interface conditions, except the Pattern condition, had significant effects on participants making more correct feature consideration judgments. However, when participants had access to the label distribution in the other three interface conditions, they made significantly more correct judgments in both questions. Compared to providing a single tool, providing a combination of tools (i.e., the Example+Distribution and Pattern+Distribution conditions) was the best approach.

Third, compared to the Baseline condition, participants having access to concrete explanations in the Example and Pattern conditions were significantly more likely to list valid scenarios. Conversely, participants had difficulty listing valid scenarios in the Distribution condition, where they were only shown the label distribution. Moreover, showing grouped examples had greater effects than ungrouped examples—the Pattern+Distribution condition significantly increased the chance of participants listing valid scenarios but the Example+Distribution condition did not.

4.3 RQ2: Perceptions

In RQ2, we investigated the effects of different interface conditions on participants' perceptions of the provided tools and their experiences. Figure 3 shows our RQ2 results. The interface condition did not have significant effects on any workload measures. Thus, the corresponding plots are omitted. Our results found four main trends.

First, none of our interface conditions helped participants achieve higher confidence in sentiment judgment and feature consideration in general. Even though participants were able to make more correct judgments in certain interface conditions, their confidence was close to the midpoint (i.e., moderately confident) in all cases. This result implies that our tools did not significantly increase participants' confidence when making these judgments.

Second, compared to the Baseline condition, the Distribution condition did not significantly improve participants' perceptions. Interestingly, participants made significantly more correct judgments in the Distribution condition (RQ1 results). This contrast suggests that while the label distribution could persuade participants to make objectively correct judgments, it was not subjectively perceived as understandable, helpful, or trustworthy.

Third, the Example+Distribution and Pattern+Distribution conditions significantly improved participants' perceptions across all measures except the confidence measure. These conditions with two tools might provide a more comprehensive view of word-sentiment relations than conditions with only one tool. It also demonstrates the necessity of providing more concrete explanations in addition to only showing the label distribution.

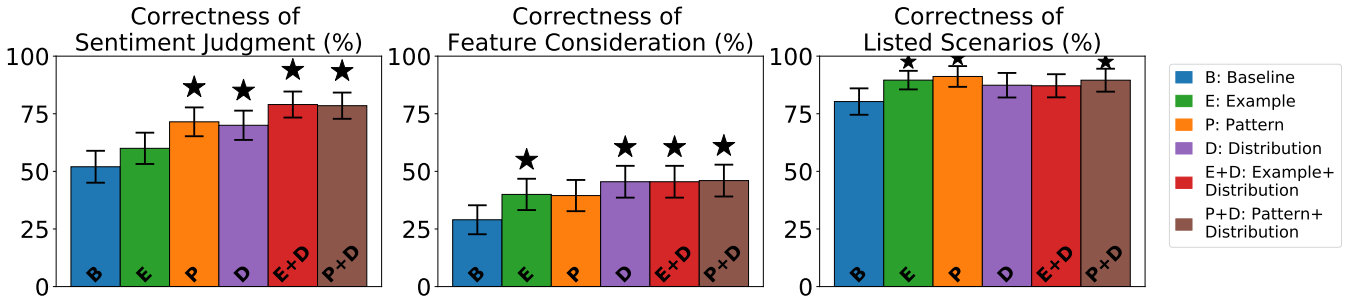

Figure 2: Effects of different interface conditions on participants' understanding with means and 95% confidence intervals. The star mark highlights interface conditions with statistically significant effects ($p < .05$) compared to the Baseline condition.

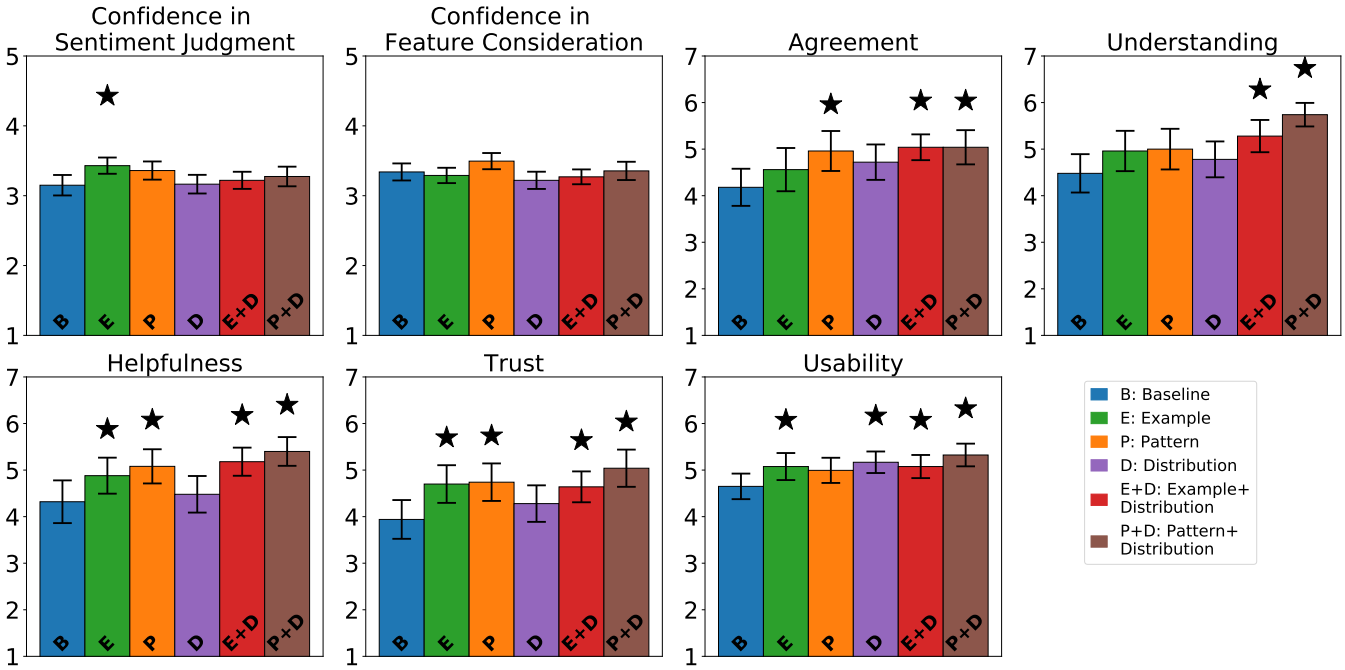

Figure 3: Effects of different interface conditions on participants' perceptions with means and 95% confidence intervals. The star mark highlights interface conditions with statistical significance ($p < .05$) compared to the Baseline condition.

Finally, none of the interface conditions had effects on workload measures compared to the Baseline condition. This implies that our tools improved participants' performance without increasing workload.

4.4 RQ3: Behaviors

In RQ3, we investigated the effects of different interface conditions on participants' behaviors during different tasks. Figure 4 shows our RQ3 results. Our results found two main trends.

First, across all interface conditions, participants made intense sentiment judgments (i.e., selecting the two endpoints on the range slider) 25% of the time or less. Participants made significantly less intense sentiment judgments in the Example and Pattern conditions. One possible reason is that participants received an equal or similar volume of explanations for both sentiments, ultimately leading to judgments of a more moderate intensity. In contrast, access to the label distribution made participants more likely to make an intense sentiment judgment.

Second, participants spent significantly more time on sentiment judgments in all interface conditions except the Distribution condition. This result is not surprising—the Distribution condition was simple as it only showed a pie chart of the sentiment label distribution. In contrast, other interface conditions provided more nuanced textual explanations and interactive functions, so participants engaged in activities such as reading and exploring the system. Regarding the time interval for the entire judgment process, only the Example condition significantly slowed participants down.

One possible reason is that participants spent time reading detailed reviews before making judgments.

5 DISCUSSION

In this section, we summarize the effects of our three tools, report on additional analyses regarding RQ1, discuss design implications, and review limitations of our study.

Summary of Results: In our study, we designed three tools to help participants understand the predictiveness of an unintuitive feature in a sentiment classifier. Our tools had different effects compared to the no-tool baseline. The Distribution tool, which showed the label distribution of training instances containing the word, represents the most abstract evidence. While it helped participants make sentiment and feature consideration judgments correctly and quickly, it did not improve perceptions. The Example tool, which provided a set of training examples containing the word, is a natural approach to explaining unintuitive features. Using this tool, participants spent significantly longer time pondering on each question but failed to make more correct sentiment judgments. The Pattern tool, which extracted contextual patterns containing the word, provided a summary of underlying phenomena associated with the unintuitive feature. While it helped participants make more correct sentiment judgments and listed more correct scenarios, its effects on feature consideration were only marginally significant ($p = .056$).

To summarize, no tool alone could help participants both: (1) understand the predictiveness of an unintuitive feature and (2) have better perceptions of the system and their experience. The best approach is to provide a combination of tools (i.e., Pattern+Distribution and Example+Distribution). Prior studies have found a similar trend. That is, providing a combination of tools or visualizations helps end-users better understand a machine learning model than providing a single tool/visualization [8, 19, 41].

Additional RQ1 Analysis: The listed scenarios in RQ1 reflected how participants rationalized unintuitive word-sentiment relations in their own words. They provide insights into different strategies participants took to explain unintuitive features. We analyzed these data regardless of interface conditions and discovered five strategies. The dominant strategy was to list concrete examples that carry the sentiment and contain the word ($N = 3,030$ out of all $3,447$ listed scenarios). We report on four other strategies as follows.

First, participants explained the predictiveness of a word by describing *semantic* patterns ($N = 347$), including a word's meaning (*"heavy means sturdy."*), its typical usage in product reviews (*"wait indicates a slow shipping speed."*), and its typical usage in natural language (*"stay is used for long-lasting and durable."*).

Second, participants explained the predictiveness of a word by describing *lexical* clues ($N = 11$). This included explaining the use of an adjective or adverb (*"completely is an amplifier typically used for negative things."*), the tense of a word (*"lasted implies that the product stopped working."*), and the differences between a word's singular vs. plural form (*"star is singular, thus less likely to be as positive [than] its plural form like 'five stars'."*).

Third, participants engaged in *pragmatics* analysis ($N = 56$). In such cases, participants thought about the intention and implication behind a word when used in a review. Examples included *"people are more likely to complain about something special [...], especially something as small as a seal."* as well as *"people don't tend to be happy if something just works."* Under the second and third strategies, participants described evidence that is highly nuanced. From the lens of the dual process theory in psychology [24], these participants exhibited more thoughtful and critical investigation, which is a different reasoning process compared to using heuristics such as a word's literal meaning.

Finally, some participants made judgments based solely on the label distribution ($N = 3$), (e.g., *"I see zipper is associated with negative reviews more."*) Such scenarios were not considered valid. However, they demonstrate that some participants took shortcuts by referring to the majority label when explaining the predictiveness of a word.

Design Implications: Our study provides two major implications for future designs of XAI tools.

First, *XAI tools that provide feature importance explanations should prepare to further explain why a feature is predictive.* In our study, we focused on predictive features in a relatively simple model (i.e., logistic regression using unigrams) and a simple task (i.e., sentiment classification). Our results found that many predictive features can be perceived as unintuitive to humans even in such a simple context. The same is likely true for more complex models and tasks. One possible explanation is that while machines learn predictive features from statistical patterns, humans understand a concept (e.g., sentiment) based on semantics, pragmatics, prior experience, and multi-step reasoning [21, 48]. For example, "minutes" was predictive of *positive* for automotive product reviews and *negative* for pet product reviews. Logistic regression models learned these statistical patterns without further asking *why*. However, for these trends to make sense to humans, it is helpful to further realize that "minutes" in automotive product reviews is typically used to describe a quick installation and that "minutes" in pet product reviews is typically used to describe a short product lifespan. These additional explanations are helpful because they provide the broader context that humans need to position and reason about an otherwise unintuitive statistical pattern (i.e., "feature x is predictive of label y").

Second, *predictive yet unintuitive features can be attributed to relevant patterns and contexts in the training data.* While prior studies employed pairwise feature interactions in a local example to explain predictive yet unintuitive unigrams [4, 22, 51], our tools explained unintuitive unigrams by tracing them back to the origin—the label distribution, relevant examples, and contextual patterns extracted from training data. Our results confirm the efficacy of this approach. Compared to participants without access to any tools (i.e., Baseline condition), participants who had access to our contextual pattern tool and distribution tool (i.e., Pattern+Distribution condition) were better able to correctly judge a word's sentiment, correctly judge a word's predictive power, describe scenarios for *why* a word is predictive, and had better perceptions of the AI system and their experiences. This finding suggests two important points supported by prior work. First, end-users benefit from being able to scrutinize *why* a feature is predictive in addition to knowing predictive features only [28, 29]. Second, communicating information about the training data is an effective approach to enhancing an end-user's understanding of and trust in a machine learning model [1].

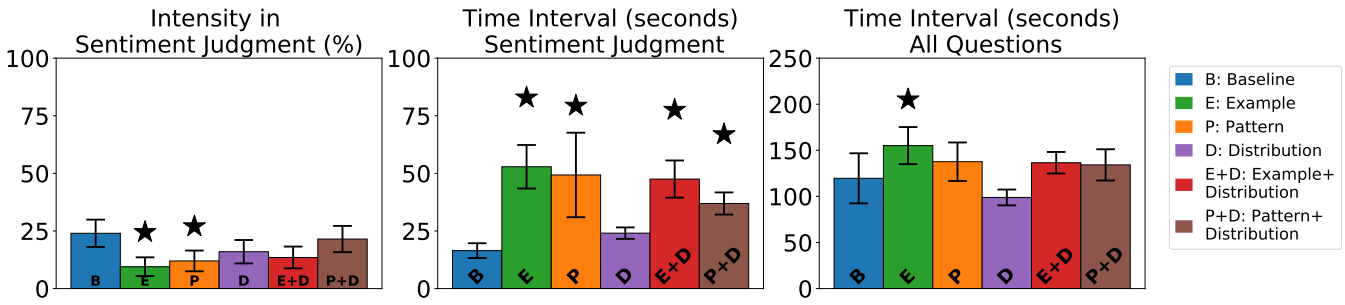

Figure 4: Effects of different interface conditions on participants' behaviors with means and 95% confidence intervals. The star mark highlights interface conditions with statistical significance ($p < .05$) compared to the Baseline condition.

Limitations and Opportunities for Future Work: Our study has several limitations. First, we focused on unintuitive features in the context of sentiment analysis. Explaining unintuitive features in more complex NLP tasks where word-label relations are more nuanced such as deception detection [30], toxicity detection [12], and sarcasm detection [23] may require new tools. Second, we focused on explaining unintuitive unigrams learned by a logistic regression classifier. Future work should explore the generalizability of our tools to other feature representations and models (e.g., non-linear models). Finally, in our study, participants were asked to scrutinize individual features that were predicted to be unintuitive. Future work could explore visualizations that nudge users to recognize that a feature is unintuitive and explore such unintuitive features based on their own curiosity.

6 CONCLUSION

Although showing predictive features is a prevalent approach to explaining machine learning models, features deemed as predictive by machines can be incomprehensible or unintuitive to humans. Predictive yet unintuitive features often represent phenomena that are not obvious without additional explanations. Our research took initial steps toward explaining unintuitive features by using sentiment analysis of product reviews as a case study. We focused on explaining the predictiveness of an unintuitive unigram feature by showing (1) label distribution, (2) sampled training examples, and (3) contextual patterns mined from training data. We conducted a crowdsourced study (N = 300) to evaluate the efficacy of our tools from different perspectives. While the quantitative label distribution could quickly convince participants to accept the unintuitive association between a feature and a label, seeing concrete examples and especially contextual patterns improved participants' qualitative understanding of the underlying phenomena and subjective perceptions of the provided tools. The best results were achieved when the tools where combined. Our research shows that it is both *possible* and *useful* to explain predictive yet unintuitive features learned by sentiment classifiers. It opens up research opportunities to investigating problems of similar nature in a wider range of tasks and models.

ETHICS STATEMENTS

Ethical considerations statement: The study was reviewed and approved by our Institutional Review Board (IRB). During the study, participants were asked to complete several tasks that were not considered to be too cognitively demanding. The study used a between-subjects design. This was partly done to help prevent a "spill-over effect" between interface conditions within a single participant.

Researcher positionality statement: This work was partly inspired by our experiences as instructors of graduate-level text data mining courses, where curious students are frequently puzzled by certain features that are learned to be important according a model but do not make immediate sense to humans. We found that such questions cannot be sufficiently answered by current explainable machine learning techniques. We therefore set out to develop and evaluate tools to bridge this gap, hoping these tool can help people learn about a predictive task and increase their understanding of and trust in machine learning models.

Adverse impact statement: In our study, we developed different tools to help people understand the predictiveness of an unintuitive feature. Our tools focused on displaying quantitative evidence (e.g., the label distribution across training instances containing the word) and qualitative evidence (e.g., contextual patterns containing the word in the training data). Our RQ1 results suggest that quantitative evidence alone can influence people to accept that a feature is strongly predictive without understanding *why*. As shown in Figure 2, participants in the Distribution (vs. Baseline condition) were significantly more likely to accept that a feature is strongly predictive but were *not* significantly more likely to list valid scenarios to justify their judgment. This result underscores the importance of showing both quantitative and qualitative evidence. Showing only quantitative evidence may influence people to superficially accept that a machine learning model is doing the "right thing" without critically understanding why.

REFERENCES

[1] Ariful Islam Anik and Andrea Bunt. 2021. Data-centric explanations: explaining training data of machine learning systems to promote transparency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[2] 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.
[3] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".
[4] Vadim Borisov and Gjergji Kasneci. 2022. Relational Local Explanations. *arXiv preprint arXiv:2212.12374* (2022).
[5] John Brooke. 2013. SUS: a retrospective. *Journal of usability studies* 8, 2 (2013), 29–40.
[6] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th international conference on intelligent user interfaces*. 454–464.
[7] Zoran Bursac, C Heath Gauss, David Keith Williams, and David W Hosmer. 2008. Purposeful selection of variables in logistic regression. *Source code for biology and medicine* 3, 1 (2008), 1–8.
[8] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*. IEEE, 160–169.
[9] Carrie J Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th international conference on intelligent user interfaces*. 258–262.
[10] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–14.
[11] Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. 335–336.
[12] 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.
[13] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 1721–1730.
[14] Valerie Chen, Q Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. 2023. Understanding the role of human intuition on reliance in human-AI decision-making with explanations. *Proceedings of the ACM on Human-computer Interaction* 7, CSCW2 (2023), 1–32.
[15] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I think i get your point, AI! the illusion of explanatory depth in explainable AI. In *26th International Conference on Intelligent User Interfaces*. 307–317.
[16] Ian Covert, Scott M Lundberg, and Su-In Lee. 2020. Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems* 33 (2020), 17212–17223.
[17] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv* (2017). https://arxiv.org/abs/1702.08608
[18] Donald E Farrar and Robert R Glauber. 1967. Multicollinearity in regression analysis: the problem revisited. *The Review of Economic and Statistics* (1967), 92–107.
[19] Ana Valeria Gonzalez, Gagan Bansal, Angela Fan, Robin Jia, Yashar Mehdad, and Srinivasan Iyer. 2020. Human evaluation of spoken vs. visual explanations for open-domain qa. *arXiv preprint arXiv:2012.15075* (2020).
[20] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.
[21] Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2022. A fine-grained comparison of pragmatic language understanding in humans and language models. *arXiv preprint arXiv:2212.06801* (2022).
[22] Joseph D Janizek, Pascal Sturmfels, and Su-In Lee. 2021. Explaining Explanations: Axiomatic Feature Interactions for Deep Networks. *J. Mach. Learn. Res.* 22 (2021), 104–1.
[23] Aditya Joshi, Pushpak Bhattacharyya, and Mark J Carman. 2017. Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)* 50, 5 (2017), 1–22.
[24] Daniel Kahneman. 2011. *Thinking, fast and slow.* macmillan.
[25] Gary M Kaufmann and Terry A Beehr. 1986. Interactions between job stressors and social support: Some counterintuitive results. *Journal of applied psychology* 71, 3 (1986), 522.
[26] Peter E Kennedy. 2005. Oh no! I got the wrong sign! What should I do? *The Journal of Economic Education* 36, 1 (2005), 77–92.
[27] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*. PMLR, 1885–1894.
[28] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*. 126–137.
[29] Todd Kulesza, Weng-Keen Wong, Simone Stumpf, Stephen Perona, Rachel White, Margaret M Burnett, Ian Oberst, and Amy J Ko. 2009. Fixing the program my computer learned: Barriers for end users, challenges for the machine. In *Proceedings of the 14th international conference on Intelligent user interfaces*. 187–196.
[30] 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.
[31] 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.
[32] J. R. Landis and G. G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159–174.
[33] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7871–7880.
[34] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).
[35] Laura L Nathans, Frederick L Oswald, and Kim Nimon. 2012. Interpreting multiple linear regression: a guidebook of variable importance. *Practical assessment, research & evaluation* 17, 9 (2012), n9.
[36] Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification. In *16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 1069–1078.
[37] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. 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)*. 188–197.
[38] Changhoon Oh, Seonghyeon Kim, Jinhan Choi, Jinsu Eun, Soomin Kim, Juho Kim, Joonhwan Lee, and Bongwon Suh. 2020. Understanding How People Reason about Aesthetic Evaluations of Artificial Intelligence. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 1169–1181.
[39] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.
[40] Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. 2020. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems* 33 (2020), 19920–19930.
[41] 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.
[42] 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*.
[43] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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)*. 3982–3992.
[44] Marco Tulio Ribeiro and GitHub Contributors. 2020. *LIME: Explaining the predictions of any machine learning classifier.* https://github.com/marcotcr/lime
[45] 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.
[46] 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.
[47] 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*. 611–636.
[48] Rita Sevastjanova and Mennatallah El-Assady. 2022. Beware the rationalization trap! when language model explainability diverges from our mental models of language. *arXiv preprint arXiv:2207.06897* (2022).
[49] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *International conference on machine learning*. PMLR, 3145–3153.
[50] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*. PMLR, 3319–3328.

[51] Michael Tsang, Youbang Sun, Dongxu Ren, Beibei Xin, and Yan Liu. 2018. Can I trust you more? Model-Agnostic Hierarchical Explanations. (2018).

[52] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt. 2020. How do visual explanations foster end users' appropriate trust in machine learning?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 189–201.

[53] Chih-Kuan Yeh, Joon Kim, Ian En-Hsu Yen, and Pradeep K Ravikumar. 2018. Representer point selection for explaining deep neural networks. *Advances in neural information processing systems* 31 (2018).

[54] Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*. Association for Computational Linguistics, 3914–3923.

[55] Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. 2021. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics* 10, 5 (2021), 593.