# Supporting Serendipitous Discovery and Balanced Analysis of Unstructured Text with Interaction-Driven Metrics and Bias-Mitigating Suggestions

MAHMOOD JASIM, University of Massachusetts Amherst

CHRISTOPHER COLLINS, Ontario Tech University

ALI SARVGHAD, University of Massachusetts Amherst

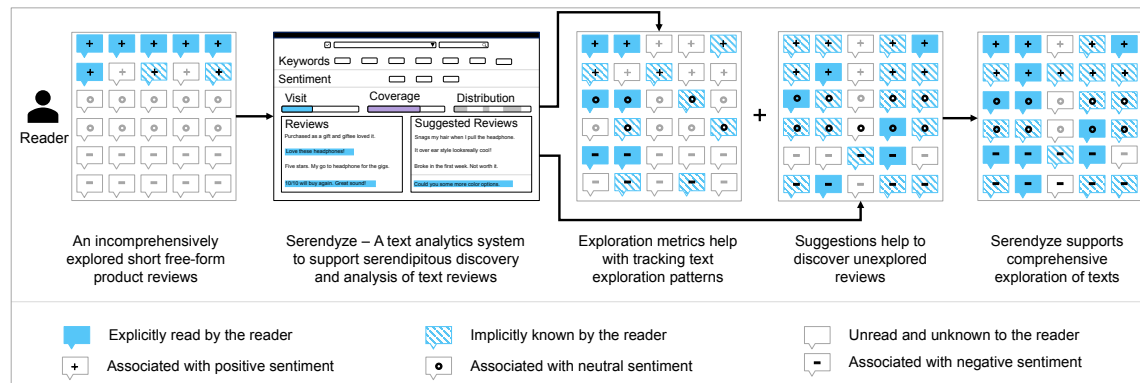NARGES MAHYAR, University of Massachusetts Amherst

Fig. 1. Serendyze is a text analytics system that uses two novel interventions — exploration metrics and a bias mitigation model — to enable readers to more comprehensively explore short free-form texts such as product reviews. The exploration metrics help readers to track their data exploration across different facets such as sentiments. The bias mitigation model suggests reviews that are semantically and sentiment-wise dissimilar to what the readers have been exploring so that they can discover a broader range of reviews. Integrated with an interactive interface, these features enable readers to gain comprehensive knowledge about the data prior to decision-making.

In this study, we investigate how supporting serendipitous discovery and analysis of short free-form texts, such as product reviews can encourage readers to explore texts more comprehensively prior to decision-making. We propose two interventions — Exploration Metrics that help readers understand and track their exploration patterns through visual indicators and a Bias Mitigation Model that maximizes knowledge discovery by suggesting readers sentiment and semantically diverse reviews. We designed, developed, and evaluated a text analytics system called Serendyze, where we integrated these interventions. We asked 100 crowd workers to use Serendyze to make purchase decisions based on product reviews. Our evaluation suggests that exploration metrics enable readers to efficiently cover more reviews in a balanced way, and suggestions from the bias mitigation model influence readers to make confident

Authors' addresses: Mahmood Jasim, University of Massachusetts Amherst, mjasim@cs.umass.edu; Christopher Collins, Ontario Tech University, Christopher.Collins@ontariotechu.ca; Ali Sarvghad, University of Massachusetts Amherst, asarv@cs.umass.edu; Narges Mahyar, University of Massachusetts Amherst, nmahyar@cs.umass.edu.

data-driven decisions. We discuss the role of user agency and trust in text-level analysis systems and their applicability in domains beyond review exploration.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**.

Additional Key Words and Phrases: serendipity, text exploration, bias mitigation model

## 1 INTRODUCTION

Short free-form online texts can considerably influence how readers acquire knowledge on critical issues [106], develop perception on current affairs [100], and make crucial data-driven decisions [118]. Especially, in the online commerce scenario, free-form text such as product reviews are often considered as the most influential factor behind sales and attitudes towards a product [35, 106]. As such, customers of online products often depend on product reviews to make purchase decisions [47, 106]. However, due to the digital data deluge [55, 65], coupled with a limited amount of time to accrue insights from raw text data [65, 83], potential customers of online products often struggle to comprehensively peruse and tease apart valuable insights from the mundane and the redundant [55, 83, 103]. As such, a broad range of commercial decisions, ranging from minor personal paraphernalia to major investments, are often made based on an incomplete exploration and understanding of the underlying product reviews [49, 65, 106].

Due to the expansive problem domain, interdisciplinary researchers from human-computer interaction, natural language processing, and information retrieval have often collaboratively investigated and proposed several methods to enable comprehensive sensemaking of texts [2, 5, 46, 114]. While these methods provide aggregated statistics and overall summaries, they often overlook unpopular and marginalized opinions by favoring popular or recent information [18, 57]. Prior works also suggest that users' innate cognitive biases often influence how they interact with analysis systems [108]. For instance, people who are oversensitive to consistency [42, 108] tends to interact with data that supports the largest encompassing hypothesis, dismissing other data. When reading product reviews, this bias may influence a reader to read reviews that are predominantly positive or negative [49] Furthermore, persistence of impressions based on discredited evidence [42, 108] often results in continuous interaction with data supporting a hypothesis that has been disproved. This bias may prompt readers to ignore reviews that highlight issues with their preferred products. These biases often manifest when users are overwhelmed with large amounts of data, resulting in them following their preconceptions, anchoring biases, and using biases as filters to explore underlying data [109]. This manifestation of innate bias is inadvertently amplified by systems that respond to users' interactions and preferences — leading to the facilitation of an incomplete, ineffective, and often biased data exploration prior to decision-making [107, 108]. As a result, current approaches towards aggregated analysis alone might not be effective in opinion analysis of product reviews where decision-making is dependent on comprehensive exploration and rigorous text analysis [65, 83, 108].

Recent interest in data exploration and discovery [64, 71] along with beyond-accuracy metrics [57] has prompted research into identifying and presenting diverse and serendipitous information to increase people's coverage and understanding of the data. Coupled with information visualization research geared towards providing navigational

cues to investigate how readers interact with visual artifacts [95, 107, 111], serendipitous information[1] — information that is yet unexplored by the readers and may add to their knowledge of the underlying data — has shown promise in expanding the depth and breadth of data exploration [79, 95, 107]. However, these existing methods that encourage data exploration by increasing data coverage and mitigating biases were not designed for texts, and their effectiveness on numerical or categorical data may not translate to texts.

Other approaches provide summaries based on text analysis methods, such as topic modeling [3, 51], keyword extraction [105], and sentiment analysis [43]. However, these summaries may not always be representative of information present in the actual text [52]. Moreover, summarization and aggregation might lead to the marginalization of unpopular opinions [57, 73]. As such, it might be beneficial to support serendipitous discovery and analysis at the text level to supplement summary-level text analysis to enable readers to explore product reviews more comprehensively.

In this work, we investigate interventions for supporting serendipitous discovery and analysis to help readers comprehensively read and tease apart valuable insights from short free-form texts in a balanced way. To that end, we investigate a two-pronged approach. First, we provide three interaction-driven exploration metrics **Visit** — a measure of text a reader has explicitly interacted with, **Coverage** — a measure of text covered by a reader implicitly, that are similar and redundant to the text they have already visited, and **Distribution** — a measure of the relation of text the reader has visited from different facets, such as sentiments, to the true distribution of that facet in the dataset. Second, we propose a *bias mitigation model* to improve knowledge discovery and balance overall text exploration. The model tracks how a reader has been visiting texts and generates suggestions that are semantically and sentiment-wise different from what they have visited already.

The interaction-driven exploration metrics act as an awareness mechanism to enable readers to understand and track their data exploration progress and patterns through visual indicators. They highlight which reviews the readers have implicit and explicit knowledge about as well as the reviews that are left unexplored. The bias mitigation model enables serendipitous discovery and offers a complementary view of texts. It helps readers to balance their holistic understanding, increase data coverage, and mitigate bias towards specific sentiments in texts by providing them with suggestions that are different from what they have visited already.

We integrated the exploration metrics and the bias mitigation model with an interactive text analytics system, *Serendyze*. We use Serendyze to investigate the following questions:

(1) **RQ1:** Does supporting serendipitous discovery and analysis help readers to perform in-depth exploration to cover more texts?

(2) **RQ2:** How do readers' text exploration behaviors change when they have access to their exploration patterns?

(3) **RQ3:** How does suggestions from unexplored text impact readers' decisions?

In this study, we used product reviews as an example dataset for large-scale short free-form online texts due to their availability and abundance. Furthermore, among myriad online products, we selected headphones as the candidate due to their ubiquitous usage [104]. To study how serendipitous discovery and analysis impact review exploration, knowledge gathering, and decision-making, we conducted a crowd-sourced between-subjects study in which 100 participants used Serendyze to select their most preferred headphones to recommend to someone.

The findings from our study demonstrate that exploration metrics and bias mitigating suggestions enable readers to make more informed and confident decisions. We found that the majority of the participants who used both exploration

---

[1]The term *Serendipity* has been defined in various ways by previous researchers [57, 105]. In this paper, we define *serendipity* as an *unexpected yet beneficial discovery that adds to the knowledge of the readers about the data they are exploring.*

metrics and suggestions were confident that they visited enough reviews to make an informed decision (16/25) as opposed to the participants who did not use these features (10/25). The majority of the participants who used the features were also confident that they had made the right decision (19/25) compared to those who did not (9/25).

Based on the collected usage logs, we also found that participants who used exploration metrics and bias mitigating suggestions *covered* an average of 234 reviews before making a decision with a 12.28 coverage to time spent ratio, compared to only 66 reviews covered by participants who did not use these features with 8.64 coverage to time spent ratio. By the term *covered*, we refer to the number of reviews the participants have explicit or implicit knowledge about. We consider a participant has *explicit* knowledge about a review if they have visited the review by marking it as read and they have *implicit* knowledge about reviews that are semantically similar to the reviews they visited. The coverage numbers suggest that readers who used exploration metrics and bias mitigating suggestions had a much broader coverage and knowledge of reviews.

The collected usage logs and responses to the post-study questionnaire also suggest that Serendyze helped the participants (18/25) to gather comprehensive knowledge from reviews by enabling them to visit reviews in a balanced way without leaning towards specific sentiments (positive, negative, or neutral). We consider a readers' review exploration as "balanced" when the sentiments visited by the reader reflect the true distribution of sentiments present in the dataset. Furthermore, the participants who used the suggestions discovered reviews of opposing viewpoints that they were not aware of before, which enriched their knowledge about the products and positively impacted their decisions.

Based on the findings from our study, we highlight our contributions as follows:

(1) A novel approach to support serendipitous discovery and analysis using three interaction-driven exploration metrics and a bias mitigation model to enable more comprehensive text exploration prior to decision-making.

(2) Empirical evidence that demonstrates the utility of an example text analytics system called Serendyze, integrated with functionalities to explore serendipitous information from product reviews. The system shows reflective metrics to readers about their data exploration patterns and suggests reviews they might not have considered otherwise to accumulate comprehensive knowledge needed for informed decision-making.

(3) Discussions on how systems that support serendipitous discovery and analysis can combat biased text exploration. We also discuss readers' agency in mixed-initiative systems and the expansion of text-level analytics systems for critical data-driven decision-making beyond product reviews.

## 2 RELATED WORK

In this section, we review existing literature on free-form texts, such as product reviews, and their impact on knowledge acquisition and decision-making. We also describe existing tools and techniques for product review analysis and how serendipitous information discovery plays a role in text exploration and understanding.

### 2.1 Impact of Online Text on Decision-making

Online short free-form text contributes the bulk of information found in various domains including online commerce, news forums, social media, and civic engagement platforms. Despite only a small part of the population (7%) generating this information [91], online text may serve as a measure of public opinion that is immediate, spontaneous and often honest [84]. Previous studies have shown how such comments have the potential to increase readers' understanding of people's opinions, the reasoning behind their opinion formulations, and how decisions are made based on such reasoning [40]. In the online commerce scenario, researchers have shown that product reviews are the most important

factor that influences sales and attitudes towards a product [35, 106]. In 2020, Qualtrics revealed that 93% of consumers mentioned how online product reviews impacted their purchasing decisions [87]. While ratings can act as an initial filter to find desirable products, it is the product reviews that enable a reader to decide which product to purchase [47].

## 2.2 Visual Analytics Approaches for Online Product Reviews

Researchers have explored various text analysis techniques such as opinion extraction, sentiment analysis, topic modeling, and trend analysis, and combined them with visualizations to enable exploration and analysis of large-scale short free-form text corpus, such as product reviews [4, 5, 24, 59, 62]. For example, to explore and analyze online product reviews, OpinionBlocks provides an aspect-based summary of product reviews using a block visualization to present an overview of positive and negative reviews [46]. Review Spotlight provides summaries of user reviews on restaurants using objective-noun pairs organized as tag clouds [116].

To facilitate comparison of opinions among different products derived from text mining across various features, Carenini et al. proposed a multimedia interface [13] to aggregate opinions using bar chart visualizations. Opinion Observer is another such system that enables comparison of people's opinions on product features based on opinion mining by providing a summary of pros and cons of the product features [70]. Chen et al. utilized term-variation patterns to identify underlying topics present in product reviews to facilitate the understanding of conflicting opinions towards online products using a host of visualizations [17].

Others have experimented with extracting and presenting affective content from product reviews. For example, Gregory et al. enabled user-directed affective content exploration in product reviews using variations of rose plots [37]. Furthermore, they experimented with thematic clustering based on keyword extraction to enable exploration of product reviews [37]. OpinionSeer enables multilevel exploration of opinion data from hotel reviews with explicit consideration towards uncertainty using augmented radial charts [114].

Prior works suggest that the majority of text analytics methods often focus on providing aggregated statistics and summaries and put less emphasis on comprehensive exploration and knowledge discovery from the actual text [4, 59, 62]. The potential impact of product reviews on the decision-making process [87, 106] necessitates the investigation of methods that may enable readers to acquire a comprehensive understanding of reviews prior to making critical decisions.

## 2.3 Tools and Techniques to Support Serendipitous Data Discovery and Analysis

Research in recommender systems, a subclass of information filtering systems, focuses on identifying data items by predicting how a user might rate the item [92]. Although research in this area has mostly focused on the accurate prediction of user preferences, there has been a recent interest in exploring methods beyond traditional accuracy-based metrics [57]. For instance, researchers have explored metrics to diversify data recommendations [101], provide novel recommendations [119] or support serendipitous discovery of data items [80]. Among various beyond-accuracy metrics explored in prior works [41, 57], serendipity has received significant attention in the last decade.

The term *serendipity* is often referred to as the process of finding valuable or surprising things that are not looked for [7, 41]. Others have defined serendipity as a combination of surprise and relevance [41]. However, existing methods adhering to such definitions have mainly focused on suggesting relevant data items and rejecting irrelevant ones [60, 75], leading to neglecting unpopular or marginalized opinions. For instance, consider a reader who is reading product reviews of headphones using a system that suggests reviews to the reader based on relevance. If the reader is reading reviews that focus on the price, they will receive more, albeit different, suggested reviews about price. They might not be suggested reviews regarding other aspects such as color or sound quality because the system may consider these

aspects to be irrelevant based on what the reader has been reading. As a result, the reader might make a purchase decision without learning about other aspects of the headphone that might be important to them. In contrast, we consider serendipity to be an *unexpected yet beneficial discovery of information that adds to the readers' knowledge.* Our goal is to support the serendipitous discovery of unexpected information that could help readers broaden and improve their knowledge acquisition instead of reinforcing their existing preconceptions with relevant data items.

Previous research in data visualization has explored ways to support serendipitous discovery and analysis of data [3, 31, 51]. For instance, Bohemian Bookshelf provides visualizations for exploring book collections that enable people to discover trends and relations within the collection in a playful manner [105]. Another work, Serendip, provides a topic modeling tool with multiple views. It focuses on intermixing different scales of data inquiry and information types by visualizing the relationships between the data items [3]. Another visualization tool that promotes serendipitous discovery is PivotPaths [31]. It enables playful and casual exploration of interlinked metadata using visual paths in enticing arrangements to motivate people to explore the information. Footprints is another analytics tool that uses multiple interconnected visualizations to help users navigate through news articles [51]. Footprints also enables people to tag the data as *Read*, *To Read*, and *Useful* to track exploration progress and data coverage.

While these tools provide functionalities to support the serendipitous exploration of documents, their effectiveness for exploring relatively large text documents, including academic papers, books, and news articles, may not translate to short free-form text such as product reviews. Furthermore, these tools often enable the exploration and analysis of large text corpora at the summary level. For instance, PivotPaths enables serendipitous discovery of relationships between facets such as author name, venue, and keywords, but not the actual text content of academic publications. Similarly, Footprints enable serendipitous discovery of topics and other metadata such as dates and sources, but not the text content of documents. The lack of systems to explore short free-form text data necessitates a deeper investigation of how providing serendipitous information at the text level impacts the data exploration and analysis process.

## 2.4 Approaches to Increase Data Coverage and Avoid Biased Exploration

Prior work suggests that users of analytics tools are often prone to biases when exploring data [33, 108]. While interacting with the data and system artifacts, a user's internal biases and presumptions towards the data can impact the exploration and analysis process [49, 107, 108]. Such biases include *oversensitivity to consistency* [42, 108], where an analyst tends to interact with data that supports the largest encompassing hypothesis and they dismiss other data. In product review domain, this bias may manifest and influence a reader to read reviews that are predominantly positive or predominantly negative based on the aggregation of reviews [24, 49]. Furthermore, biases such as *persistence of impressions based on discredited evidence* [42, 108] influence analysts to continue interacting with data that supports a hypothesis but has been disproved already. This bias can influence readers to make biased decisions based on their brand or product preference, even when there are reviews that highlight issues with their preferred products. One approach to mitigating such biases could involve exposing the differences between the data a user has explored and the overall characteristics of the complete underlying data, making users aware of the existence of their innate biases that are potentially injected during their data exploration [20, 49, 108].

Existing systems designed towards combating such biases often provide visual and navigational cues on how the user has been exploring the data and interacting with the system to inform users of potentially biased interactions and exploration [51, 95, 107]. For instance, Sarvghad et al. proposed a visual analytics tool to provide analysis history to highlight the dimension coverage of data dimensions explored by the user [95]. These data dimensions are comprised of different attributes present in tabular data. The tool employed a variation of scented widgets to assist analysts in

forming questions based on their past data exploration patterns. Wall et al. [107] also experimented and modeled users' potential biased behavior while using scatterplots based on the history of their data exploration patterns.

While these tools, methods, and experiments shed light on the potential of providing navigation cues to avoid biased exploration and increase data coverage, they are mostly focused on ordinal, categorical, or numerical data. Furthermore, these tools were not designed to investigate how providing such information may impact readers' knowledge acquisition prior to decision-making. As such, the effects of supporting serendipitous discovery and analysis of reviews to help readers explore, cover more information, and gather knowledge prior to decision-making remain largely unexplored.

## 3 SERENDYZE

Serendyze is designed and developed as an interactive text analytics system to propel readers to more comprehensively explore and analyze short free-form texts prior to decision-making. For this study, we focused on product reviews from Amazon, but Serendyze can be used with any text corpora, including social media posts and comments posted on engagement platforms. Here, we describe different components and functionalities integrated with Serendyze along with the exploration metrics and bias mitigating model, which support serendipitous discovery and analysis.

### 3.1 Exploration Metrics

In this work, we propose three interaction-driven exploration metrics — **Visit**, **Coverage**, and **Distribution**. The exploration metrics are designed to enable readers to track their data exploration progress and patterns (see Fig. 2).

*3.1.1 Visit.* **Visit** is a measure of reviews a reader has explicitly interacted with. To measure the Visit metric, Serendyze maintains a list of reviews that the reader has marked as read as the *visited list*, $V$. Visit is simply the percentage of reviews marked as read by the reader from the total number of reviews for the product ($N$) using equation 3.1.

$$Visit = \left\lceil \frac{|V|}{N} \right\rceil \cdot 100 \tag{3.1}$$

*3.1.2 Coverage.* We define **Coverage** as a measure of reviews the reader has knowledge of either explicitly or implicitly. We assume that a reader has *explicit* knowledge about a review if they have visited the review and *implicit* knowledge about a review ($x$) if they have already visited another review ($y$) that is semantically similar to the review ($x$) [21]. For instance, consider two reviews on the same product: "*Good Headphones, Great for price. The headphones work quite well. They don't feel like great headphones but they have held up pretty well and produce good sound.*" and "*Great sound, affordable. Great sound for the price and seem like they will last for a while. A good value for the price as well.*" These reviews are sufficiently semantically similar that they can be considered redundant. As such, if a reader visits one of these reviews by marking it as read, we conclude that they have *covered* the other review. The Coverage metric thus tracks the percentage of reviews the reader has either explicit (visit) or implicit (semantically similar) awareness of.

To measure Coverage, we first convert each review to a vector representation which embeds semantic information using Doc2Vec [68]. While Doc2Vec is a generalization of the popular Word2Vec [78] embedding, Doc2Vec's advantage over Word2Vec is its applicability on variable-length documents, making Doc2Vec suitable for embedding short free-form texts such as reviews. We decided to use Doc2Vec over other bi-directional language models, such as BERT [27] and Elmo [86] as it is more interpretable and less computationally expensive [67] for measuring the semantic similarity among reviews in zero-shot environments. However, due to the modular design of Serendyze, Doc2Vec can be replaced with more contemporary transformer-based models for appropriate tasks.
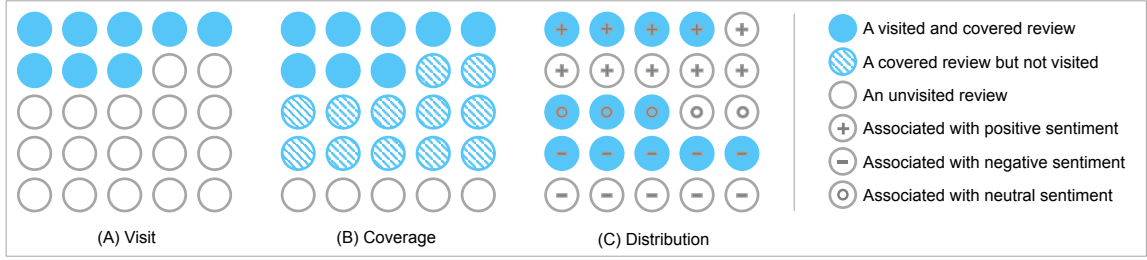
Fig. 2. Three exploration metrics: A) **Visit** - a measure of text the reader has directly interacted with, B) **Coverage** - a measure of texts covered by the reader implicitly through semantic similarity and redundancy, and C) **Distribution** - a measure of the relation of text the reader has visited from different facets, such as sentiments, to the true distribution of that facet. A filled cyan circle represents explicit knowledge: a review the reader has directly interacted with (visited). A striped cyan circle represents implicit knowledge: a review that the reader has not interacted with directly but has *covered* through direct interaction with another semantically similar review. An unfilled gray circle represents a review the reader has not interacted with and has no implicit or explicit knowledge about.

Serendyze maintains three live lists of Doc2Vec vectors of reviews: visited ($V$), unvisited ($U$), and covered ($C$). When a review is visited, pairwise cosine similarity [36] between $V$ and $U$ is measured. Based on experiments and pilot studies, we use a normalized similarity score of 0.8 as the threshold to determine if a review is similar enough to be considered redundant and placed in the covered list ($C$). Any review from the unvisited list with a similarity value of at least 0.8 with any review from the visited list is added to the covered list ($C_L$). Finally, the Coverage value is measured as the percentage of covered reviews — reviews that a reader has implicit or explicit knowledge about — from the total number of reviews ($N$) for the product using equation 3.2. Note that the visited list ($V$) is a subset of the covered list ($C$) as the latter contains all visited reviews with additional redundant reviews.

$$Coverage = \left\lceil \frac{|C|}{N} \right\rceil \cdot 100 \tag{3.2}$$

*3.1.3   Distribution.* We define **Distribution** as a measure of the relation of reviews the reader has visited from different facets, such as sentiments, to the true distribution in the dataset. For this study, we considered sentiments (positive, neutral, and negative) as the facet to measure Distribution. However, these facets can be customized to include star ratings, sentiments, topics, or other metadata or text mining results. In our study, Distribution is a measure of consistency and equilibrium of a reader's review exploration of various sentiments. For instance, if a reader focuses heavily on positive reviews while ignoring negative or neutral ones, we consider such exploration patterns as not well-distributed. To measure Distribution, Serendyze counts the total number of positive, neutral, and negative reviews for a product. During use, Serendyze maintains separate lists of positive, neutral, and negative reviews that a reader has visited. As the reader continues to visit reviews, the proportions of visited sentiments are calculated using equation 3.4, where $V_X$ is the visited list of sentiment $X$ and $U_X$ is the unvisited list of sentiment $X$. $X$ can be positive, neutral, or negative.

The Distribution metric is designed to help readers understand how well their visit history reflects the true distribution of sentiments. For example, if a dataset contains vastly more positive reviews than other categories, an unbiased sample of the data would also contain more positive reviews. Because it measures proportions, by aiming for the Distribution measure for each sentiment to be equal, the reader could ensure their understanding is reflective of the dataset.

$$Distribution_X = \frac{|V_X|}{|V_X \cup U_X|} \tag{3.4}$$

While reading reviews, if a reader's Distribution metric for a sentiment exceeds the Distribution metrics of other sentiments by more than 7%, it is flagged as a tendency to lean towards that sentiment. Note that Distribution detects imbalance based on the proportions of reviews visited for each sentiment, not the absolute number. For example, if a reader is focusing on positive reviews to the point where the proportion of positive reviews visited exceeds the proportion of negative and neutral by 7% or more, we consider the reader's exploration is skewed towards positive reviews. In this way, the measure helps readers stay aware of how their visited reviews reflect the true distribution.

The threshold of 7% was determined in pilot testing. We found that a threshold lower than 7% too aggressively penalized exploration of a certain sentiment. In contrast, a higher value allows readers to neglect other sentiments for a longer period of time. The strictness of this threshold is fully customizable based on the facets and dataset used.

## 3.2 Bias Mitigation Model

In this work, we propose a heuristic bias mitigation model to extract and present suggestions to readers based on their interactions with reviews. The model is designed to focus on supporting serendipitous discovery and balanced analysis of reviews by providing suggestions that encourage readers to visit more reviews and improve their knowledge acquisition about the products. To that end, the model suggests unvisited reviews that are semantically and sentiment-wise dissimilar to the reviews the reader has visited already. The suggestions are generated to mitigate biased exploration and guide readers to an understanding of the data, which is reflective of the true distributions of the semantic and sentiment diversity in the reviews. The complete algorithm to generate the model is presented in Algorithm 1.

There are two major components of the model: (1) The **dissimilarity measure** that calculates how dissimilar the suggestion is from the reviews that the reader has already visited and (2) The **sentiment measure** that calculates if the reader is focusing too much on a specific sentiment and neglecting others. The algorithm is called to generate bias mitigating suggestions for every review visited and marked as read by the reader using Serendyze. Serendyze maintains several lists, including lists of Doc2Vec vectors of visited ($V$) and unvisited ($U$) reviews, and a list of suggestions the reader has visited ($S$). The list of visited suggestions also contains flags about the primary reason a suggestion was made (to maximize dissimilarity or unbias sentiment).

For each prospective suggestion $u$, the projected distribution of sentiments is calculated (lines 7–10). Serendyze calculates the coefficient of variation ($CoV$, line 11), a measure of relative variability measured by the ratio of standard deviation to mean of the visited review proportions of different sentiments. A coefficient of variation of less than 1 indicates that the reader is exploring reviews of different sentiments in a distributed fashion. Higher values indicate a greater degree of variability and unbalanced exploration. Then, Serendyze calculates pairwise cosine similarity measurement from ($u$) to every review in the visited list ($V$) to generate a maximum dissimilarity score.

When suggesting $u$ would not result in a high $CoV$, the sentiment score $s$ for $u$ is assigned as $1 - CoV$ (lines 13–14). This results in a relatively high score of $s$, which is appropriate as we prefer suggestions that do not introduce sentiment distribution biases. When suggesting $u$ would unbalance the sentiment distribution $CoV > 1$, the sentiment score $s$ is inversely related to the proportion of $u$'s sentiment already visited. As a result, unvisited reviews with sentiments that have not been visited (lower proportion value) will now be scored higher.

For example, if a reader has been exploring too many positive reviews, they will gradually start to receive negative and neutral reviews as suggestions. This will increase the chances of an unvisited review with a potentially neglected sentiment to be ranked higher by the model, increasing the reader's chance of receiving diverse suggestions. The final score is a weighted combination of $s$ and $d$. The default weighting is equal (line 18), and the adjustment of weighting factors is discussed below. The top 5 scoring suggestions are returned.

---

**Algorithm 1** Bias Mitigation Model

---

1: **procedure** Get-Suggestion($U, V, S$)  ▷ U, V, S are arrays of unvisited, visited, and visited suggested reviews

2: $\quad M \leftarrow 1 - \frac{|S_Y|}{|S|}$  ▷ Calculate score modifiers. $Y \in [\text{Dissimilarity, Sentiment}], |M| = 2$

3: $\quad T \leftarrow \phi$  ▷ List of objects to store candidate reviews and their scores in tuples

4: $\quad$ **for** u in $U$ **do**

5: $\qquad d \leftarrow \phi$  ▷ minimum dissimilarity score

6: $\qquad s \leftarrow \phi$  ▷ sentiment score

7: $\qquad V' = V + u$  ▷ Add candidate review to temporary $V$

8: $\qquad P_{pos} \leftarrow |V'_{pos}|/|V_{pos} \cup U_{pos}|$  ▷ proportion of positive reviews visited

9: $\qquad P_{neut} \leftarrow |V'_{neut}|/|V_{neut} \cup U_{neut}|$  ▷ proportion of neutral reviews visited

10: $\qquad P_{neg} \leftarrow |V'_{neg}|/|V_{neg} \cup U_{neg}|$  ▷ proportion of negative reviews visited

11: $\qquad CoV \leftarrow$ Coefficient-of-Variation ($P_{pos}, P_{neut}, P_{neg}$)  ▷ Prospective CoV if $u$ is visited

12: $\qquad d \leftarrow 1 - (min(CosineSimilarity(u, V)))$  ▷ Find max dissimilarity from already visited reviews

13: $\qquad$ **if** $CoV < 1$ **then**  ▷ Adding $u$ results in distributed reading

14: $\qquad\quad s \leftarrow 1 - CoV$  ▷ Give $u$ a higher sentiment score

15: $\qquad$ **else if** $CoV > 1$ **then**  ▷ Adding $u$ results in an unbalanced reading

16: $\qquad\quad s \leftarrow 1 - P_X$  ▷ $X \in [pos, neut, neg]$ associated with $u$

17: $\qquad$ **if** $CoV < 1 \, \&\& \, M = \phi$ **then**  ▷ $M = \phi$, when the reader has not visited any suggestion

18: $\qquad\quad T[u].score \leftarrow 0.5 * d + 0.5 * s$  ▷ Default case

19: $\qquad$ **else**

20: $\qquad\quad T[u].score \leftarrow M[Dissimilarity] * d + M[Sentiment] * s$

21: $\qquad T[u].review \leftarrow u$

22: $\qquad$ **if** $s > d$ **then**  ▷ Store the dominating component for choosing the suggestion

23: $\qquad\quad T[u].component \leftarrow Sentiment$

24: $\qquad$ **else**

25: $\qquad\quad T[u].component \leftarrow Dissimilarity$

26: $\quad$ Sort($T$) by $T.score$

27: $\quad$ $Suggestions \leftarrow T[0:5]$  ▷ The first five elements of candidate review list

28: $\quad$ **return** $Suggestions$

---

To balance between the two major components of the model, so that one component does not dominate the other while ranking unvisited reviews as suggestion candidates, Serendyze calculates two score modifiers ($M$), where $M[Dissimilarity]$ is the dissimilarity modifier and $M[Sentiment]$ is the sentiment modifier. When a reader visits suggestions, the modifiers track the proportion of suggestions that were primarily made for each component ($\frac{|S_Y|}{|S|}$, where $Y \in [Dissimilarity, Sentiment]$) (line 2). The primary component guiding a suggestion is set on lines 22–25.

Thus, once some suggestions have been visited, the default scoring formula is replaced with modifier values (line 20). With these modifiers, the unvisited reviews are scored in a way that ensures that one component will not dominate the scores. For example, if a reader is visiting suggestions whose scores are dominated by dissimilarity, the sentiment modifier ($M[Sentiment]$) will gradually increase in value and start to dominate the score. As a result, the reader will receive suggestions geared towards different sentiments from what they have been visiting instead of the semantic dissimilarity of visited reviews. This extension is critical for the readers to receive diverse suggestions that support serendipitous review discovery and develop an unbiased understanding of texts.
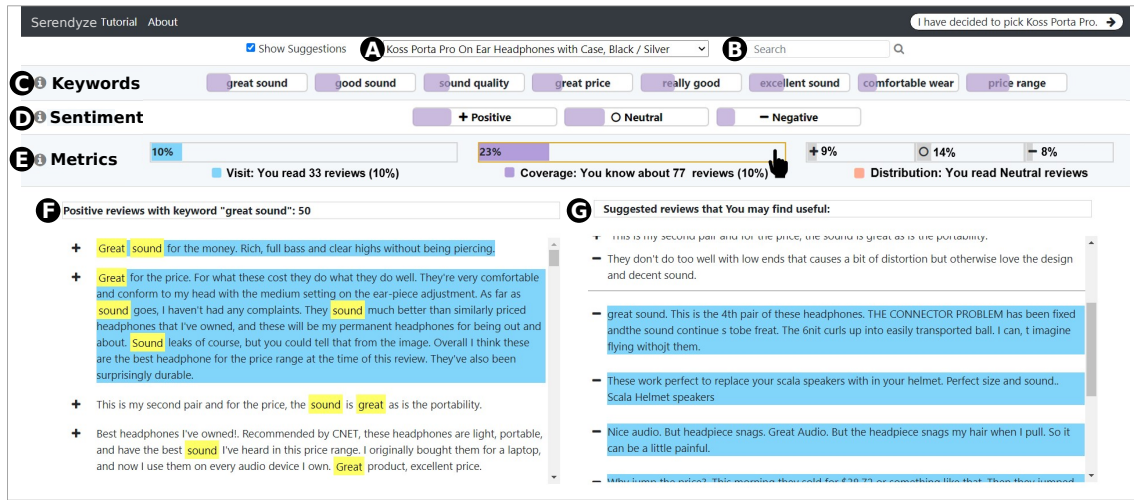
Fig. 3. Different components in the Serendyze interface: A) a dropdown option for selecting a product, B) a search bar to search for any word present in the reviews, C) a set of filters corresponding to representative keywords, D) filters for positive, neutral, and negative reviews, E) the exploration metrics - Visit, Coverage, and Distribution, F) all product reviews, and G) suggested reviews generated by the bias mitigation model that the readers may find interesting.

## 3.3 User Scenario

We present an example scenario to motivate the design and integration of exploration metrics and bias mitigation model with Serendyze. Consider Naomi, who is planning to purchase headphones for her brother as a present. She wants to find the best option within her limited budget. So, she prefers to explore headphones online with many available options and product reviews to evaluate their values. However, from her previous experiences of purchasing products online, she lacks confidence in gathering enough knowledge about different headphones to make the right decision.

Naomi decides to use Serendyze to explore headphone reviews. She starts by selecting a headphone. Then she reads several reviews and marks them as read. While looking through the suggestions, she finds one that talks about the value of the headphone given the price point. She hovers over the Coverage bar and finds out from the scented widgets embedded within the keywords that she has not visited any reviews regarding the headphone price. She uses the appropriate keyword to filter reviews that mention price. At some point during the exploration, she realizes by looking at the Distribution bar that she has been mostly visiting positive reviews. She filters the reviews by Negative and finds reviews that show the deficiencies of the headphone, balancing out her overall impression of the headphone. Since Serendyze keeps a record of her review exploration, she keeps switching between different headphones and learns more about them without the risk of losing her exploration progress. She gradually narrows down to a headphone that is best suited for her needs. She hovers over the metrics bars and sees that she has covered aspects important to her, and she has also visited a balanced distribution of positive, neutral, and negative reviews. She proceeds to purchase the headphones with confidence that she is informed enough about different headphones to make the best decision.

## 3.4 System Description

Serendyze is an interactive text analytics system designed and developed to enable readers to explore, analyze, and gather knowledge from short-free form texts, such as product reviews. We compartmentalized the Serendyze interface

into several components including a dropdown option for selecting a product (Fig. 3(A)), a search bar to search for any word present in reviews (Fig. 3(B)), a set of filters corresponding to the most frequently occurring keyword pairs (Fig. 3(C)), filters for positive, neutral, and negative reviews (Fig. 3(D)), the exploration metrics including Visit, Coverage, and Distribution (Fig. 3(E)), and finally, two sets of reviews — all product reviews (Fig. 3(F)), and suggestions generated by the bias mitigation model (Fig. 3(G)). In this section, we describe the functionalities of these components.

*3.4.1    Keywords and Search.* Serendyze extracts keywords from reviews by identifying all word pairs that co-occur at the document level, where a document is one complete review. For visual clarity, we used the top-8 most frequent word pairs as representative keywords for each product. These keywords can be used as filters to explore relevant reviews (Fig. 3(C)). Serendyze extracts relevant reviews by performing an approximate string search [8] to identify reviews that contain one or both words from the keyword pairs [15]. After filtering the reviews, it highlights all occurrences of the words present in the selected keyword pair (Fig. 3(F)).

The Search functionality is implemented as an extension of the keyword filters. The readers can use the search bar (Fig. 3(B)) to search for any word that might be present in the reviews for the selected product. Upon a successful hit, Serendyze filters the reviews based on the search query and highlights the search word in the reviews.

Serendyze is designed to be modular and customizable with the option to be outfitted with contemporary topic modeling and keyword extraction methods [54]. However, due to their probabilistic nature, potential uncertainties present in such systems might pose a threat as a confounding factor. As such, we decided to follow a deterministic and explainable method to extract keywords.

*3.4.2    Sentiments.* In Serendyze, each review is considered as an individual document and the reviews were categorized using the associated star rating at the document level. We categorized reviews that gave the product 1-star or 2-star rating as negative (−), 3-star rating as neutral (o), and 4-star and 5-star rating as positive (+). We did not use the star rating as facets, nor incorporate them into the interface directly, because previous studies have shown that when presented visually, star ratings have an undue cognitive impact compared to sentiments [97]. We do not claim that star ratings are wrong or unreliable. However, they are not appropriate to be presented visually in our study. A review with two stars might be perceived more negatively than a positive review with four stars [97]. In addition to keywords, these positive, negative and neutral sentiments associated with reviews can also be used as filters (Fig. 3(D)).

Serendyze could be outfitted with an off-the-shelf, state-of-the-art, or novel sentiment analysis method [115]. However, to avoid algorithmic misclassification and maintain transparency, we refrained from using automated sentiment analysis.

*3.4.3    Exploration Metrics.* In Serendyze, we present three interaction-driven exploration metrics — **Visit**, **Coverage**, and **Distribution** — using a set of bar charts. Readers can use these bar charts to access their data exploration patterns. We used horizontal bar charts to visualize Visit and Coverage metrics as they represent percentage values for data visit and data coverage. We represent Distribution using a set of bar charts that depict the proportion of available positive, neutral, and negative reviews visited by the reader (Fig. 3(E)). Each exploration metrics bar is annotated with an explanation of the reader's exploration patterns. For example, in Fig. 4, the text below the Visit bar suggests that the reader has explored 33 reviews which is 10% of the total reviews for this particular product and the text below Distribution suggests that while reading 33 reviews, the reader has been focusing mostly on Neutral reviews.

The Visit and Coverage bars can be interacted with in two ways. First, hovering over these bars transforms the keywords and sentiment filters into scented widgets [111], providing visual cues of exploration metrics for each keyword pair and sentiment category. For example, in Fig. 4 when the reader hovered over the Coverage bar, the keyword pairs
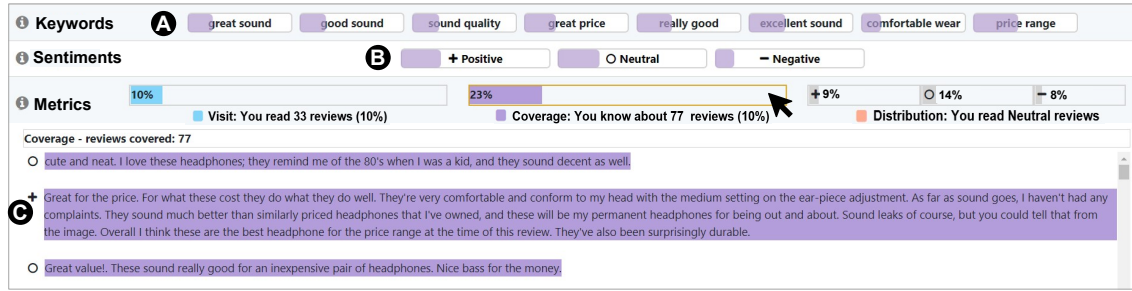
Fig. 4. Hovering over the Visit or Coverage metric bars reveals data exploration scented widgets embedded in the keyword (A) and sentiment filters (B). Here, the reader hovered over the Coverage bar, and the scented widgets show the keyword pair "comfortable wear" and the negative reviews are underexplored compared to other keyword pairs and sentiments. Clicking on the bar filters and shows reviews (C) relevant to the exploration metric selected.

(A) and sentiments (B) filled up corresponding to the reader's exploration progress at that time. Serendyze follows the visual information seeking mantra [102] to trigger the scented widgets on demand to reduce interface clutter and avoid cognitive overload when delivering visual information. The second interaction allows readers to drill down and read relevant reviews in detail by clicking on the metrics bars. For example, clicking on the Coverage bar allows readers to filter and see the reviews covered (Fig. 4(C)).

*3.4.4 Product reviews.* Serendyze provides two sets of reviews — all reviews from the selected product (Fig. 3(F)), and a set of 5 suggestions (Fig. 3(G)) generated by the bias mitigation model. The model promotes serendipitous discovery and analysis by providing readers with suggestions that introduce them to features, attributes, or other knowledge related to the selected product that they have not considered or experienced at a rate representative of the true data.

In this work, we used a heuristic bias mitigation model as proposed in 3.2 to suit our study objectives. However, we developed Serendyze as a modular and customizable platform where the bias mitigation model could be replaced with another model that could be used to generate suggestions suitable for other study tasks and domains. For example, the heuristic bias mitigation model used in this study that focuses on semantic and sentiment-wise dissimilarity could be replaced with a neural model to suggest similar, popular, or relevant reviews.

The exploration metrics and bias mitigation model rely on the reader to mark the reviews they have visited already. We enabled two ways to mark a review as read. The readers can click on any review or hover over a review to mark it as read. Previous research on user interaction with interface artifacts suggests that mouse movement is correlated with eye-tracking [26, 93, 98]. They also suggest that readers are often prone to hovering instead of clicking with interface artifacts [38], probing us to include such an alternative. The amount of time needed to hover over a review to mark it as read is dynamic and depends on the length of the review. In this work, we used a dynamic range from 1 sec to 5 sec to register the hover time to mark a review as read based on the average reading speed of adults [89] and the length of each review. When a reader marks a review as read, the bias mitigation model is called and Serendyze renders an updated set of suggestions. Based on the feedback from the pilot study, we retained the suggested reviews that are marked as read below the new suggestions in a chronologically descending order to enable users to keep track of their work (Fig. 3(G)). Serendyze saves a readers' review exploration by session. As a result, switching between products does not remove the reviews marked as read.

### 3.5 Implementation Details

We developed Serendyze as a web application with an HTML, CSS, and JavaScript front-end and a Python backend. The Doc2Vec embedding, cosine similarity, and other natural language processing functionalities, including identifying representative keyword pairs, are calculated using the gensim library [90]. During the study, the Python scripts were hosted in a freely available server [1]. Upon interaction with the reviews, the front-end fires a request with the list of visited and unvisited reviews, and the server returns the coverage, distribution, and suggestions.

The scalability of the system is dependent on the number of reviews per product. We stress-tested the system with over 10000 reviews across 10 products, each containing over 1000 reviews. Measured over 100 attempts, it takes Serendyze an average of $3.11 \pm 0.85$ seconds to return suggestions for a product with 1000 product reviews. We performed the tests on a laptop with an Intel Core i5 7th generation processor (7300HQ) and 8 gigabytes of RAM, running on localhost. The source code is publicly available for viewing[2].

### 3.6 Pilot Study

Before deploying Serendyze in the real world to study how people use the system to explore online product reviews, we performed a pilot study simulating the same experience with 12 participants (Pi-1 to Pi-12). We recruited participants (8 males and 4 females, $28 \pm 4$ years of age on average) using word of mouth and email across different countries. The goal of the two-week-long pilot study was to simulate and assess the system workflow, identify potential interface issues, and whether participants could use the functions provided in Serendyze to explore the data comprehensively.

This pilot study helped us to better realize and solidify operational procedures to perform real-world deployment of Serendyze. Based on the feedback from the pilot study, we modified the system interface and tuned the threshold values for similarity and distribution. We modified the interface by revising the Distribution visualization and used three distinct bars to represent the true distribution of positive, neutral, and negative reviews instead of one aggregate value. We also added functionalities to display the already visited suggestions in chronologically descending order below the newly generated suggestions. Finally, we fixed several interaction issues, including adding a loading symbol to provide visual feedback that the bias mitigation model is generating new suggestions. We also adjusted the hover time needed to mark a review as read and made other small improvements.

## 4 EVALUATION

To evaluate the viability of supporting serendipitous discovery and analysis of product reviews using exploration metrics and bias mitigating suggestions, we performed a user study with 100 crowd workers. The study was approved by the institutional review board. In this section, we explain the study conditions, participants, procedure, and findings.

### 4.1 Conditions

The study was between subjects with four conditions, as presented in Fig. 5. Condition B is the *baseline*, condition M is the Serendyze version with exploration *metrics* only, condition S is the Serendyze version with the *suggestions* only, and condition M&S is the Serendyze version with both *metrics and suggestions*. Each of these conditions has a set of basic components in common — the option to select products, the representative keywords, the positive, neutral, and negative sentiment categories, and the reviews. All conditions enabled users to filter reviews based on keywords and/or sentiments and mark reviews as read by clicking or hovering on them.

---

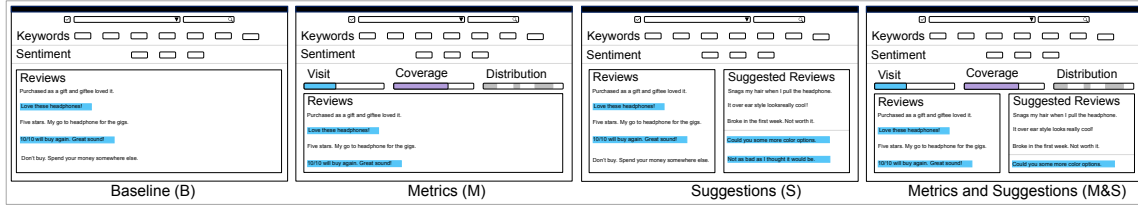[2]https://osf.io/jmqx2/?view_only=144115224a204dea8e2104cb829b9606

Fig. 5. This figure depicts all Serendyze components and which features were available with the four conditions (B, M, S, and M&S). The dropdown for product selection, search options, keywords and sentiment filters, and reviews were available for all four conditions (B, M, S, and M&S). The exploration metrics (Visit, Coverage, and Distribution) were only available for conditions M and M&S. The suggestions generated by the bias mitigation model were only available for conditions S and M&S.

We designed conditions M and S to remove confounding factors by evaluating features independently. Condition M&S is the culmination of the Serendyze system with all functionalities.

## 4.2 Participants

We recruited crowd-worker participants through Amazon Mechanical Turk [23], a popular crowdsourcing platform used to conduct studies requiring human intelligence. All of our participants were *Amazon Master Workers* who received the qualification for consistent demonstration of a high degree of success in performing a wide range of tasks across many requests. 25 Master Workers were assigned to each condition. Each participant was compensated with USD $15.

We asked participants to fill out a pre-study questionnaire to help us understand their online shopping practices and preferences. The response of the questionnaire suggested diverse shopping practices across our participants. For example, 1 participant had never purchased a product online, 37 participants purchased 1–5 products per week, while 31 participants purchased 6–10 products, and another 31 participants purchased more than 10 products. Furthermore, 22 participants spent less than 10 minutes, 29 participants spent 10–20 minutes, 22 participants spent 20–30 minutes, and a final 27 participants spent more than 30 minutes reading product reviews before making purchase decisions. 8 participants were a little dependent on product reviews while 43 participants were moderately and 42 participants were significantly dependent. Another 7 participants were completely dependent on product reviews.

## 4.3 Dataset

We used a subset of publicly released Amazon product reviews [39] as an example corpus of online short free-form texts. Among numerous products, we selected headphones as the candidate due to their ubiquitous usage [104]. Among thousands of headphones in the dataset, we selected three random headphones that had over 5000 reviews each, with an average star rating between 4.5 and 4.6. We chose these conditions to select popular headphones that are not obviously superior or inferior to each other. These three headphones were Koss Porta Pro, Sony MDRV6 Studio, and Sennheiser HD280PRO. We removed all reviews that contained HTML content or languages other than English. As mentioned in Section 3.4.2, we assigned these reviews to positive, negative, and neutral sentiments based on the associated star ratings. To keep the number of reviews reasonable for the study participants to read and make decisions, we randomly sampled 120 reviews each for positive, neutral, and negative sentiments for each headphone, aiming for approximately 1000 reviews in total. Based on previous work [112] and Amazon's product review guidelines [6], We then removed reviews that are less than 10 words long and more than 100 words long to maintain the length of the reviews at a reasonable level — suitable for the participants to read and make timely decisions. Finally, we ended up with 880

reviews with 338 reviews for Koss Porta, 277 reviews for Sony, and 265 reviews for Sennheiser. There were 340 positive, 277 neutral, and 263 negative reviews in total across all three headphones. Our sampling did not follow the actual distribution of sentiments present based on star-ratings, since the headphones chosen were rated mostly positively, and that would result in a dataset with too few neutral and negative reviews. This would make it difficult to reasonably study participants' review exploration across different sentiments. Rather, the dataset was constructed to match the study design so that participants could not immediately distinguish among three headphones based on sentiment distribution and they had to rely on reading reviews to make their decision. We used the same dataset with all four conditions.

## 4.4 Procedure

We asked participants to explore reviews of each of the three headphones using assigned versions of Serendyze and make a decision to refer one of the headphones to someone they know. We asked them to recommend one of the headphones to others instead of buying for themselves in an attempt to motivate participants to learn about these headphones beyond personal preferences. We randomized the procedure of assigning conditions to participants by providing a single link to all crowd workers who participated in the study. This link would then redirect the participant randomly to one of the four conditions. We also kept a record of studies performed with each condition, and when a condition reached 25 studies, we randomized the remaining redirections to the conditions still not exhausted.

Each study procedure began with participants' agreement to sign the consent form. After signing the consent form, participants were asked to answer a pre-study questionnaire that asked questions about their prior online product review exploration and purchase experiences, including the time spent, number of reviews read, and products purchased. We also asked their headphone brand or feature preferences to see if their preferences influenced their decisions.

After the pre-study questionnaire, we directed the participants to the tutorial section, featuring a required video explaining the procedures and functionalities of the Serendyze condition assigned to them. These videos lasted up to 3 minutes, based on the condition. Participants could rewind the video but could not skip forward. At the end of the video tutorial, the participants proceeded to the study task. An extended tutorial with annotated figures was also provided to participants and was accessible anytime from the navigation bar. In both tutorials, we only presented the features and functionalities for the condition's components. We did not disclose the goal of our study or demonstrate any pre-defined exploration patterns to avoid biasing participants' review exploration.

We instructed the participants to thoroughly read the reviews for all three headphones during the study and decide on a product to refer to others. They were also instructed to spend at least two minutes on each headphone. During these instructions, we did not inform the participants about the goal or hypotheses of the study and did not provide them with any pre-defined exploration patterns or hints. In contrast, the participants were instructed to explore the reviews in any way they wanted using the features provided in their study condition to make recommendation decisions. After they were decided, we asked them to finalize their decisions and proceed to the post-study questionnaire.

In the post-study, we asked participants open-ended questions to learn about their experiences using Serendyze to explore reviews before decision-making. In the post-study questionnaire, we added attention checks to identify whether the participants' answers match their activity during the study. We also asked them questions about their usage of exploration metrics and suggestions, their ease of use, how useful they found exploration metrics and suggestions, and how they utilized them while exploring reviews. Furthermore, we asked them what they liked and disliked and what issues they faced while working with Serendyze. For both the pre- and post-study questionnaires, we asked the participants to answer all questions and ensure they passed the attention checks in order to be compensated. The attention checks included questions to verify if the participants could recall information about the headphones.
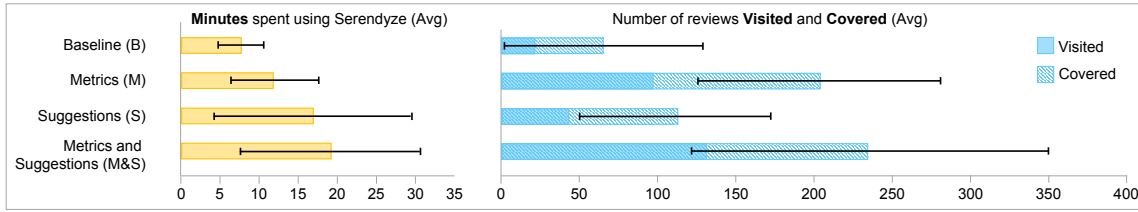
Fig. 6. Statistics on how long participants used Serendyze and their knowledge about the data. Participants who used condition M&S spent the most time (19.05 ± 11.46 minutes on average) reading reviews with Serendyze. Participants who used condition M&S visited the most reviews (130 ± 88.87 reviews on average) and had explicit knowledge about these reviews. Participants who used condition M&S covered the most reviews (234 ± 113.9 reviews on average) and have implicit knowledge about these reviews.

## 4.5 Data Collection and Analysis

We collected usage logs containing the participants' timestamped interactions with all components of Serendyze and stored them for later analysis. We collected the participants' responses to the pre- and post-study questionnaires, the time they spent answering the questions, and the time they spent on the study. We embedded the questionnaires with the study platform so that the participants did not have to traverse multiple websites to participate in the study.

We analyzed the collected data both quantitatively and qualitatively. We used parametric and non-parametric inferential statistics, to analyze the quantitative data. We analyzed the qualitative data collected from the responses to pre- and post-study questionnaires using an open-coding method [12]. Two annotators independently coded the questionnaire data collected from condition M&S as it contains all interventions and potentially most variable data. The annotators discussed and reached an agreement to consolidate their codes into a representative set of codes. The inter-coder reliability measured using Krippendorff's alpha [61] was 0.86. Based on these codes, one annotator annotated the remaining data collected from conditions B, M, and S and the other annotator verified the annotations. The data collected from pre- and post-study and the codes for qualitative analysis is provided as supplementary materials.

## 4.6 Findings

Our research questions investigate whether supporting serendipitous discovery helps readers to explore reviews more comprehensively, how their exploration behaviors change with access to their exploration patterns, and how suggestions of unexplored reviews impact their decision-making. We formulated the following hypotheses to answer these questions:

(1) **H1 - Comprehensive**: Participants who had access to the exploration metrics will cover more reviews.
(2) **H2 - Unbiased**: Participants who had access to the exploration metrics will read a more balanced distribution of reviews.
(3) **H3 - Confident**: Participants who had access to both the exploration and suggestions will have greater confidence in their decision.

To evaluate our hypotheses, we analyzed the collected quantitative and qualitative data from 100 crowd workers across four different conditions (B, M, S, M&S). We present the findings from our analysis in this section.

**H1: Participants who had access to the exploration metrics covered more reviews.** We collected the number of reviews across different products that the participants had covered explicitly or implicitly. We posited that the participants had explicit knowledge about any review that they visited, and implicit knowledge about all reviews that

Table 1. Analysis of reviews covered across conditions. The results of a two-way ANOVA where Conditions and Products are the independent variables and the number of reviews covered is the dependent variable shows that there is a statistically significant difference ($p<.05$) across conditions but no significant difference across products nor the interaction between conditions and products. Post-hoc Tukey tests indicate a statistically significant pairwise difference in the average number of reviews covered among all pairs of conditions where one condition provides exploration metrics, and the other does not.

| Factors | Degree of Freedom | Mean Sum of Squares | F-value | Pr(>F) |
|---|---|---|---|---|
| Conditions | 3 | 51042 | 39.21 | $2e^{-16}$ |
| Products | 2 | 3177 | 2.44 | 0.09 |
| Conditions : Products | 6 | 759 | 0.58 | 0.75 |

| Condition Pairs | Difference | Lower-bound | Upper-bound | P-adjusted |
|---|---|---|---|---|
| Metrics (M) - Baseline (B) | 45.93 | 30.71 | 61.16 | $2e^{-11}$ |
| Suggestions (S) - Baseline (B) | 15.69 | 0.47 | 30.92 | .04 |
| Metrics and Suggestions (M&S) - Baseline (B) | 56.16 | 40.93 | 71.39 | $5e^{-12}$ |
| Suggestions (S) - Metrics (M) | −30.24 | −45.47 | −15.01 | $2e^{-11}$ |
| Metrics and Suggestions (M&S) - Metrics (M) | 10.22 | −4.99 | 25.45 | .31 |
| Metrics and Suggestions (M&S) - Suggestions (S) | 40.47 | 25.24 | 55.69 | $4e^{-8}$ |

were semantically similar and redundant to the reviews they had explicit knowledge about. Fig. 6 presents the number of reviews covered on average across all conditions and suggests that the participants who used conditions (M and M&S) with exploration metrics covered more reviews on average ($203 \pm 77$ and $234 \pm 114$) compared to the conditions (B and S) without exploration metrics ($66 \pm 64$ and $113 \pm 61$). The average coverage to time-spent ratio for conditions B, M, S, and M&S are 8.64, 17.28, 6.72, and 12.28, respectively. These ratios show that the participants who used conditions M and M&S had a higher average coverage to time-spent ratio than those who used conditions B and S. A higher coverage to time-spent ratio suggests that the participants spent less time covering more reviews. As such, our results suggest that the participants who had access to exploration metrics covered more reviews efficiently, by spending less time to gain more knowledge about the products.

To evaluate **H1**, since the coverage value for all conditions passed the Shapiro-Wilks test, we performed a two-way ANOVA test with 2 factors — conditions and products. Table 1 presents the results of the test. The results suggest that there is a statistically significant difference of average reviews covered by participants across conditions. Furthermore, there are no statistically significant differences among different products and the interactions between the conditions and the products. A Tukey posthoc test revealed statistically significant pairwise differences between the conditions that provide exploration metrics (M and M&S) and the conditions that do not (B and S). We posit that the pair Metrics and Suggestions (M&S) - Metrics (M) are not statistically significant because they both provide exploration metrics. Based on the results, we conclude that people with access to exploration metrics covered more reviews, and consider **H1** to be supported. The participants' responses also corroborate these results. Participants in condition M explained how they used the exploration metrics: P31 mentioned, "*I used exploration metrics to review enough to make a satisfying estimate of value of each product. [Coverage] showed me which reviews I hadn't covered yet, so I could read [those reviews] to get a better opinion.*" Another participant (P45) said, "*I wanted to read a good amount of all kinds of reviews – positive, neutral, and negative. The Distribution helped me see if I was doing that. [. . . ] I also used Coverage to save time so I didn't read too many redundant reviews.*" The participants who used condition M&S highlighted how having access to exploration metrics helped them read reviews more comprehensively before making decisions. P90 said, "*The exploration metrics helped me to see what percentage of reviews I had really read to see if I was getting a full picture or not. [. . . ] It helped me to*

*make sure that I know about enough reviews before making a decision.*" Another participant (P85) mentioned, "*I used exploration metrics to make sure I was looking at the various types of reviews, including neutral and negative. I wanted to make sure I had read enough of each of these types to make a final decision.*"

**H2: Participants who used exploration metrics read positive, neutral, and negative reviews in a balanced way.**
For each participant, we collected the number of positive, neutral, and negative reviews they visited. The distribution of the percentage of positive, neutral, and negative reviews visited by participants is presented in Fig. 7. The figure indicates a notable difference between conditions. The participants who used conditions M and M&S with exploration metrics visited reviews from all sentiments in a balanced manner, reflecting the true distribution of sentiments present in the dataset used. Among 2411 reviews visited by participants in condition M, on average, each participant visited 35 ± 24 positive, 27 ± 15 neutral, and 33 ± 19 negative reviews. The participants in condition M&S visited 3268 reviews where on average, each participant visited 49 ± 44 positive, 37 ± 24 neutral, and 43 ± 23 negative reviews. In contrast, the participants in condition B visited 521 reviews. On average, 12 ± 13 were positive, 4 ± 8 were neutral, and 5 ± 7 were negative reviews. Finally, among 1067 reviews visited by participants of condition S, on average, each participant visited 11 ± 11 positive, 9 ± 7 neutral, and 17 ± 14 negative reviews. Fig. 7 shows that without access to exploration metrics, participants visited more positive (B) or more negative (S) reviews. Despite not mentioning the term "balanced" during the tutorial, task assignment, and questionnaires, the qualitative responses for participants suggest that the exploration metrics helped participants explore reviews in a more *balanced* way — as termed by the participants. 15/25 participants who used condition M and 18/25 participants who used condition M&S mentioned that they used exploration metrics to balance out how they were reading different kinds of reviews so that they did not lean towards one specific sentiment. P28 (condition M) mentioned, "*I did not want to look at only negative reviews. I used the exploration metrics to make sure I was reading enough reviews of each type.*" P49 remarked, "*I made sure that I have read a fairly even distribution of all kinds of sentiments, not just positive or negative.*" P100, who used condition M&S mentioned how the exploration metrics enabled them to notice an imbalance in their work and seek out other reviews: "*I noticed that I had read a lot of positive reviews, so I then read some neutral and negative reviews to balance it out so I got a fuller picture.*" P77 said, "*It was good to be aware that I was viewing a range of review types, and not solely focusing on only positive or only negative reviews.*" These findings suggest that our intervention enabled participants to overcome *oversensitivity to consistency* [42, 108] and allowed them to explore all sentiments. However, 4/25 participants who used condition M and 3/25 participants who used condition M&S decided not to use exploration metrics. While 2 of them (P32, P84) mentioned that they "*did not feel the need to*", 4 other participants mentioned (P36, P44, P50, P88) that they, "*explored the data on their own, using their own strategy*".

**H3: Participants who used both exploration metrics and suggestions were confident in their decisions.** To understand how exploration metrics and bias mitigating suggestions impact participants' decision-making process, we asked each participant how confident they were about reading enough reviews and making the right decision in the post-study questionnaire (See Figure 8). 76% of the participants (19/25) in condition M&S were highly or completely confident that they had read enough reviews. Furthermore, 64% of the participants (16/25) in condition M&S and 76% of the participants (19/25) in condition S were also highly or completely confident that they had made the right decision. However, we did not find similar high confidence among participants who used conditions B or M. This result suggests that bias mitigating suggestions might have played a role in invigorating participants' confidence in reading enough reviews and making the right decision.

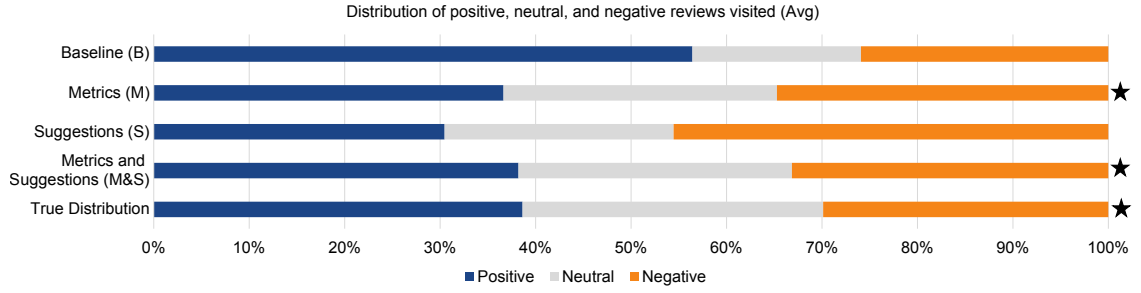Distribution of positive, neutral, and negative reviews visited (Avg)



Fig. 7. This figure depicts the average distribution of positive, negative, and neutral reviews as visited by participants. The figure suggests that when participants in conditions M and M&S, annotated by (★), explored different sentiments in a balanced manner, which is reflective of the true distribution of sentiments present in the dataset. However, in the other two cases, we see an imbalance where participants in condition B visited a relatively larger number of positive reviews, and the participants in condition S visited a relatively larger number of negative reviews.
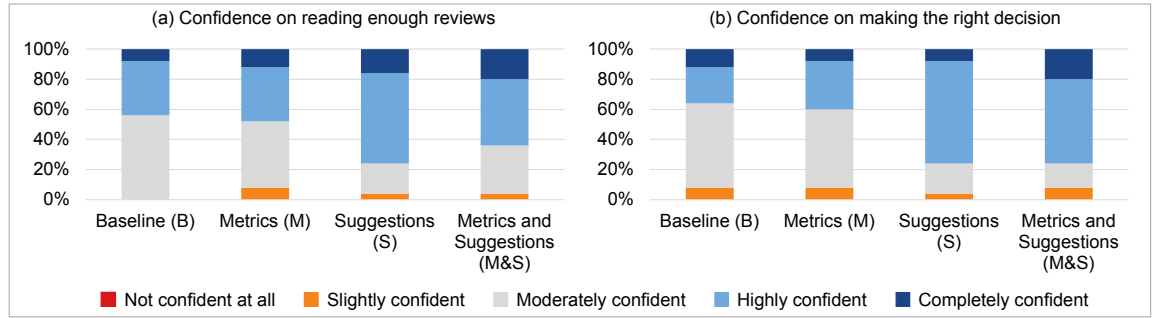


Fig. 8. This figure shows participants' confidence (a) in reading enough reviews prior to decision-making and (b) in making the correct decision for all Serendyze conditions (B, M, S, M&S). The figure suggests that in both cases, for conditions S and M&S, they were more confident in reading enough reviews and also on the decisions that they made compared to other conditions.

Table 2. This table presents the pairwise Mann-Whitney U test results between condition M&S and the other three conditions (B, M, and S). The results indicate a statistically significant pairwise difference in the confidence among condition M&S and conditions B and M at an alpha of .05. While condition M&S provides both exploration metrics and suggestions based on the bias mitigation model, conditions B and M do not provide suggestions.

| Condition Pairs | z-score | p-value |
|---|---|---|
| Metrics and Suggestions (M&S) - Baseline (B) | −2.11 | .03 |
| Metrics and Suggestions (M&S) - Metrics (M) | −2.11 | .03 |
| Metrics and Suggestions (M&S) - Suggestions (S) | −0.48 | .62 |

To evaluate **H3**, we performed a Kruskal-Wallis ranked sum test, which is a non-parametric test on the distribution of confidence level among participants, since confidence level distribution for all conditions failed the Shapiro-Wilk normality test. The results suggest a statistically significant difference of confidence among the participants who used different conditions (B, M, S, and MS). For four conditions, the degree of freedom was 3, the critical Chi-Squared value was 10.05, and the p-value was $p = .02 < .05$. Hence, we rejected the null hypothesis and followed up by performing a pairwise Mann-Whitney U test with condition M&S against other conditions (B, M, and S). The result of this test
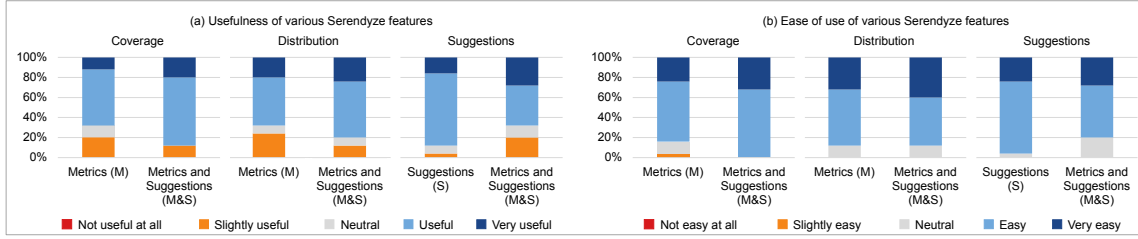
Fig. 9. This figure depicts how the participants perceived the (a) usefulness and (b) ease of use of Coverage, Distribution, and suggestions based on the conditions where the participants had access to these features. The figure suggests that the usefulness of all features is over 68%, and ease of use is over 80% across all relevant conditions.

is presented in Table 2. The statistically significant pairs are highlighted in boldface with a gray background. The results indicate a statistically significant difference in the confidence of participants on making the right decisions between those who used condition M&S compared to the participants who used condition B ($p = .03$) or M ($p = .03$). However, there is no significant difference among the confidence of participants who used condition M&S compared to participants who used condition S ($p = .62$). This lack of statistical significance further supports the observation that the suggestions based on the bias mitigation model might influence participants' confidence when making decisions. Based on these results, we consider **H3** to be partially supported. Although we did not account for it in the study, the pre-study questionnaire suggest that 8/25 participants had either Sony (3 participants) or Sennheiser (3 participants) brand preference. However, the post-study questionnaires suggest only 1 participant from each group decided to select the headphone of their preferred brand, further suggesting that participants could overcome *persistence of impressions based on discredited evidence* [42, 108] and make more confident decisions that do not mirror their preconceived preferences.

Participants' feedback also suggests that using exploration metrics and suggestions improved their confidence in their decisions. P87 (condition M&S) mentioned, "*...[Serendyze] allowed me to process a lot of information quickly. I could search for a specific feature for each headphone product and feel confident about it because of the large amount of positive and negative reviews.*" P83 contrasted their experience of using Serendyze with their regular product review patterns, saying, "*I appreciated the Distribution a lot. I am very guilty of reading reviews that back up my existing opinion - justifying a purchase rather than really learning about the product... [Exploration metrics] helped me avoid that.*" P93 mentioned how suggestions helped them understand reviewers' perspectives, saying, "*It was a quick way for me to see how others felt about the products, and they gave me more information based on the other reviews that I have already read.*" P81 also mentioned how suggestions enabled them to make the right decision by helping them compare between products: "*The suggested reviews let me pro and con better as I made my decision.*"

**Participants found suggestions to be useful but had mixed feelings about unexpected suggestions.** During analysis of the post-study questionnaire, we found that the participants found suggested reviews useful for making decisions. Fig. 9(a) presents how participants perceived the usefulness of different features provided in Serendyze. The figure suggests that 80% of participants (20/25) who used condition S and 68% of participants (17/25) who used condition M&S found the suggestions to be useful or very useful. Responses from the post-study questionnaire suggest suggestions also impacted participants' decision-making (15/25 for condition S and 13/25 for condition M&S).

We used qualitative responses collected from participants to understand why and how the suggestions helped participants before decision-making. A considerable number of participants in conditions S (13/25) and S&M (12/25)

believed suggestions helped them to gain deeper knowledge from reviews. For example, P65 mentioned, "*I decided to choose the headphones that I chose because they had high marks in regards to audio quality based off of the suggested reviews that I was shown.*" Furthermore, 9/25 participants who used condition S and 6/25 participants who used condition S&M highlighted that suggestions helped them find unexpected information that they may not have thought about yet. P90 (condition M&S) mentioned, "*I liked that it was different from what I was reading. They [suggestions] helped me back up my opinion of the product*". P60 said, "*I was suggested reviews I didn't see in the regular one, especially the ones around comfort and styles. I didn't think about those at the beginning*". Other participants (5/25 from S and 9/25 from M&S) mentioned that they found the suggested reviews useful for gaining perspectives on opposite opinions and combat biased exploration. P77 mentioned how they leveraged suggestions to get opposite viewpoints, saying, "*I would look to the suggestions when I was done reading a particular review and I wanted to read the opposite view. It helped to make sure I didn't get biased towards a product*".

However, not all participants preferred the unexpectedness of the suggestions generated by the bias mitigation model. It is should be mentioned that during the introduction, tutorial, task assignment, and pre-study questions, we did not mention to participants how the suggestions were generated. In the conditions (S and M&S) where participants had access to suggested reviews, the interface provided reviews as only the suggested reviews *they may like* (see Fig. 3). Diving deep into the responses of participants who were taken aback by the unexpected suggestions, we found that they often considered the suggestions unhelpful or not detailed enough. P76 mentioned, "*It [suggestions] gave me unhelpful reviews that made me not to trust the system.*" P61 explained, "*suggested reviews were mostly comprised of short reviews [. . .] that could have been easily generated by bots, so I didn't trust them.*"

Some other participants didn't find the suggestion useful because they thought that "*it did not do a good job*" (P76), or they "*could not figure out why they [suggestions] are being suggested*" (P80). These responses suggest that some participants may have different expectations from the suggested reviews feature when using Serendyze. These expectations might be an artifact of a priming effect. People often expect recommendations or suggestions to be similar to what they are exploring rather than offering diversity [57]. We further discuss this observation and its ramifications in Section 5.

**Participants heavily Used Serendyze interventions to perform text-level analysis of reviews.** Figure 9 suggests that the majority of participants who had access to exploration metrics and suggestions found them to be useful and easy to use. However, we wanted to explore deeper and learn how the participants leveraged these features to learn more from the data prior to decision-making. We used time-stamped interaction logs to analyze participants' use of Serendyze features to model their exploratory and decision-making strategies (see Fig. 3). We should emphasize that while designing the tutorial, task description, and pre-study questions, we paid careful attention to not bias participants towards using any particular feature. The product review task, identical across all conditions, asked participants to explore the reviews and decide which headphones they would refer to someone.

Four node-link graphs in Figure 10 show the participants' usage of and transitions between the six primary components of Serendyze — product selection, keywords, sentiments, exploration metrics, reviews, and suggestions —for conditions B, M, S, and M&S). The total number of interactions were 970 for the baseline condition (B), 4452 for the condition with exploration metrics only (M), 2137 interactions for the condition with suggestions only (S), and 5519 for condition M&S, which contains both exploration metrics and suggestions. It is worth mentioning that the operational granularity of these interactions is not symmetric. For instance, from an interaction perspective, interacting with a keyword might impact a set of reviews but interacting with a review impacts only that particular review.

Fig. 10. Node-link diagrams that show how participants interacted among the six major Serendyze components across all conditions — product selection, keywords, sentiments, exploration metrics, reviews, and suggestions. Arrows depict a transition from one component to another. For brevity and clarity, the connections between two components are shown only when the number of interactions among them is more than 1% of all interactions for that condition. The overall use of components is double encoded in the border thickness and background saturation. The orange lines show connections with a higher frequency of interactions. For brevity and clarity, we only draw connections between two components if the number of interactions among them is more than 1% of all interactions for that condition. The figures suggest that participants heavily used exploration metrics and suggested reviews whenever available.

Fig. 10(a) suggests that the participants who used condition B often used keywords and sentiments to filter reviews. The majority of the participants (13/25) preferred the cascaded filters to filter reviews by a keyword and a sentiment in conjunction. One participant (P12) mentioned, "*I loved the keyword filter to focus on what was important to me and I could then see exactly how many positive and how many negative for that particular keyword. That was amazing!*" The

participants (10/25) also preferred the option to search custom keywords and the highlighting of the searched keyword on the reviews. P19 said, "*I really liked the way that you could easily search for sentiments. For example, I cared most about reasonable pricing and sound quality. It was easy for search and have those things highlighted in individual reviews.*"

Fig. 10(b) suggests a prominent interaction trend where the participants transitioned between exploration metrics and reviews. There are also traces of interactions from exploration metrics to sentiments. In essence, the exploration of reviews by participants who used condition M revolved around exploration metrics as 15/25 participants mentioned they used exploration metrics to balance the types of sentiments they were exploring from the reviews. We found similar trends in condition S (Fig. 10(c), where participants explored both the suggested and regular reviews and went back and forth between them. This observation is also supported by the participants' post-study feedback soliciting how they used the suggested reviews. 13/25 participants mentioned that they started reading regular reviews; after a while, they started to read suggestions and kept alternating between the two. Finally, in Fig. 10(d), we see that the trends from conditions M and S repeat with the participants having access to all features of Serendyze. These diagrams suggest that the participants heavily used the proposed interventions to perform their tasks whenever available.

The qualitative responses we collected from the participants also reflect their desire to use exploration metrics and suggestions. For instance, P93 mentioned, "*Serendyze provides a nice collection of useful information and features in order to compare products. It is something I would use while looking for products online.*" P61 highlighted how suggested reviews helped them to keep track of reviews that were important,

> *The feature that I liked the most was the tracking of the "Suggested reviews that you have visited already."*
> *It was really convenient for me to keep track of the reviews that had made an impression on me. It was*
> *really easy to add reviews there, and this was important because I've found that it is easy to lose track of*
> *specific reviews with a really important detail that was not in other reviews, and this feature is a great*
> *solution to prevent losing track of important or personally useful reviews.*

P30 mentioned how exploration metrics helped them to learn how much data they have explored, "*I used these features [exploration metrics] now and then to get an idea of what information I had already covered. I found them to be very interesting and something I'd like to see on all review pages!*" The responses for the aesthetics of the Serendyze interface were mixed. While some participants (P44, P49) preferred the "*color coded interface*", others (P43, P29) though the interface was "*plain and needed more color and designs.*"

## 5 DISCUSSION

The findings from the evaluation of Serendyze suggest that exploration metrics helped participants to explore reviews more comprehensively and cover more reviews compared to systems without such interventions. It also allowed them to balance their text exploration across all sentiments and perspectives present in reviews. We also found that the bias mitigation model provided participants with useful suggestions, enabling them to gain deeper knowledge about the products and raising their confidence in making informed decisions. Participants also found unexpected suggestions that often impacted their decision by providing evidence. However, some participants did not find suggestions useful due to expectation mismatch. Overall, the evaluation highlights the necessity for text-level exploration functionalities to complement summary-level exploration and readers' eagerness to leverage such functionalities for data-driven decision-making. In this section, we further unfold the findings from the evaluation, discuss implications of such findings, and speculate how knowledge gained from this work can be integrated to improve and evolve current approaches.

### 5.1 The Impact of Preconception and Expectation Mismatch on Accepting Suggestions

The findings from our evaluation of Serendyze demonstrate that the participants in conditions S and M&S found suggestions to be useful (see Fig. 9(a)) and a catalyst for increasing confidence in making decisions (see Fig. 8(b)). However, some participants (4/25 each for S and M&S) did not feel that the suggestions were useful as they expected suggestions to be similar to the reviews they visited. These participants who were seeking similar reviews from the suggestions were disillusioned and felt disconnected, leading towards their eventual disinterest and discontinuation of reading suggested reviews. In contrast, others who embraced the unexpectedness found the suggestions to be serendipitous, compelling, and conducive to gaining insights from the reviews.

In this work, we experimented with a heuristic bias mitigation model to generate suggestions that are most dissimilar to what participants visited — both in terms of semantic similarity and the sentiments associated with the reviews. Such approaches that focus beyond accuracy metrics to measure the utility and performance of generating suggestions are relatively new [57, 101]. Similar to Serendyze, the inner workings of the majority of the available suggestion-generation systems, such as Amazon product recommendations, Facebook and YouTube content recommendations, and Netflix and Spotify entertainment recommendation, are hidden from the users. Such systems often rely on similarity or relevance among user [99] or product attributes [69] based on a users' exploration patterns to suggest new data [57]. These available systems might play a role in priming and shaping user perceptions and heavily impact how and what the users expect from the suggestions generated by an automated system [32, 56, 58]. The participants who actively sought similar reviews did not perceive that the suggestions were reaffirming their gained knowledge and this expectation mismatch likely resulted in their disconnection from using the dissimilar suggestions.

To address such issues, systems such as Serendyze could be designed to have the functionality to alternate between providing suggestions that are homogeneous to what the readers have been reading and suggestions that are dissimilar to what they have been reading. The first approach might help them reaffirm their decisions based on what they read, and the second approach can help them gather serendipitous, broader, and diverse knowledge. Such duality might provide flexibility for users to explore the data as they prefer. Furthermore, more clarifications can be added to explain how suggestions are generated. To that end, visual indications of scores and ranks associated with suggestions [85] might improve the readers' perceptions towards suggestions, provide transparency, and combat confusion [94].

### 5.2 Perception of Detailed Reviews and their Impact on Trust in the System

Our evaluation demonstrated how suggestions enabled the participants to gather knowledge from opposing perspectives, access reviews that they had not thought about beforehand, and make confident decisions. While suggestions were heavily used by participants (Fig. 10(c) and (d)), some participants were not pleased with suggestions as they did not find them to be helpful. For some of these participants, the unhelpful suggestions were sufficient to induce mistrust in the system, and they stopped using suggestions. Digging deeper, we made two observations: (1) some participants perceived that the suggestions were too brief and did not have enough information to gather deeper knowledge; (2) some considered the lack of details in suggestions to be artifacts of random text generating agents or *"bots"* (P61).

People have a complex relationship with how they interact with and use a system, what data they get out of the system, how they interpret such data, and how they establish trust in the system [18]. While mistrust of a system can often derive from unexpectedness and uncertainty, the quality of the data provided, and the manner in which they are provided, may also have a role to play in how users perceive the data presented to them [30]. Serendyze generates suggestions based on the dissimilarities in semantics and the sentiments associated with each review. While capable of

generating reviews that can enable readers to diversify their knowledge gained from the reviews via serendipitous discovery and analysis of interesting reviews, the bias mitigation model is not designed to assess the quality of the reviews being suggested. Adding quality assessment functionalities to measure the quality of generated reviews could help mitigate the mistrust induced by suggested reviews in text analytics systems like Serendyze.

One approach to assess whether the suggestions contain details that might be desirable to the readers could be to modify the bias mitigation model to identify and leverage the latent aspects present in the reviews [29, 88]. For instance, during the scoring of candidate suggestions (see Algorithm 1), the model could assess whether certain aspects for a headphone, such as price, longevity, value, sound quality, or other aspects desirable to the reader, are present in the candidate suggestion. Furthermore, the presence and absence of the desired aspects could be provided to readers using visual cues [51, 95] to help readers decide whether they want to read the suggestion. Providing cues to missing aspects could also benefit readers by helping them assess whether the suggestions are useful to suit their needs [96].

## 5.3 User Agency and Trust in Mixed-Initiative Systems

For evaluating Serendyze, we asked participants to perform an open-ended task of reading reviews and making a purchase decision. The participants were free to approach the task any way they wanted, and they used Serendyze organically to complete their tasks without provocation to use specific features. However, some participants who used condition M&S (4/25) chose to depend solely on exploration metrics. They justified their choice by highlighting their preference to not be influenced by machine-generated and algorithmically-curated suggestions and wanted to "*analyze the data on their own*" (P98). They also felt that the visualization of the exploration metrics was the result of their tangible interactions with the system, and the metrics mapped and presented their behavioral patterns without manipulation. Such observations open up bigger questions around user agency in mixed-initiative systems such as Serendyze, where users and automated systems work in tandem to achieve a goal [44].

User agency is a critical concept in mixed-initiative systems, and in human-computer interaction, in general [117]. It often dictates whether the users will adapt to the functionalities provided by the system [74]. A user is considered to have user agency when they perceive that they are responsible for their interactions with the system and they own the consequences of their actions on the system [22, 117]. Our evaluation suggests that the exploration metrics with the visual cues based on scented widgets [111] enabled participants to feel in control of their exploration process by allowing them to follow their own review exploration strategy. This observation aligns with previous works where data visualization has been shown to be effective in conveying information regarding exploratory analysis and open-ended tasks [50] to instill a sense of transparency and trust in users [30]. In contrast, algorithmic curation of online text, including news articles, social media posts, and reviews, is often associated with a lack of transparency and is conducive to generating mistrust in users due to their closed nature as black-box solutions [25, 28].

Due to this research area being under-explored, it is challenging to definitively design solutions that balance user agency in mixed-initiative systems. However, the debate remains on how to address the volatility of user agency in mixed-initiative systems with inevitable black-box components and algorithmically curated system responses. One might argue that the system should enable users to have agency and have the capability to support users' choices of rejecting features that they do not feel comfortable adapting to. The others might advocate providing additional features and guidance to make automated systems more transparent [14]. These questions and viewpoints demand the attention of researchers from multiple disciplines, including human-computer interaction, machine learning, and visual analytics. We extend the call to future researchers to investigate these questions and devise solutions to how the dichotomy between user agency and trust in mixed-initiative systems can be balanced.

## 5.4 Approaches such as Serendyze can Facilitate Deeper Fine-Grained Knowledge Acquisition

Participants were keen to use text-level analysis features available to them to read and analyze reviews in detail. Apart from the exploration metrics and suggestions, three features that were prominently mentioned by participants across all conditions are: (1) the ability to search for any keyword they wanted and to see them highlighted in the filtered reviews, (2) the ability to filter reviews by different sentiments (positive, neutral, and negative), and (3) the provenance tracking where they could mark the reviews they read. The free-form keyword search and highlighting provided users with the freedom to explore the reviews by focusing on what is important for them. The sentiments gave readers a nuanced sense of reviewers' disposition towards a product, which is different from star ratings, as star ratings are not an alternative to affinity represented on reviews [97]. Finally, the provenance tracking enabled them to track their review exploration without the need for mental notes, reducing cognitive effort for decision-making. Participants across different conditions expressed their desire to see functionalities provided in Serendyze on *"Amazon"* (P12, P49) or similar *"online sites"* (P18, P97).

Two overarching insights can be extracted from this observation. First, there is a lack of available functionalities and options to analyze reviews. Major online commerce websites (Amazon, Etsy, eBay, etc.) host numerous products with thousands of reviews per product, but do not provide powerful features to analyze the free-form text reviews. While there are filters such as price range, warranty, color, etc., these are product-level filters enabling analysis among products based on metadata attributes. They are often not connected with reviews for the product, and the readers seeking to purchase a product based on others' reviews have to painstakingly read through the reviews, often make mental notes, and make decisions based on incomplete knowledge [87].

While Amazon and eBay provide keywords extracted from reviews, the exploration capability they provide is limited. This leads us to the second insight: the necessity of fine-grained analysis at the text level. Participants' appreciation towards these seemingly rudimentary features suggests the utility of text-level analytics where the analysis can be performed on the review contents and highlights the necessity of integrating such features on available platforms.

However, such necessities lead us to questions around identifying the appropriate granularity [102] of information to present to readers for exploring and analyzing text. Some of these questions involve how to enable readers to analyze text content more efficiently while negating redundancy and how to combine visualization and computational approaches to disseminate information at multiple levels of granularity. In future, researchers from HCI, visualization, natural language processing, and information retrieval could collectively explore paradigms of information seeking when text-level analytics is integrated with summary-level overviews.

## 5.5 Application of Serendyze in Other Domains beyond Product Reviews

Our study revealed an opportunity for text-level analysis of product reviews to help readers learn more from the data prior to making data-driven decisions. This necessity may manifest in domains where comprehensively exploring text data is critical for decision-making. One such domain is civics, where decision-makers depend on large-scale public input to gain an understanding of public perception before making critical policy decisions [52, 73]. They often use analytics tools that enable analysis of public-generated data — predominantly text data as comments, ideas, and opinions — to measure the temperature of public perception [73]. While tools designed for analyzing redundant and often ambiguous public input help decision-makers get high-level overviews of public opinions, marginalized and unpopular opinions are often neglected due to the scale of public input and lack of analysis tools to identify such opinions [72], especially at text-level [52]. Since these decisions directly impact peoples' lives, effective analysis to ensure the perspectives

of all citizens are addressed is critical in this domain [73]. Interventions such as the exploration metrics can help decision-makers to identify and extract insights from redundant information and track whether their public input exploration is skewed towards certain agendas, topics, or sentiments. Furthermore, the bias mitigation model can serendipitously suggest opinions and feedback that might have remained hidden under more popular opinions. As such, these interventions could enable decision-makers in the civic domain to not only gain a holistic understanding of public input but also enhance their accountability and transparency [53], when making policy decisions.

Another domain where text-level analysis tools for short free-form text analysis, such as Serendyze can be expanded to is social media analysis. While there exist a plethora of tools and techniques to analyze social media texts [45, 76], the issues regarding aggregation and summarization of opinions also manifest in this domain [113]. Such issues are especially pertinent due to the concerns around algorithmic filtration and curation of social media content based on users' digital footprints [9, 81]. These curating algorithms often decide what social media contents the readers should be exposed to [11, 66], which might result in inadvertently creating filter bubbles [82]. For many people who use social media as a source of news and current affairs, such curation and presentation of catered data might promote homophily [10] and render the readers oblivious to the bigger picture of current affairs in virtual social spaces [34, 77]. Text-level analytics tools such as Serendyze can combat the formation of echo chambers via serendipitous suggestions of social media content that are dissimilar from the posts that a reader is exploring in social media. For instance, if a reader is mostly exploring content that originated from sources aligned with liberal ideas, they could be suggested content from sources that are inclined towards conservative thoughts. While social media readers have the agency to decide which ideas they align with, such text-level intervention can enable them to be introduced to opposing ideas that might lead to a better understanding of arguments prior to establishing social alignments.

## 6  LIMITATIONS AND FUTURE WORK

**Limitations.** One of the limitations of Serendyze is the latency associated with performing the pairwise comparison of visited and unvisited reviews to measure the similarity scores and generate suggestions. While the system performed well in a local system and during the pilot study, during the study with crowd workers, with up to 72 participants working simultaneously, the freely available server [1] used to perform the calculations was overwhelmed with traffic, and as a result, some of the participants (6/100) felt that the system worked slower than they expected. We emphasize that the latency is an outcome of logistical challenges and could have been mitigated with a more powerful back-end server or batch-wise distribution of tasks among crowd workers. In future, we will optimize Serendyze to perform more efficiently in low-resource environments. The other limitation involves the interface and the associated complexity. Some participants, especially the ones who used condition M&S (4/25), found some components of Serendyze to be confusing due to interface clutter. To mitigate this issue, the Serendyze interface could be improved by enabling participants to hide not just the suggestions but any component that they might not want to see. Although the inner workings of generating suggestions were not explained to participants due to study purposes, in future, the participants can be informed by adding an explanation to remove confusion and increase transparency.

Serendyze is designed as a customizable and modular web application. For this study, instead of probabilistic machine learning approaches, we used deterministic approaches to analyze reviews that included using keywords extractions based on co-occurrence and using star-rating as the foundation for sentiments. We emphasize that Serendyze can be outfitted with advanced computational methods to generalize it for tasks and domains where probabilistic classifications are acceptable and desired for scalability. However, in this study, we focused more on the interaction design and less on the computational approaches. As such we adopted deterministic approaches to identifying keywords, sentiments, and

similarities among reviews. In future, we plan to deploy Serendyze as a longitudinal study to track participants' purchase behavior across a month on multiple online products. Such experiments will enable us to further study the long-term impact of exploration metrics and bias mitigating suggestions on people's review exploration, holistic understanding, and data-driven decision-making.

**Future Work.** There are several avenues to explore in the future to improve Serendyze. We will study the utility of Serendyze in real-world scenarios by deploying it as a companion web application or browser extension that can enable readers to utilize Serendyze features to explore reviews on online commerce sites. In these real-world deployment, Serendyze will be outfitted with product reviews that mirror the real distribution of facets across product reviews. Before deploying Serendyze in a real-world setting, we will augment it with several functionalities based on this study and the knowledge we gained from the participant responses. For instance, we will add clarifying information to explain all components and optimize Serendyze to improve the scalability. We will also integrate and enable the readers to hot-switch between different suggestion-generating models to account for their exploration preferences during text exploration. Serenedyze's modular and customizable design (see Section 3) will allow us to experiment with various text analytics methods to enable exploration of various facets present in the data, including subjectivity [16], stance [63], and latent aspects [29, 88].

Some argue that product review distributions in online commerce websites are often inherently biased based on self-selection biases such as purchasing bias and under-reporting bias [49]. Such biases often result in the review distribution being bi-modal or non-normal, leaning more towards positive or negative reviews [24, 49]. One avenue to explore in future works is to study the effect of presenting suggestions that negate word of mouth on decision-making [48]. The modular and customizable design of Serendyze will enable us to replace the bias mitigation model with other statistical models appropriate for such studies. We also plan to study people's exploration patterns if they were limited to reading a fixed number of reviews, a fixed amount of time [110], or a fixed organization of suggestions.

In future, Serendyze could also be outfitted with features to disseminate and allow exploration and analysis of various product and review attributes, including product specifications, pictures, price, warranty information, peer rating, etc. Peer rating could also be used to weight the suggestions to provide recommendations based on how others valued a product. In addition, Serendyze could be improved by adding features to compare between two or more products in juxtaposition. Further improvements can be made by adding note-taking functionalities for the readers to further reduce mental load prior to decision-making. The Serendyze interface could also be updated with improved aesthetics and accessibility features to make it more presentable.

Another avenue to explore in future is to broaden the investigation and assess the applicability of text analytics tools like Serendyze in other domains. For instance, in the digital civics domain, exploration metrics and bias mitigating suggestions could be used to help decision-makers identify marginalized or unpopular perspectives among often redundant public-generated data. Furthermore, Serendyze could be used to analyze social media posts of contentious or divergent topics to help combat echo chambers [19]. In the future, we will collaborate with government and non-government organizations (NGO) — who collect, analyze, and make decisions based on public-generated data — to deploy and study how Serendyze can help them gain insights and a holistic understanding of public-generated texts.

The exploration metrics and bias mitigating suggestions in Serendyze could be also expanded beyond text to other media types including photos or videos. For instance, in applications such as Yelp, the bias exploration metrics and bias mitigating suggestions might help viewers identify distinct popular dishes, attractions, or places of interest among often redundant photos posted by people who have already experienced these commodities.

## 7  CONCLUSION

In this study, we investigated how supporting serendipitous discovery and analysis of short free-form text, such as product reviews, enables readers to explore product reviews more comprehensively to tease apart valuable insights from the mundane and the redundant. To that end, we proposed two interventions to enable readers to explore reviews more comprehensively prior to decision-making. These interventions include three exploration metrics — Visit, Coverage, and Distribution — that help readers to keep track of what reviews they have explicitly read, what redundant reviews they have implicit knowledge about via semantic and sentiment similarity, and how they have been exploring different facets of reviews such as sentiments compared to the true distributions of these facets present in the data. The second intervention is a bias mitigation model that generates suggestions based on what the readers have been exploring by focusing on identifying and suggesting reviews that are semantically and sentiment-wise dissimilar to the reviews the readers have read already. At any given point, the model generates suggestions to help readers mitigate biased exploration, guide readers to an understanding of the data, which is reflective of the true distributions of the semantic and sentiment diversity in the reviews, and enhance their knowledge discovery. We integrated these interventions with a text analytics system, Serendyze. Our evaluation with 100 crowd workers suggests that the exploration metrics enable readers to cover more reviews in a balanced way. We also found that the suggestions generated by the bias mitigation model were influential in enabling readers to make confident decisions. Furthermore, the study reveals a general tendency, necessity, and eagerness to utilize text-level exploration features to gather deeper knowledge about short free-form text. We discuss the impact of readers' preconceptions and perception on accepting suggestion from a system and how user agency in mixed-initiative systems play a significant role in how users trust interventions that generate guidance for them on what can or should do while using such a system. We also discuss how interventions to support text-level serendipitous discovery and analysis of short free-form text have utility and applicability in other domains beyond product reviews, where deeper exploration of text prior to data-driven decision-making is critical.

## REFERENCES

[1] 2021. Python Anywhere. Retrieved August 20, 2021 from https://www.pythonanywhere.com/

[2] Ahmed Abbas, Yilu Zhou, Shasha Deng, and Pengzhu Zhang. 2018. Text analytics to support sense-making in social media: A language-action perspective. *MIS Quarterly* 42, 2 (2018).

[3] Eric Alexander, Joe Kohlmann, Robin Valenza, Michael Witmore, and Michael Gleicher. 2014. Serendip: Topic model-driven visual exploration of text corpora. In *Proc. IEEE Conf. on Visual Analytics Science and Technology (VAST)*. IEEE, 173–182.

[4] Mohammad Alharbi and Robert S Laramee. 2019. Sos textvis: An extended survey of surveys on text visualization. *Computers* 8, 1 (2019), 17.

[5] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saied Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. 2017. Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268* (2017).

[6] Amazon. 2021. Community Guidelines. Retrieved September 1, 2021 from https://www.amazon.com/gp/help/customer/display.html?nodeId=GLHXEX85MENUE4XF

[7] Pek van Andel. 1994. Anatomy of the unsought finding. serendipity: Orgin, history, domains, traditions, appearances, patterns and programmability. *The British Journal for the Philosophy of Science* 45, 2 (1994), 631–648.

[8] Ricardo Baeza-Yates and Gonzalo Navarro. 1996. A faster algorithm for approximate string matching. In *Annual Symposium on Combinatorial Pattern Matching*. Springer, 1–23.

[9] Rahul Bhargava, Anna Chung, Neil S Gaikwad, Alexis Hope, Dennis Jen, Jasmin Rubinovitz, Belén Saldías-Fuentes, and Ethan Zuckerman. 2019. Gobo: A system for exploring user control of invisible algorithms in social media. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*. 151–155.

[10] Halil Bisgin, Nitin Agarwal, and Xiaowei Xu. 2012. A study of homophily on social media. *World Wide Web* 15, 2 (2012), 213–232.

[11] Engin Bozdag. 2013. Bias in algorithmic filtering and personalization. *Ethics and information technology* 15, 3 (2013), 209–227.

[12] Philip Burnard. 1991. A method of analysing interview transcripts in qualitative research. *Nurse Education Today* 11, 6 (1991), 461–466.

[13] Giuseppe Carenini and Lucas Rizoli. 2009. A multimedia interface for facilitating comparisons of opinions. In *Proceedings of the 14th international conference on Intelligent user interfaces*. 325–334.

[14] Davide Ceneda, Alessio Arleo, Theresia Gschwandtner, and Silvia Miksch. 2021. Show Me Your Face: Towards an Automated Method to Provide Timely Guidance in Visual Analytics. *IEEE Trans. on Visualization and Computer Graphics* (2021).

[15] Jonathan Chang, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems*, Vol. 22. Citeseer, 288–296.

[16] Iti Chaturvedi, Edoardo Ragusa, Paolo Gastaldo, Rodolfo Zunino, and Erik Cambria. 2018. Bayesian network based extreme learning machine for subjectivity detection. *Journal of The Franklin Institute* 355, 4 (2018), 1780–1797.

[17] Chaomei Chen, Fidelia Ibekwe-SanJuan, Eric SanJuan, and Chris Weaver. 2006. Visual analysis of conflicting opinions. In *2006 IEEE Symposium On Visual Analytics Science And Technology*. IEEE, 59–66.

[18] Jason Chuang, Daniel Ramage, Christopher Manning, and Jeffrey Heer. 2012. Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proc. SIGCHI Conf. on Human Factors in Computing Systems*. 443–452.

[19] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences* 118, 9 (2021).

[20] Christopher Collins, Natalia Andrienko, Tobias Schreck, Jing Yang, Jaegul Choo, Ulrich Engelke, Amit Jena, and Tim Dwyer. 2018. Guidance in the human–machine analytics process. *Visual Informatics* 2, 3 (2018), 166–180.

[21] Courtney D Corley and Rada Mihalcea. 2005. Measuring the semantic similarity of texts. In *Proc. ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*. 13–18.

[22] David Coyle, James Moore, Per Ola Kristensson, Paul Fletcher, and Alan Blackwell. 2012. I did that! Measuring users' experience of agency in their own actions. In *Proc. SIGCHI Conf. on Human Factors in Computing Systems*. 2025–2034.

[23] Matthew JC Crump, John V McDonnell, and Todd M Gureckis. 2013. Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS One* 8, 3 (2013), e57410.

[24] Kushal Dave, Steve Lawrence, and David M Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*. 519–528.

[25] Jenny L Davis. 2017. Curation: A theoretical treatment. *Information, Communication & Society* 20, 5 (2017), 770–783.

[26] Urška Demšar and Arzu Çöltekin. 2017. Quantifying gaze and mouse interactions on spatial visual interfaces with a new movement analytics methodology. *PloS One* 12, 8 (2017), e0181818.

[27] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[28] Nicholas Diakopoulos and Michael Koliska. 2017. Algorithmic transparency in the news media. *Digital Journalism* 5, 7 (2017), 809–828.

[29] Ying Ding, Changlong Yu, and Jing Jiang. 2017. A neural network model for semi-supervised review aspect identification. In *Pacific-Asia Conf. on Knowledge Discovery and Data Mining*. Springer, 668–680.

[30] Marian Dörk, Patrick Feng, Christopher Collins, and Sheelagh Carpendale. 2013. Critical InfoVis: exploring the politics of visualization. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*. 2189–2198.

[31] Marian Dörk, Nathalie Henry Riche, Gonzalo Ramos, and Susan Dumais. 2012. Pivotpaths: Strolling through faceted information spaces. *IEEE Trans. on Visualization and Computer Graphics* 18, 12 (2012), 2709–2718.

[32] Michael D Ekstrand, F Maxwell Harper, Martijn C Willemsen, and Joseph A Konstan. 2014. User perception of differences in recommender algorithms. In *Proc. ACM Conf. on Recommender Systems*. 161–168.

[33] Geoffrey Ellis. 2018. *Cognitive Biases in Visualizations*. Springer.

[34] Seth Flaxman, Sharad Goel, and Justin M Rao. 2016. Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly* 80, S1 (2016), 298–320.

[35] Kristopher Floyd, Ryan Freling, Saad Alhoqail, Hyun Young Cho, and Traci Freling. 2014. How online product reviews affect retail sales: A meta-analysis. *Journal of Retailing* 90, 2 (2014), 217–232.

[36] Wael H Gomaa, Aly A Fahmy, et al. 2013. A survey of text similarity approaches. *Int. J. Computer Applications* 68, 13 (2013), 13–18.

[37] Michelle Gregory, Nancy Chinchor, Paul Whitney, Richard Carter, Elizabeth Hetzler, and Alan Turner. 2006. User-directed sentiment analysis: Visualizing the affective content of documents. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*. 23–30.

[38] Sunwoo Ha, Adam Kern, Melanie Bancilhon, and Alvitta Ottley. 2020. Expectation Versus Reality: The Failed Evaluation of a Mixed-Initiative Visualization System. In *Proc. IEEE Workshop Celebrating the Scientific Value of Failure (FailFest)*. IEEE, 1–5.

[39] Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proc. Int. Conf. on World Wide Web*. 507–517.

[40] Natalie Henrich and Bev Holmes. 2013. Web news readers comments: Towards developing a methodology for using on-line comments in social inquiry. *J. Media and Communication Studies* 5, 1 (2013), 1–4.

[41] Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. 2004. Evaluating collaborative filtering recommender systems. *ACM Trans. on Information Systems (TOIS)* 22, 1 (2004), 5–53.

[42] Richards J Heuer. 1999. *Psychology of intelligence analysis*. Center for the Study of Intelligence.

[43] Enamul Hoque and Giuseppe Carenini. 2016. Multiconvis: A visual text analytics system for exploring a collection of online conversations. In *Proc. Int. Conf. on Intelligent User Interfaces*. 96–107.

[44] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proc. SIGCHI Conf. on Human Factors in Computing Systems*. 159–166.

[45]  Mengdie Hu, Krist Wongsuphasawat, and John Stasko. 2016. Visualizing social media content with sententree. *IEEE Trans. on Visualization and Computer Graphics* 23, 1 (2016), 621–630.

[46]  Mengdie Hu, Huahai Yang, Michelle X Zhou, Liang Gou, Yunyao Li, and Eben Haber. 2013. OpinionBlocks: a crowd-powered, self-improving interactive visual analytic system for understanding opinion text. In *IFIP Conf. on Human-Computer Interaction*. Springer, 116–134.

[47]  Nan Hu, Noi Sian Koh, and Srinivas K Reddy. 2014. Ratings lead you to the product, reviews help you clinch it? The mediating role of online review sentiments on product sales. *Decision Support Systems* 57 (2014), 42–53.

[48]  Nan Hu, Paul A Pavlou, and Jennifer Zhang. 2006. Can online reviews reveal a product's true quality? Empirical findings and analytical modeling of online word-of-mouth communication. In *Proceedings of the 7th ACM conference on Electronic commerce*. 324–330.

[49]  Nan Hu, Jie Zhang, and Paul A Pavlou. 2009. Overcoming the J-shaped distribution of product reviews. *Commun. ACM* 52, 10 (2009), 144–147.

[50]  Jessica Hullman and Nick Diakopoulos. 2011. Visualization rhetoric: Framing effects in narrative visualization. *IEEE Trans. on Visualization and Computer Graphics* 17, 12 (2011), 2231–2240.

[51]  Ellen Isaacs, Kelly Damico, Shane Ahern, Eugene Bart, and Mudita Singhal. 2014. Footprints: A visual search tool that supports discovery and coverage tracking. *IEEE Trans. on Visualization and Computer Graphics* 20, 12 (2014), 1793–1802.

[52]  Mahmood Jasim, Enamul Hoque, Ali Sarvghad, and Narges Mahyar. 2021. CommunityPulse: Facilitating Community Input Analysis by Surfacing Hidden Insights, Reflections, and Priorities. In *Proc. ACM Conf. on Designing Interactive Systems*. 846–863.

[53]  Mahmood Jasim, Pooya Khaloo, Somin Wadhwa, Amy X Zhang, Ali Sarvghad, and Narges Mahyar. 2021. CommunityClick: Capturing and reporting community feedback from town halls to improve inclusivity. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–32.

[54]  Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. 2019. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications* 78, 11 (2019), 15169–15211.

[55]  Quentin Jones, Gilad Ravid, and Sheizaf Rafaeli. 2004. Information overload and the message dynamics of online interaction spaces: A theoretical model and empirical exploration. *Information Systems Research* 15, 2 (2004), 194–210.

[56]  Iman Kamehkhosh and Dietmar Jannach. 2017. User perception of next-track music recommendations. In *Proc. Conf. on User Modeling, Adaptation and Personalization*. 113–121.

[57]  Marius Kaminskas and Derek Bridge. 2016. Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Trans. on Interactive Intelligent Systems (TiiS)* 7, 1 (2016), 1–42.

[58]  Gaurav Kapoor and Selwyn Piramuthu. 2009. Sequential bias in online product reviews. *Journal of Organizational Computing and Electronic Commerce* 19, 2 (2009), 85–95.

[59]  Khairullah Khan, Baharum Baharudin, Aurnagzeb Khan, and Ashraf Ullah. 2014. Mining opinion components from unstructured reviews: A review. *Journal of King Saud University-Computer and Information Sciences* 26, 3 (2014), 258–275.

[60]  Denis Kotkov, Shuaiqiang Wang, and Jari Veijalainen. 2016. A survey of serendipity in recommender systems. *Knowledge-Based Systems* 111 (2016), 180–192.

[61]  Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability. (2011).

[62]  Kostiantyn Kucher and Andreas Kerren. 2015. Text visualization techniques: Taxonomy, visual survey, and community insights. In *Proc. IEEE Pacific Visualization Symposium*. IEEE, 117–121.

[63]  Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)* 53, 1 (2020), 1–37.

[64]  Matevž Kunaver and Tomaž Požrl. 2017. Diversity in recommender systems–A survey. *Knowledge-based Systems* 123 (2017), 154–162.

[65]  Bum Chul Kwon, Sung-Hee Kim, Timothy Duket, Adrián Catalán, and Ji Soo Yi. 2015. Do people really experience information overload while reading online reviews? *Int. J. Human-Computer Interaction* 31, 12 (2015), 959–973.

[66]  Anja Lambrecht and Catherine Tucker. 2019. Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. *Management Science* 65, 7 (2019), 2966–2981.

[67]  Jey Han Lau and Timothy Baldwin. 2016. An empirical evaluation of doc2vec with practical insights into document embedding generation. *arXiv preprint arXiv:1607.05368* (2016).

[68]  Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Int. Conf. on Machine Learning*. PMLR, 1188–1196.

[69]  Greg Linden, Brent Smith, and Jeremy York. 2003. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing* 7, 1 (2003), 76–80.

[70]  Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*. 342–351.

[71]  Jie Lu, Dianshuang Wu, Mingsong Mao, Wei Wang, and Guangquan Zhang. 2015. Recommender system application developments: a survey. *Decision Support Systems* 74 (2015), 12–32.

[72]  Narges Mahyar, Mahmood Jasim, and Ali Sarvghad. 2020. Designing Technology for Sociotechnical Problems: Challenges and Considerations. *IEEE Computer Graphics and Applications* 40, 6 (2020), 76–87.

[73]  Narges Mahyar, Diana V Nguyen, Maggie Chan, Jiayi Zheng, and Steven P Dow. 2019. The Civic Data Deluge: Understanding the Challenges of Analyzing Large-Scale Community Input. In *Proc. ACM Conf. on Designing Interactive Systems*. 1171–1181.

[74]  Stephen Makonin, Daniel McVeigh, Wolfgang Stuerzlinger, Khoa Tran, and Fred Popowich. 2016. Mixed-initiative for big data: The intersection of human+ visual analytics+ prediction. In *Proc. Hawaii Int. Conf. on System Sciences (HICSS)*. IEEE, 1427–1436.

[75] Gary Marchionini. 2006. Exploratory search: from finding to understanding. *Commun. ACM* 49, 4 (2006), 41–46.

[76] Adam Marcus, Michael S Bernstein, Osama Badar, David R Karger, Samuel Madden, and Robert C Miller. 2011. Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proc. SIGCHI Conf. on Human Factors in Computing Systems*. 227–236.

[77] Solomon Messing and Sean J Westwood. 2014. Selective exposure in the age of social media: Endorsements trump partisan source affiliation when selecting news online. *Communication research* 41, 8 (2014), 1042–1063.

[78] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[79] Chris North, Remco Chang, Alex Endert, Wenwen Dou, Richard May, Bill Pike, and Glenn Fink. 2011. Analytic provenance: process+ interaction+ insight. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*. 33–36.

[80] Kensuke Onuma, Hanghang Tong, and Christos Faloutsos. 2009. Tangent: a novel,'surprise me', recommendation algorithm. In *Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*. 657–666.

[81] Orestis Papakyriakopoulos, Juan Carlos Medina Serrano, and Simon Hegelich. 2020. Political communication on social media: A tale of hyperactive users and bias in recommender systems. *Online Social Networks and Media* 15 (2020), 100058.

[82] Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you.* Penguin UK.

[83] Do-Hyung Park, Jumin Lee, and Ingoo Han. 2006. Information overload and its consequences in the context of online consumer reviews. *Proc. PACIS* (2006), 28.

[84] Do-Hyung Park, Jumin Lee, and Ingoo Han. 2007. The effect of on-line consumer reviews on consumer purchasing intention: The moderating role of involvement. *Int. J. Electronic Commerce* 11, 4 (2007), 125–148.

[85] Mateus M Pereira and Fernando V Paulovich. 2020. RankViz: A visualization framework to assist interpretation of Learning to Rank algorithms. *Computers & Graphics* 93 (2020), 25–38.

[86] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. arXiv:1802.05365 [cs.CL]

[87] Qualtrics. 2020. Online reviews statistics to know in 2021. Retrieved July 7, 2021 from https://www.qualtrics.com/blog/online-review-stats/

[88] Toqir A Rana, Yu-N Cheah, and Tauseef Rana. 2020. Multi-level knowledge-based approach for implicit aspect identification. *Applied Intelligence* 50, 12 (2020), 4616–4630.

[89] Keith Rayner, Timothy J Slattery, and Nathalie N Bélanger. 2010. Eye movements, the perceptual span, and reading speed. *Psychonomic Bulletin & Review* 17, 6 (2010), 834–839.

[90] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proc. LREC Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, 45–50.

[91] Pew Research. 2009. Project for Excellence in Journalism. Retrieved July 7, 2021 from http://worldcat.org/identities/lccn-no2004027960/

[92] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2011. Introduction to recommender systems handbook. In *Recommender Systems Handbook*. Springer, 1–35.

[93] Kerry Rodden and Xin Fu. 2007. Exploring how mouse movements relate to eye movements on web search results pages. (2007).

[94] Avi Rosenfeld and Ariella Richardson. 2019. Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems* 33, 6 (2019), 673–705.

[95] Ali Sarvghad, Melanie Tory, and Narges Mahyar. 2016. Visualizing dimension coverage to support exploratory analysis. *IEEE Trans. on Visualization and Computer Graphics* 23, 1 (2016), 21–30.

[96] Sheikh Muhammad Sarwar, Felipe Moraes, Jiepu Jiang, and James Allan. 2021. Utility of Missing Concepts in Query-biased Summarization. In *Proc. 44th International ACM SIGIR Conf. on Research and Development in Information Retrieval*. 2056–2060.

[97] Jacquelyn L Schreck and Matthew G Chin. 2019. Online Product Reviews: Effects of Star Ratings and Valence on Review Perception among Those High and Low in Need for Cognition. In *Proc. Human Factors and Ergonomics Society Annual Meeting*, Vol. 63. SAGE Publications Sage CA: Los Angeles, CA, 401–405.

[98] Urban Sedlar, Janez Bešter, and Andrej Kos. 2007. Tracking mouse movements for monitoring users' interaction with websites: Implementation and applications. *Elektrotehniški vestnik* 1, 74 (2007), 31–36.

[99] Bracha Shapira, Lior Rokach, and Shirley Freilikhman. 2013. Facebook single and cross domain data for recommendation systems. *User Modeling and User-Adapted Interaction* 23, 2-3 (2013), 211–247.

[100] Rui Shi, Paul Messaris, and Joseph N Cappella. 2014. Effects of online comments on smokers' perception of antismoking public service announcements. *J. Computer-Mediated Communication* 19, 4 (2014), 975–990.

[101] Yue Shi, Xiaoxue Zhao, Jun Wang, Martha Larson, and Alan Hanjalic. 2012. Adaptive diversification of recommendation results via latent factor portfolio. In *Proc. ACM SIGIR Conf. on Research and Development in Information Retrieval*. 175–184.

[102] Ben Shneiderman. 1996. The eyes have it: A task by data type taxonomy for information visualizations. In *Proc. IEEE Symp. on Visual Languages*. 336–343.

[103] Pedro Soto-Acosta, Francisco Jose Molina-Castillo, Carolina Lopez-Nicolas, and Ricardo Colomo-Palacios. 2014. The effect of information overload and disorganisation on intention to purchase online. *Online Information Review* (2014).

[104] Statista. 2021. Unit shipments of headphones worldwide from 2013 to 2020. Retrieved December, 2021 from https://www.statista.com/statistics/236075/revenue-of-headphone-shipments-in-the-united-states/

[105] Alice Thudt, Uta Hinrichs, and Sheelagh Carpendale. 2012. The bohemian bookshelf: supporting serendipitous book discoveries through information visualization. In *Proc. SIGCHI Conf. on Human Factors in Computing Systems*. 1461–1470.

[106] Bettina von Helversen, Katarzyna Abramczuk, Wiesław Kopeć, and Radoslaw Nielek. 2018. Influence of consumer reviews on online purchasing decisions in older and younger adults. *Decision Support Systems* 113 (2018), 1–10.

[107] Emily Wall, Arup Arcalgud, Kuhu Gupta, and Andrew Jo. 2019. A markov model of users' interactive behavior in scatterplots. In *Proc. IEEE Visualization Conference (Short Papers)*. IEEE, 81–85.

[108] Emily Wall, Leslie M Blaha, Lyndsey Franklin, and Alex Endert. 2017. Warning, bias may occur: A proposed approach to detecting cognitive bias in interactive visual analytics. In *Proc. IEEE Conf. on Visual Analytics Science and Technology (VAST)*. IEEE, 104–115.

[109] Emily Wall, Leslie M Blaha, Celeste Lyn Paul, Kristin Cook, and Alex Endert. 2018. Four perspectives on human bias in visual analytics. In *Cognitive Biases in Visualizations*. Springer, 29–42.

[110] Ruolin Wang, Zixuan Chen, Mingrui Ray Zhang, Zhaoheng Li, Zhixiu Liu, Zihan Dang, Chun Yu, and Xiang'Anthony' Chen. 2021. Revamp: Enhancing Accessible Information Seeking Experience of Online Shopping for Blind or Low Vision Users. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[111] Wesley Willett, Jeffrey Heer, and Maneesh Agrawala. 2007. Scented widgets: Improving navigation cues with embedded visualizations. *IEEE Trans. on Visualization and Computer Graphics* 13, 6 (2007), 1129–1136.

[112] Max Woolf. 2014. A statistical analysis of 1.2 million amazon reviews. *Max Woolf's Blog* (2014).

[113] Yingcai Wu, Nan Cao, David Gotz, Yap-Peng Tan, and Daniel A Keim. 2016. A survey on visual analytics of social media data. *IEEE Transactions on Multimedia* 18, 11 (2016), 2135–2148.

[114] Yingcai Wu, Furu Wei, Shixia Liu, Norman Au, Weiwei Cui, Hong Zhou, and Huamin Qu. 2010. OpinionSeer: interactive visualization of hotel customer feedback. *IEEE Trans. on visualization and computer graphics* 16, 6 (2010), 1109–1118.

[115] Ashima Yadav and Dinesh Kumar Vishwakarma. 2020. Sentiment analysis using deep learning architectures: a review. *Artificial Intelligence Review* 53, 6 (2020), 4335–4385.

[116] Koji Yatani, Michael Novati, Andrew Trusty, and Khai N Truong. 2011. Review spotlight: a user interface for summarizing user-generated reviews using adjective-noun word pairs. In *Proc. SIGCHI Conf. on Human Factors in Computing Systems*. 1541–1550.

[117] Guo Yu and Alan Blackwell. 2017. Effects of timing on users' agency during mixed-initiative interaction. BCS.

[118] Kem ZK Zhang, Sesia J Zhao, Christy MK Cheung, and Matthew KO Lee. 2014. Examining the influence of online reviews on consumers' decision-making: A heuristic–systematic model. *Decision Support Systems* 67 (2014), 78–89.

[119] Yi Zhang, Jamie Callan, and Thomas Minka. 2002. Novelty and redundancy detection in adaptive filtering. In *Proc. ACM SIGIR Conf. on Research and Development in Information Retrieval*. 81–88.