Teachable Facets: A Framework of Interactive Machine Teaching for Information Filtering

Swati Mishra[*]
Cornell University
Ithaca, USA
swati@infosci.cornell.edu

Matt Ryerkerk
Bloomberg
New York, USA
mryerkerk@bloomberg.net

Yitzchak D. Lockerman
Bloomberg
New York, USA
ylockerman@bloomberg.net

David Eis
Bloomberg
New York, USA
deis@bloomberg.net

Jeffrey M. Rzeszotarski
Cornell University
Ithaca, USA
jeffrz@cornell.edu

ABSTRACT

Interactive tools help users filter relevant information from massive online sources, like news feeds and online discussion forums, by enabling them to externalize their preferences. However, users' information goals and preferences are often complex and are comprised of data attributes and a user's subjective judgements over these attributes. For instance, when filtering news articles based on their newsworthiness, the system must capture both data attributes like recency and shareability of the article, along with the user's personal and flexible assessment of news sentiment. While most interactive tools enable users to externalize goals that are expressible as true/false statements, they do not support incorporating subjective, loosely structured judgements of data attributes which fulfill complex goals. In this paper, we introduce Teachable Facets (TF), widgets that users can create on the fly to filter relevant information to improve the sense-making of analysts. These teachable widgets employ a Machine Teaching (MT) framework to enable users to formulate personalized filtering criteria for complex, multi-dimensional, loosely indexed, and unstructured data; teach a filtering criterion using representative samples; apply these filters to new data streams; and assess the relevance of outcomes. Through a user study, we evaluate the performance of these filters based on their ability to discover relevant items and the expressibility they offer to the users in teaching criteria. In our discussion, we identify ways this approach might improve future systems and delineate implications should such systems be deployed broadly.

CCS CONCEPTS

• **Information systems** → **Personalization**; • **Human-centered computing** → *User interface design.*

KEYWORDS

Interactive Machine Learning, Faceted Search, User Interface Design, Machine Teaching, Information Filtering

ACM Reference Format:
Swati Mishra, Matt Ryerkerk, Yitzchak D. Lockerman, David Eis, and Jeffrey M. Rzeszotarski. 2024. Teachable Facets: A Framework of Interactive Machine Teaching for Information Filtering. In *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '24), March 10–14, 2024, Sheffield, United Kingdom.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3627508.3638289

1 INTRODUCTION

Filtering relevant information from large data streams can be challenging due to the limitations of "vocabulary" offered by search interfaces and the inherent complexity of user preferences [1]. For instance, consider news feed applications that are commonly available as part of search engine platforms. These employ filters to extract newsworthy articles from large news feeds based on criteria like popular topics, exclusivity, or most read articles. Using these pre-defined faceted search interfaces [51], users can create dynamic queries like "show the most recent news on topic X", or "show most frequently read articles". However, these pre-defined filters often limit users' vocabulary. Complex and nuanced information needs, such as the newsworthiness of a story or its potential economic impact, cannot be easily operationalized as toggleable filters. This is because determining elements of complex filtering goals, such as the newsworthiness of an article, involves a variety of sub-criteria like the *exclusivity* of the event it reports, *proximity* of the event to the reader, *eliteness* of the people involved, its *shareability*, *impact*, or use of linguistic features like *sentiment* [5, 19]. Some of these factors are easy to quantify and encode as a filter – for instance, *proximity* may be represented using the geographic location tag of the reader's device, and *shareability* may be represented by the number of readers likely to share a given news topic on social media platforms. On the other hand, factors such as news *sentiment* and its impact involve qualitative judgements, are highly personalized, and evolve as a user's preferences change. This makes concepts such as newsworthiness difficult to model and incorporate as part of a traditional faceted search interface (see Figure 1).

Failure to capture these complex user preferences can result in unintended consequences in information filtering tasks. If a system solely relies on quantifiable sub-components such as its shareability

[*] The author is currently affiliated with McMaster University, Canada and can be contacted at mishrs23@mcmaster.ca

CHIIR '24, March 10–14, 2024, Sheffield, United Kingdom

ACM ISBN 979-8-4007-0434-5/24/03
https://doi.org/10.1145/3627508.3638289

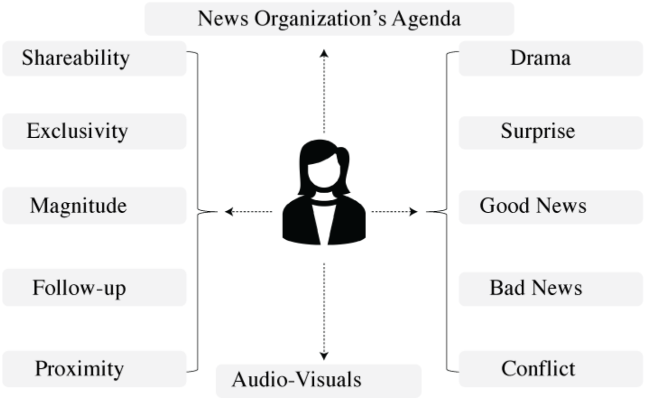


Figure 1: The complex definition of newsworthy articles as described in [19]. The left side shows the quantifiable news factors, while the right side shows qualitative and highly personal news factors.

(i.e., filtering based on "most shared articles"), it risks trapping readers in homogeneous viewpoints [31] or skewing their impression of the information space. Furthermore, designing interfaces that promote usage of few specific filtering criteria may also be risky for journalistic organizations. If news interfaces always promote "most read" headlines, organizations have perverse incentives to write on topics that are most shared as opposed to most informative. Prior research has addressed these challenges by employing Machine Learning (ML) to model complex user behavior and preferences [47, 61]. These models work behind the scenes, learning from implicit user feedback including items clicked and read, topics selected and liked, location, and user demographics. Using these representations of user behavior, models automatically discover relevant items from large news feeds and make them available as a recommended list. However, recent studies have demonstrated that these implicitly derived user interest models often introduce a homogenization effect among users with diverse opinions, and users do not approve of these agents for "hard" news topics due to their opacity [32, 34, 62]. For critical decision-making tasks like identifying stocks to trade or news topics of journalistic value, users prefer to externalize their complex information needs over automated techniques even if they are more arduous or less effective [52].

Though faceted search with dynamic query filter widgets has specific limitations with regards to complex and nuanced criteria such as newsworthiness, it provides a higher degree of transparency and explicit feedback, thereby addressing some of the concerns that users have about complex models. These benefits come from the way faceted search allows users to navigate through complex information hierarchies [20], make reversible changes, see adjustments reflected immediately [1], and refine queries based on an iterative sensemaking loop [28]. The mechanism that currently falls short in cases such as newsworthiness is the Boolean nature of most dynamic query filters (i.e., turned on or off) and their lack of agility in the face of changing criteria. Likewise, the lack of transparency, conceptual complexity, and homogenization of unitary machine learning recommendation systems falls short for these complex criteria. In this paper, we propose a human-in-the-loop ML approach which blends these opposing strategies in the form of *Teachable Facets*.

Teachable Facets (TF) are ML-based filtering widgets that users can create on the fly to filter data based on personalized and complex criteria. These filters incorporate a learning algorithm that can be *taught* complex user preferences through qualitative user judgments. The teaching process helps users understand their criteria more deeply and build rapport with the system. TFs can be used as stand-alone lenses through which to filter large, constantly updating information streams or can be combined with pre-defined filtering widgets to express preferences. We employ a Machine Teaching approach [33, 49, 54], to enable users to create these filters, apply them in news filtering tasks, and revisit and modify them when their preferences evolve.

In this paper, we describe the design of Teachable Facets and implement it in the context of a news filtering task. Personalizing news feeds based on subjective criteria (e.g., filtering "good news about topic X") can be especially hard for users of traditional browsing interfaces [22]. First, we decompose these subjective criteria based on qualitative features that may define them (e.g., good news may be defined by the presence of positive sentiments like joy and surprise). We then provide mechanisms to manipulate these features in order to create custom definitions of filtering criteria. Using the TF framework, users can *formulate* personalized filtering criteria for multi-dimensional, loosely indexed, and unstructured data like news headlines; *teach* a filtering criterion using representative samples; *apply* these filters to incoming data streams; and *assess* the relevance of the outcomes in terms of visible items in the news feed. We deploy our interactive tool in a crowdsourcing environment, where users are given the task of creating news filters that represent subjective criteria shown on the right side of Figure 1. For all participants, we gather system usage data via an interaction logger and analyze it to identify interactive information filtering behavior. Over the course of a session, users develop a number of switchable filters that capture different aspects of their criteria, enabling them to quickly customize and refine their feed.

Our work generalizes to broader information filtering tasks that feature complex and nuanced criteria that are hard for users to externalize. This includes applications such as product search (matching ill-defined user preferences), literature surveys (where keyword search may be insufficient), anomaly detection in event streams (anomalies may be hard to characterize by heuristics), and collaborative sensemaking (managing competing group criteria). We contribute a novel technical approach that leverages anomaly detection and machine teaching frameworks to quickly and effectively construct complex content classification models. We evaluate the success of our system based on the expressibility that Teachable Facets offer to the users to externalize complex preferences and their ability to discover relevant items as they learn user criteria. Finally, we discuss limitations and extensions of our approach, especially as it relates to broader use cases in ML model-mediated content curation and management.

2 RELATED WORK

Capturing complex filtering goals is a widely researched topic in areas like recommendation systems [63], user modeling [59], and search [36]. A majority of this work focuses on employing Machine Learning (ML) techniques that automatically learn user preferences from their interaction history and multi-faceted user profiles incorporating demographics, location, activity on other platforms, etc. [11, 47, 61]. More related to our work is literature on interactive interfaces that provide control and transparency to users and help them externalizing their information goals and interests [35]. One popular approach to filtering relevant information is by employing hierarchical facets [20], which provide comprehensive categories to represent information. For instance, Cat-a-cone [21] employs a faceted search paradigm that makes it easier to find categories in a medical datasets, like searching for documents on Radiation Therapy based on keywords and labels. Many recent web-based tools for applications like e-commerce, music, travel, etc., employ a similar paradigm. They enable users to filter relevant items using data properties represented through widgets. However, these traditional tools fail in scenarios where data properties change rapidly based on context, timeliness, and relevance, for instance, in news filtering tasks [8, 29].

Finding newsworthy topics is a constant challenge in journalistic inquiry [31]. With thousands of news articles produced every day, it has become increasingly difficult for individuals to identify news items that are worthy of their attention [62]. Both information consumers and distributors struggle to identify newsworthy articles; while the consumers need to prioritize which headlines to click and read, the distributors (like online news aggregation platforms) need to decide which headlines should be pushed to the top of the page [4]. One popular approach to filtering relevant news is to use collaborative filtering [50] in which items are filtered similarly for groups of users with similar interests. Prior systems have combined these techniques with faceted search to help users stay abreast of the most popular viewpoints. For instance, a news recommendation platform might organize news headlines based on content and their popularity among users who follow similar topics [43]. However, these systems suffer from issues like lack of transparency and explicit user feedback in news filtering, "cold-start" and "grey sheep" problems, and continuous changes in user interests due to topic divergence, evolution of preferences, and sharing of devices [15, 41, 56]. In this research, we address these challenges by designing mechanisms to augment faceted search. Through interactive real-time feedback, users can teach their complex preferences and create facets in real-time that represent their filtering criteria.

Prior research has proposed various mechanisms to augment faceted search by providing users with agency to specify custom criteria during a search session, thus improving their overall experience [6, 40, 58]. For instance, researchers observed that introducing interactive faceted query suggestions into filtering tasks helps to improve directed situated navigation [51]. Another system, FacetBrowser [58], enables users to build complex "stories" with sequences of related searches using facets and organize their search tasks effectively. Interactive systems like SearchLens [6] build upon this technique and employ a keyword-based mechanism that enables users to create a collection of "Lenses" that reflect their different latent interests across different contexts. Our research builds upon this vast body of work by engaging the user in creating their own preference profiles and externalizing their intent through a machine teaching paradigm. This approach offers the advantage of learning more flexible, loosely-defined preference criteria and provides the ability to save these filters as "checkpoints" that can be further refined when user preferences evolve.

Research shows that incorporating real-time user feedback in ML systems can lead to better system performance on challenging tasks [2, 13, 23]. For instance, interactive data clustering tools like AppGrouper [7] which enable users to edit and modify the size of data clusters produced by the algorithm, yield more coherent and useful data clusters compared to algorithms alone. Prior research has focused extensively on designing interfaces that help users provide feedback on data input and model outcome [16, 24, 38]. One popular approach to leveraging human feedback is to provide representative samples to help model learn the user's decision boundary [54, 69]. By providing feedback on instances selected by a learning model, the user can guide the ML model towards a desired behavior. For instance, in order to improve the relevance of top-k documents ranked by an algorithm, Intent Radar [44] incorporates both negative and positive relevance feedback to improve the specificity of model outcome. With these interactive systems, users primarily engage in improving a single ML model outcome for a given task, and focus their effort on providing relevant feedback while assessing the outcome. However, it is difficult to capture complex and evolving user preferences within the scope of a single model for filtering incoming news streams. In this research, we provide an opportunity for end users to select important features, provide feedback on representative samples for training the model, and build multiple models that capture different decision criteria for different tasks over the same dataset. Our system supports users' holistic engagement with the entire ML process using an interactive teaching paradigm, and incorporate it within their complex information workflows.

We deploy our interactive tool in a crowdsourcing environment to gather data on how people use Teachable Facets for a news filtering task. Crowdsourcing platforms have been adopted widely for ML tasks like data generation, annotation, and model training [57]. When making ML models accessible to users with limited expertise, it is important to explain model results in a way they can understand. Research in explainable AI provides substantial evidence that lack of interpretability of model outcome and its decision-making process adversely impacts the system's usability and reliability [37, 65]. Within ML systems, researchers have discussed the impact of model prediction and explanation complexity on the user's mental models [27, 39]. In this research, we incorporate simple explanations to improve transparency of the filtering algorithm.

3 SYSTEM DESIGN

Faceted filters are comprised of labels that represent categories to which an entity might belong. These labels are characterized by data dimensions (or features); for instance, news headline containing features like players, game, score, and events are likely to be categorized as sports news. In our approach, we preserve this spirit of faceted navigation and decompose loosely-defined labels into users' own *conceptual* features. Therefore, *good news* may incorporate words that communicate sentiments like happy and surprise, *bad news* may incorporate fear, anger and sadness, and *drama* may incorporate words that have mixed sentiments like both happy and angry [46]. Our Teachable Facets (TF) system seeks to help users create information filtering criteria that capture their news preferences, modify them as their preferences evolve, and apply them to filter relevant data items from constantly updating information streams. In TF, we employ a simple, yet powerful, machine teaching paradigm, which employs a *select-teach-review* iterative loop in lieu of forming a training and test set of ground truth prior to training (as in traditional ML models). Users *select* the news features they consider important, *teach* by explicitly associating their preference to read and share a news headline with the features they selected, and *review* the recommended items in the news feed based on their feedback. Over multiple TF iterations, users teach a specific faceted view one aspect of their preferences. Across a session, users create new facets or modify existing facets in order to personalize and diversify their news feed, eventually forming a stable of customized facets which can be deployed like any other traditional faceted browsing widget.

3.1 Use Case

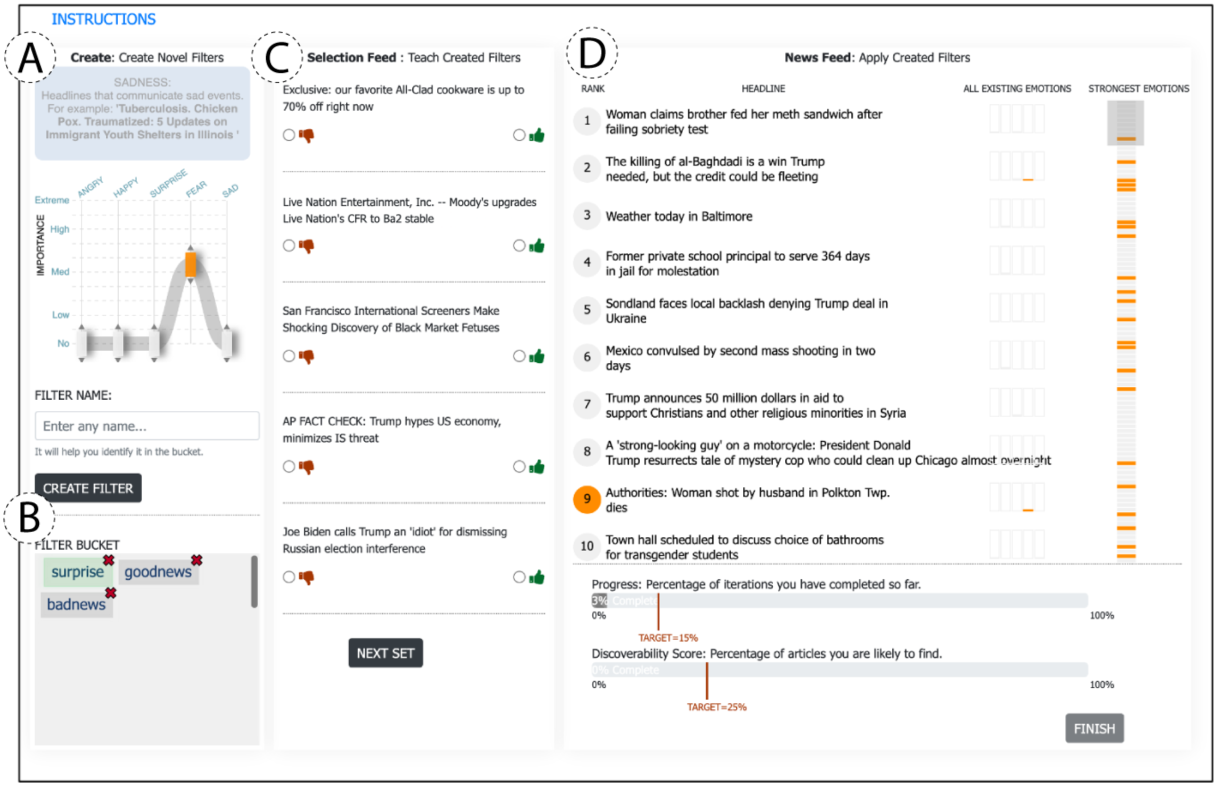
To demonstrate how this system could be used, we present a hypothetical user: Isolde. Isolde is an independent trader who regularly invests in stocks based on the sentiments of news coverage for a company. Related research has noted that professionals now increasingly make financial decisions like which stock to trade based on qualitative news factors like sentiment [53, 67]. To this end, Isolde keeps track of specific news topics and wants to organize her news feed based on target sentiment. She is frustrated with her news aggregation website which does not allow her to create qualitative filters. In the past, Isolde has also observed that the kinds of news filters she uses shape her future news feed without any explicit input from her. While facets help users navigate large databases using pre-defined categories that are coherent and complete [60, 68], these interactions often predispose users to translate their complex preferences using limited vocabulary. In Isolde's cases, this neglected her subjective judgment of news headlines (like sentimens and impact), which was equally important.

Isolde is particularly interested in news that communicate sentiments like good (G), conflict (C) and surprise (S). A part of her trading strategy is to invest in high volume stocks when news coverage has positive and conflicting news sentiment [53]. While she can search for keywords like "happy," or "shocking news" and toggle on different regional filters, it doesn't really allow her to find articles which are positive *in tone*, and match her specific idea of what is surprising. Isolde moves to the Teachable Facets interface (Figure 2). She is greeted by a welcome page explaining how she can create a filter by teaching specific definitions to the system. While this bootstrapping problem is a general downside of machine teaching interfaces, it overcomes "cold-start" and "grey sheep" problems associated with automated systems [17]. The TF interface shows her some initial news to give a sense that it is starting to curate. Isolde creates her first filter S, thinking that it would be an effective way to judge system performance.

To kickstart the teaching process, she first adjusts sliders on the left panel of the interface (Figure 2 A) that help to specify the features that define the filter S. She then names this filter, which saves this configuration as a facet in the view. The sliders serve two specific purposes in our design. First, it helps the users prioritize those features that are more or less important for a criteria, and thus distinguish between labels that share overlapping features. Several features are expected to be present in a given text, and the slider provides an opportunity to externalize how the user defines a given filtering criteria. For instance, a *surprise news* filter S, may prioritize news that communicates either happy or sad sentiments (like occurrence of a sudden natural disaster), but *good news* filter G must prioritize news that communicates high happy sentiments (like neighboring communities helping in rescue efforts after a sudden natural disaster). This is aided by the inclusion of color-coding throughout the interfaces. In Figure 2 B, Isolde has created filter S using features like surprise and happy; the highlighted colors indicate where these features have been found by the model. Once configured, these features function as *reference points* for both the user and the algorithm; the user can review whether the instances shown in the news feed match their specified preference criteria, while the algorithm narrows down the feature space of the data instances for learning the filtering criteria.

After specifying filtering criteria using preferred data features, Isolde begins teaching the model about her specific definition of *surprisingness* using the middle panel of the interface (Figure 2 C). TF presents a limited number of news headlines from which the users can select the *best matching* and *worst matching* headlines corresponding to the filtering criteria they specified. Using both positive and negative instances for feedback is a useful technique for learning accurate criteria [44]. In Isolde's case, she identifies some headlines that are genuinely surprising to her, and others which may be surprising to her friend Marke, but are not at all surprising based on her knowledge of the world. As she teaches the filtering criteria, it is automatically applied to the incoming news feed (Figure 2 D) to demonstrate how her preferences have been reflected. The progress bar at the bottom also helps her judge her progress through the teaching process, and how well the model is learning her preferences. Once she has hit a point where the model seems to understand what she wants, Isolde pauses and saves it as a new facet in her view.

Isolde then follows a similar process to find stories that have conflict; having sentiments that are both happy and sad. She notices that the kinds of headlines she receives as she teaches are different, and that this iterative teaching loop is also helping her get a sense of her own definition of what a "happy news story" may be. Once she has created several facets in the view, she can turn them on or off and adjust the view to create a news feed that fits her specific interests. She can also repeat the teaching process on her filter to tune it based on her new interests or patterns in how journalists compose headlines. Isolde might then check out how her filters have fared with some recent world changes, and may find that her own specific interests and definition of surprisingness have since changed in the process. She can then repeat the teaching process on her filter to tune it to her new interests or new patterns in how



SADNESS:
Headlines that communicate sad events.
For example: 'Tuberculosis. Chicken Pox. Traumatized: 5 Updates on Immigrant Youth Shelters in Illinois '

Enter any name...

It will help you identify it in the bucket.

Exclusive: our favorite All-Clad cookware is up to 70% off right now

Live Nation Entertainment, Inc. -- Moody's upgrades Live Nation's CFR to Ba2 stable

San Francisco International Screeners Make Shocking Discovery of Black Market Fetuses

AP FACT CHECK: Trump hypes US economy, minimizes IS threat

Joe Biden calls Trump an 'idiot' for dismissing Russian election interference

1 Woman claims brother fed her meth sandwich after failing sobriety test

2 The killing of al-Baghdadi is a win Trump needed, but the credit could be fleeting

3 Weather today in Baltimore

4 Former private school principal to serve 364 days in jail for molestation

5 Sondland faces local backlash denying Trump deal in Ukraine

6 Mexico convulsed by second mass shooting in two days

7 Trump announces 50 million dollars in aid to support Christians and other religious minorities in Syria

8 A 'strong-looking guy' on a motorcycle: President Donald Trump resurrects tale of mystery cop who could clean up Chicago almost overnight

9 Authorities: Woman shot by husband in Polkton Twp. dies

10 Town hall scheduled to discuss choice of bathrooms for transgender students

Progress: Percentage of iterations you have completed so far.

3% Complete

Discoverability Score: Percentage of articles you are likely to find.

0% Complete

Figure 2: Our interactive news filtering interface designed to incorporate the Teachable Facets framework. Users can A) specify their complex criteria using feature importance, B) create multiple filters, C) the interface presents qualified news headlines that might be relevant for teaching by selecting one newsworthy and one not newsworthy headline, D) apply the created filters on an incoming news feed by clicking on the filter. This view also explains the various headline features prioritized by the filter.

journalists compose headlines. Isolde observes that, by applying these facets, she is able to find news articles that help make her financial decisions based on media coverage a company is receiving. While the TF system implemented for this paper functions on a static dataset [64] (so as to control exposure when conducting evaluations), future iterations one might also easily import new data streams and include elements of traditional faceted filters, such as keyword search and topics.

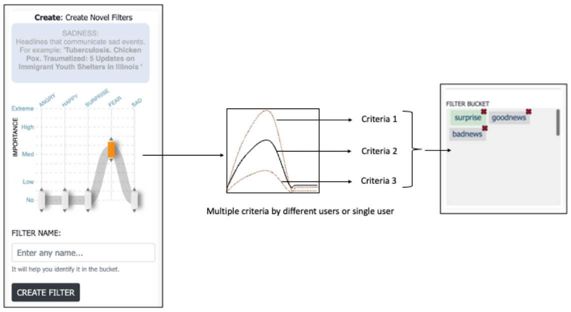
3.2 Implementation

There are three major components of our Teachable Facets system. The first element focuses on assisting users in teaching through examples. To support this, we designed a *Detector* unit comprising of an active sampling technique that selects headlines useful for the algorithm and which match a user's criteria. The second element, called the *Learning to Rank* unit, focuses on providing ranking based on trained filters, and ensuring that users can flexibly alternate between different filters. Our final element, referred to as the *Performance* unit, considers how to provide quick summaries of ranking feed performance so users can analyze the efficacy of filters they created.

3.2.1 *The Detector.*

The first challenge faced by the system is in selecting examples to present to the user for teaching (Figure 2 C). Headlines are selected for presentation to users by a Detector designed to perform two tasks to address this challenge. It first decomposes each incoming headline into a feature vector N=[e_1,e_2,.., e_n] where n= number of sentiments being considered. Here, we set n=5, corresponding to 5 prominent types of sentiment used for filtering news [53, 67] and employ a simple text-2-emotion converter that uses word synthesis to extract sentiment [9]. Using the set of news headline sentiment vectors, it assigns each one a score between 0 and 1 based on their relative variance across sentiment dimensions. This first pass finds headlines that have significant identifiable sentiment, as there maybe headlines that are either too noisy (e.g., parsing errors) or too neutral to provoke sentiments for teaching (e.g., "Samsung won't support Linux on DeX once Android 10 arrives") and are therefore ill-suited for TF's feed.

To operationalize this, the Detector employs Isolation Forest [30], a technique that uses feature trees to isolate data points that have significantly different distribution of features. We train our Isolation Forest to consider headlines with low sentiment scores (e_n < 0.3) to be part of a normal distribution, and to flag any headlines that fall outside this distribution. This results in all incoming headlines being assigned a score S between 0 and 1, depending upon the intensity of sentiments they communicate (highest score

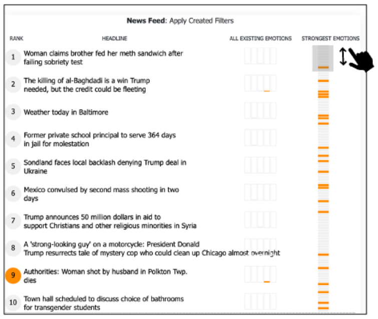


Create: Create Novel Filters

SADNESS:
Headlines that communicate sad events.
For example: 'Tuberculosis. Chicken Pox. Traumatized: 5 Updates on Immigrant Youth Shelters in Illinois '

FILTER NAME:
Enter any name...
It will help you identify it in the bucket.

Figure 3: Creating new filters based on unique preference criteria. Users may specify multiple filtering preferences.



News Feed: Apply Created Filters

RANK	HEADLINE	ALL EXISTING EMOTIONS	STRONGEST EMOTIONS
1	Woman claims brother fed her meth sandwich after failing sobriety test		
2	The killing of al-Baghdadi is a win Trump needed, but the credit could be fleeting		
3	Weather today in Baltimore		
4	Former private school principal to serve 364 days in jail for molestation		
5	Sondland faces local backlash denying Trump deal in Ukraine		
6	Mexico convulsed by second mass shooting in two days		
7	Trump announces 50 million dollars in aid to support Christians and other religious minorities in Syria		
8	A 'strong-looking guy' on a motorcycle: President Donald Trump resurrects tale of mystery cop who could clean up Chicago almost overnight		
9	Authorities: Woman shot by husband in Polkton Twp. dies		
10	Town hall scheduled to discuss choice of bathrooms for transgender students		

Figure 4: Scrolling through headlines produced by applying a given filter. Showing rank, headline, features, and overview.

for headlines communicating highest intensity sentiments). The Detector then selects 5 headlines from this pool, the highest ranked (S), the lowest ranked, and 3 headlines that are close to the mean score and correspond to the user's criteria. The intuition is that we want the user to see both extreme and typical cases during teaching.

It is important to note that the problem of identifying data points corresponding to a sentiment may also be formulated as a classification task that can be achieved using a variety of (ML) classification algorithms. However, framing this problem as an anomaly detection (AD) task (which we accomplish using Isolation Forest) offers two major advantages over traditional classification tasks. First, AD algorithms are equipped to handle unbalanced data better than classification algorithms [26]. This allows them to perform reliably with highly skewed distributions that occur frequently in real-world data streams (for instance, news headlines from a particular source might be dominated primarily by only select sentiments like anger, anticipation or surprise, with only a few positive headlines occurring occasionally in the data stream). Second, it helps handle settings such as ours where no convenient ground truth data is available to train or test the algorithm in a supervised fashion.

3.2.2 *Learning to Rank.*

An important design goal of our system is to provide a sustainable way of creating and maintaining facets, support reusability, and, most importantly, refine them when a user's decision criteria changes or evolves. We operationalize this by employing a representative sample-based teaching paradigm. By leveraging both positive and negative feedback, our system tries to capture not just the user preference but also the relative difference between preferred and non-preferred items. With every iteration, the detector presents items that have vectors far apart in the feature space to maximize information gained in each training iteration. When a user continues to teach or re-visits a filter which they wish to refine, they can adjust the feature importance and select new representative samples. For instance, the user may first create a *conflict* filter in combination with a *geolocation* filter, to identify all news articles reporting challenges faced by a local community, and then refine the same *conflict* filter to identify challenges within a particular city by selecting additional representative data instances.

To support this evolution process, we introduce the functionality of *Filter Buckets*, which save each filter facet state as an intermediate "checkpoint." Each saved state comprises of four parts; 1) the initial state of the model, 2) the data instances selected to teach it, 3) the intermediate/final state of the model, and 4) its predicted ranked scores on the incoming news feed. These saved checkpoints are activated when the user clicks on a given filter, and are applied to the most current data stream. In this setting, the model begins to train from the last checkpoint, thus learning more nuances from additional data points. Our system also allows users to delete a filter altogether if it no longer is useful and considerable concept drift has occurred between the filter's learned boundary and a user's decision criteria. In the future, our state storage system might also allow an 'undo' capability or additional visualizations for explainability.

For our learning to rank algorithm, we implemented a pointwise learning approach [25] using a logistic regression estimator that can learn from mini batches using a sequential optimizer in non-stationary classification setting [45]. The algorithm is simple, yet powerful, and provides transparency in terms of feature weights and importance. Being an online learner, the model can operate successfully in a machine teaching setting where all the data necessary to learn a decision boundary are not available, but are gradually accumulated instead as users encounter new instances.

3.2.3 *Ranking and Explaining Feed Performance.*

As users are teaching the model, it is critical to give them a sense of how well the model is learning. One method is to simply show the headlines to the user and let them be the adjudicator (Figure 3). Another method we pursue is to provide metrics which help users judge the efficacy of their new filter (Figure 2 D). Besides visualizing the raw number of teaching iterations, we also estimate the *discoverability factor* of the current filter. The filters with high discoverability are better at reaching desired articles, giving a general sense of how well the filter might perform in displaying relevant headlines at the top of the news feed.

To compute this metric, our scoring function predicts the ranking scores on a test set of 1,000 news headlines, out of which 100 news headlines are available in the user's news feed. For each teaching iteration, we calculate the rank gain of a headline using its rank ρ_i after current iteration and rank ρ_{i-1} immediately before the current iteration using this formula: $\lambda_i = log(\rho_i/\rho_{i-1})$. As the pool of headlines is N=1000, the maximum possible rank gain is $\lambda_{i,max} = 3$ and minimum possible rank gain is $\lambda_{i,min} = -3$. We define an item to be discoverable when $\lambda_i \neq 0$ for a given iteration, and its rank changes across each iteration. When rank gain $\lambda_i > 0$, the item becomes more discoverable as it moves up in ranking while if $\lambda_i < 0$, it becomes less discoverable across each iteration as result of user action. A successful filter would rank all items that correspond to a user-defined criteria higher than the others. Therefore, the cumulative rank gain for relevant items should be greater than the cumulative rank gain for irrelevant items. With this definition, if a user prefers to see news filtered based on criteria G on topic X in their news feed, then all news headlines matching the feature dimensions with $\lambda_i > 0$ will make the news feed more relevant.

Another key goal of our system is to provide transparency at every stage of the filter creation process. When engaging end users with ML systems, it is important to provide support for interpreting model results, visualizing the outcome and helping them estimate how their decisions are impacting the model's outcome. To achieve this goal, we incorporated both visualizations of the outcome of each filter and the feature distributions of each headline it was prioritizing. The news feed incorporates a ranking summary visualization that provides an overview of how their filter is ranking headlines with different sentiments (Figure 4), which is updated after every iteration. We incorporate color channels to counter change blindness that might result in oversights [55]. New, more relevant headlines are pushed to the top of the ranked list after new filters are applied, in addition to refining previously applied facets.

4 EVALUATION

Teachable Facets offer flexibility to users in determining what items should be prioritized over others. The framework enables users to establish a filtering criteria and provide examples of the best and worst matches pertaining to this criteria. The goal of our system is to augment the implicit feedback about user preferences (e.g., clicks) with explicitly articulated preference feedback (e.g., selected features and positive and negative examples corresponding to these features). The outcome of our system is a ranked list of news headlines that are filtered from a large pool based on the criteria specified by the user.

Evaluating ranking algorithms is a topic of ongoing research, comprising of techniques that assess ranking systems based on metrics like accuracy, relevance, agency and fairness [12, 14, 18, 66]. In our work, we employed a pointwise learning-to-rank algorithm that has been well evaluated for its performance in various contexts [25]. Therefore, our goal in this work was not to re-evaluate our algorithm on benchmark criteria. Additionally, evaluating most ranking systems requires ground truth, which is expensive to annotate for constantly updating streams.

Instead, our goal was to understand how well our system responds to individual preferences and produces relevant results. This motivated our analysis towards the broader question of *do teachable facets personalize well to individual preferences*? One way to evaluate whether the system outcome is likely to be personalized to an individual's preferences is to investigate if for a given filtering criteria (e.g., good news) and personalized user defined configurations of feature dimensions (e.g., sentiments like happy and surprise) produce different, yet relevant, ranked outcomes. We conduct a user study to investigate the above question, which we describe in detail in the next section.

4.1 User Study Design

We recruited a group of people (N=72) to participate in the pilot studies to test the usability of our system, clarity of instructions, performance of the underlying model, and stability of the system when multiple people use it simultaneously. For the main study, we recruited another group of participants (N=32), the results of which are presented below. All users were recruited via a popular crowd sourcing platform. Before beginning the task, each participant first took a pre-task survey in which they provided background about their news reading habits and demographic information. Participants then took a screening survey, where they were presented with 2 questions corresponding to each of the 5 sentiment dimension (a total of 10 questions). In every question, they were presented with a sentiment followed by a group of 5 news headlines. Their task was to identify 1 headline that represented the sentiment in question (e.g., *identify which of the following headline communicates "happy news" using positive words*). The goal of this survey was to capture a snapshot of how they assess the sentiments using the linguistic features of a headline, and to help them prepare for the task ahead. These questions were drawn from a corpus of news headlines curated by expert annotators [3].

Following this survey, users were presented with detailed instructions on how to use the interactive tool to create novel filters. They were allowed to freely explore the tool and create exploratory filters to get familiar with the system. They were each assigned a task to create and teach two (2) filters randomly sampled from a group of three filtering criteria *good news (G)*, *surprise news (S)* and *news with drama (D)*. They were provided with detailed definitions of the news factors from a journalistic standpoint, along with examples. Using these reference points, they first selected the feature dimensions and defined their importance for each of these filters, assigned a name to their filter, and taught them iteratively by choosing representative samples that the system selected for them. They were each allotted minutes, although they could leave anytime during the session, not exceeding 1 hour. The participants were assigned a modest target of 15 iterations and a discoverability score of 0.3 (presented as a percentage) to be achieved during the session. They were compensated with a reward of $12 per hour and a bonus for achieving discoverability score > 0.75. In the below section, we share the results of the study.

4.2 Results

Of the 32 participants in the study, 15 were male, 16 were female, and 1 did not disclose their gender identity. 18 participants were below the age of 30, 10 participants were between the ages of 30 to 50, and 4 participants were over 50. Participants had varied educational backgrounds, with a majority of them having a college degree (N=20), some had a Master's degree (N=8), 1 had a Ph.D., and 1 had a high school education. All participants were frequent consumers of local news from a variety of online and offline sources. The 32 participants created 84 unique filters with varied feature dimensions and importance. Of these 84 filters, 49 were designed with the specific task instructions (G=25 D=13 S=11), while the remaining filters were built for exploration purposes. These had various names including factors like *fun*, *positivity*, *unhappy*, and *sad news*, *fake news* and *neutral*.

To analyze the difference between the various feature dimensions selected, we consider a baseline filter B that has all 5 feature dimensions selected and set to medium importance, i.e., B=[0.5, 0.5, 0.5, 0.5, 0.5]. This baseline corresponds to creating a filter that does not have any user preferences and uses all features with equal (medium) importance to rank incoming news feeds. We computed the Euclidean distance E [10] of the feature dimensions defined for each filter (G,S,D) with the baseline filter B, where $E_{min} = 0$ and $E_{max} = 1.12$ (owing to the 5 features we used). We observed that the 49 filters corresponding to the task were different from the baseline in terms of features and their importance scores selected; filter G (min=0.48, max=1.12, std=0.17), filter D (min=0.48, max=1.02, std=0.20), and filter S (min=0.55, max=1.11, std=0.15). We further observed that participants also engaged in refining their filtering criteria throughout the course of the session. 27 out of 32 participants engaged in modifying their definitions (or filtering criteria) they selected more than once during the session. They often switched between different filters and applied varied criteria to explore how they affected their news feeds.

We considered the first ranked list presented to the participants at the beginning of the task as baseline. The participant's goal was to personalize this news feed by selecting appropriate headlines to read. Using a Kendall Tau (KT) ranked correlation measure, we computed the correlations between the baseline news feed and final news feed created by each of the 49 filters. A KT score of 1.0 would indicate the feeds were exactly the same as the baseline for each given filter, while a value of -1.0 indicates highly dissimilar feeds. We computed KT (τ) measures using two criteria; the first was comparing the ranks of actual headline. This enabled us to measure whether the outcome of each filter applied to the news feed resulted in the exact same headlines. We observed that τ values for filter G (Mean $\tau = 0.11$, $std = 0.42$), filter S (Mean $\tau = -0.2$, $std = 0.33$) and filter D (Mean $\tau = 0.03$, $std = 0.07$) indicated weak correlations with the baseline, suggesting that they were different from the starting news feed. The second correlation test measured ranked correlations using the sentiments communicated by each ranked headline. This allowed us to measure, for a group of filters, the similarity between the sentiments of the ranked headlines. We observed that τ values for filter G (Mean $\tau = 0.15$, $std = 0.36$), filter S (Mean $\tau = -0.009$, $std = 0.12$) and filter D (Mean $\tau = 0.03$, $std = 0.092$) suggested slightly stronger correlations with the baseline. These observations suggests that users who created filters with similar goals, but different definitions in terms of feature selection, were able to see news headlines that are not exactly the same and yet communicate relevant and similar sentiments.

5 DISCUSSION

In the previous sections, we described our interactive interface, user study design, and experimental results. In the following section, we discuss several important implications related to the design of interactive Machine Teaching systems.

Teachable Facets provides a flexible framework for filtering information based on complex user preferences. We observed that users engaged in creating diverse filtering criteria and refining them to extract the most relevant news headlines. Formulating queries based on data features and representative samples enabled them to explicitly impact the news ranking algorithm. For instance, when filtering news headlines with surprising news, they manipulated the feature importance when the algorithm was unable to prioritize relevant headlines with the initially specified features. This interaction provided context switching between formulating queries and exploring query results, and encouraged users to create news filters that weren't specified in the task. For instance, they used the slider features to create news filters like fake news, sad news, entertainment, etc.

Prior research demonstrates that users prefer faceted search interfaces for exploratory tasks [42]. Our research provides key evidence that faceted search can also encourage data exploration. This is particularly useful for information access tasks like news reading, where it is important for users to be familiar with diverse viewpoints. It is important to note that the features used to decompose the target criteria play important roles in this exploratory search. They function as "anchor points" through which users specify exploratory filtering criteria, find navigation leads, and investigate their goals. These features are used to also gauge the relevance of the retrieved information and evaluate the usability of the facets. Therefore, it is important that the choice of these features be grounded carefully in the user's goals. In future works, it will be important to study how complexity of the feature set may further impact teaching and creating unique facets.

Teachable Facets enable users to engage in machine teaching by specifying information filtering goals using representative samples and feature manipulation. Teaching involves adjusting the input to the learning agent based on the understanding of the outcome, and is a cognitively demanding process. Therefore, it was critical to incorporate algorithmic transparency for teacher success. During *pilot studies*, participants provided evidence that they were aware of the need for ML explanations to measure their teaching success; P23 mentioned *"... thought I would see signs that AI was learning from me, but nothing to me indicated that was the case..."*

In our final design, we introduced feature-based explanations that faithfully represented model performance over every iteration. We observed that this significantly improved success rates and encouraged users to engage in exploration beyond the task specifications. However, teaching is an iterative process that requires users to iteratively access machine explanations. We observed that a majority of participants were able to complete the task effectively (P16 mentioned *"Everything was easy to understand and complete,"* while P8 said *"Very nice filter mechanism with the story tone [news sentiments] table."*).

To engage users in teaching, it is not only important to design explanations that are interpretable and actionable, but also reduce the time and cognitive load. Machine Teaching interfaces that provide a one-to-one mapping between input and output data features can help to simplify teaching tasks that are cognitively demanding. While significant research has been focused on how to design reliable and interpretable ML explanations in various application settings [65], there is little research on the cognitive load associated with explanations. Additionally, our work also poses interesting research questions on uncovering the temporal effects of the explanation language, and what explanation design patterns present lesser cognitive load to their users.

We incorporated an instance-based teaching mechanism that enabled users to select news headlines that represented their specified criteria. This mechanism was adopted successfully in prior research in machine teaching [39, 48] and is particularly advantageous for filtering applications. We introduced a segregation between the news items participants used to teach and those they saw in the feed. This design choice was intended to provide an intuitive segregation between training and the test set, and was critical for users to *edit* the facets they created. However, since all participants were familiar with traditional news recommendation interfaces which do not provide a clear and transparent distinction between the instances used to learn user preferences, it caused a bit of learning curve with our system. P9 mentioned *"...I thought they should be the most newsworthy and most interesting to me as first, then second, third, fourth and fifth least. Instead, the third option was one I didn't want to read but would share and the fourth was one I did want to read but not share..."* However, the nature of our interaction design led users to eventually break from prior beliefs and mental models, and successfully create news filters using their custom definitions. P4 mentioned *"I think overtime I could make this work as I got used to the system and could train it better."*

These observations suggest one of the major challenges MT systems might face in the wild; *enabling users to move from the role of passive data providers to active teachers of their own systems.* Modern algorithmic systems increasingly deploy automated systems, thus, limiting end user agency. While they provide short term convenience in terms of reduced effort, they increase user reliance on algorithmic decisions. Therefore, it is important for system designers to consider mechanisms that will reduce cognitive load, but offer high agency for Machine Teaching systems in high stakes settings.

6 LIMITATIONS AND FUTURE WORK

Our research has several limitations which open pathways to future research in this area. First, while we demonstrate how Teachable Facets could be useful to support personalized information filtering and capturing complex user preferences, we did not explore how this framework could be used in combination with static (or traditional) facets. For instance, a good news filter may be created for all news headlines corresponding to topic X. In future work, we intend to extend this framework to faceted search interfaces that include traditional facets. Second, we analyzed system performance in terms of the relevance of headlines generated, and the user experience of teaching complex preferences. However, we did not perform a detailed investigation of cognitive load on users and the impact of prolonged use of these tools on news reading habits. In future work, we intend to focus on a much detailed evaluation of the tool that includes investigating whether teachable facets introduce news readers with diverse opinions, and help them break filter bubbles [32]. Third, we studied the design of TF interfaces only in context to identifying newsworthy headlines from constantly updating streams on news aggregation websites, without considering factors like source, volume, distribution, etc. Not all sources produce news articles with the same volume. For instance, larger news organizations can produce significantly larger number of articles everyday than smaller news organizations. It will be interesting, and eventually critical, to study how teachable facets might impact reader bias towards certain sources. Finally, the tests we conducted involved approximately 110 participants recruited through crowdsourcing. While this demographic represents frequent users of news aggregation websites, it will be useful to design user studies geared toward professionals like traders and journalists.

7 CONCLUSION

In this research, we first described the challenges associated with capturing complex user preferences. We then propose the design of Teachable Facets, an interactive machine teaching framework for information filtering tasks. We implement TF in context to an online news aggregation platform and deploy them in a crowd sourced environment. We systematically evaluate the design and outline implications that can highlight the benefits associated with TF framework.

ACKNOWLEDGMENTS

This research is supported by a grant from the Bloomberg Data Science Ph.D. Fellowship Program. It was conducted while the first author was a Doctoral Candidate at Cornell University, USA. She is now affiliated with McMaster University, Canada, as an Assistant Professor. We thank the members of the Bloomberg AI Engineering Group for their valuable feedback and continued support in improving this work. We thank the conference reviewing committee for their insightful comments and detailed suggestions.

REFERENCES

[1] Christopher Ahlberg and Ben Shneiderman. 1994. Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, Massachusetts, USA) *(CHI '94)*. Association for Computing Machinery, New York, NY, USA, 313–317. https://doi.org/10.1145/191666.191775

[2] Stuart Berg, Dominik Kutra, Thorben Kroeger, Christoph N Straehle, Bernhard X Kausler, Carsten Haubold, Martin Schiegg, Janez Ales, Thorsten Beier, Markus Rudy, et al. 2019. Ilastik: interactive machine learning for (bio) image analysis. *Nature methods* 16, 12 (2019), 1226–1232.

[3] Laura Ana Maria Bostan, Evgeny Kim, and Roman Klinger. 2020. GoodNewsEveryone: A Corpus of News Headlines Annotated with Emotions, Semantic Roles, and Reader Perception. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 1554–1566. https://aclanthology.org/2020.lrec-1.194

[4] Mark Boukes, Natalie P Jones, and Rens Vliegenthart. 2022. Newsworthiness and story prominence: How the presence of news factors relates to upfront position and length of news stories. *Journalism* 23, 1 (2022), 98–116. https://doi.org/10.1177/1464884919899313 arXiv:https://doi.org/10.1177/1464884919899313

[5] Helen Caple and Monika Bednarek. 2016. Rethinking news values: What a discursive approach can tell us about the construction of news discourse and news photography. *Journalism* 17, 4 (2016), 435–455.

[6] Joseph Chee Chang, Nathan Hahn, Adam Perer, and Aniket Kittur. 2019. SearchLens: Composing and Capturing Complex User Interests for Exploratory Search. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 498–509. https://doi.org/10.1145/3301275.3302321

[7] Shuo Chang, Peng Dai, Lichan Hong, Cheng Sheng, Tianjiao Zhang, and Ed H. Chi. 2016. AppGrouper: Knowledge-Based Interactive Clustering Tool for App Search Results. In *Proceedings of the 21st International Conference on Intelligent User Interfaces* (Sonoma, California, USA) *(IUI '16)*. Association for Computing Machinery, New York, NY, USA, 348–358. https://doi.org/10.1145/2856767.2856783

[8] Po-Huan Chiu, Gloria Yi-Ming Kao, and Chi-Chun Lo. 2010. Personalized blog content recommender system for mobile phone users. *International Journal of*

Human-Computer Studies 68, 8 (2010), 496–507. https://doi.org/10.1016/j.ijhcs.2010.03.005 Measuring the Impact of Personalization and Recommendation on User Behaviour.

[9] Nicholas Cummins, Shahin Amiriparian, Sandra Ottl, Maurice Gerczuk, Maximilian Schmitt, and Björn Schuller. 2018. Multimodal Bag-of-Words for Cross Domains Sentiment Analysis. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Calgary, AB, Canada, 4954–4958. https://doi.org/10.1109/ICASSP.2018.8462660

[10] Per-Erik Danielsson. 1980. Euclidean distance mapping. *Computer Graphics and Image Processing* 14, 3 (1980), 227–248. https://doi.org/10.1016/0146-664X(80)90054-4

[11] Abhinandan S. Das, Mayur Datar, Ashutosh Garg, and Shyam Rajaram. 2007. Google News Personalization: Scalable Online Collaborative Filtering. In *Proceedings of the 16th International Conference on World Wide Web* (Banff, Alberta, Canada) *(WWW '07)*. Association for Computing Machinery, New York, NY, USA, 271–280. https://doi.org/10.1145/1242572.1242610

[12] Sarah Dean, Sarah Rich, and Benjamin Recht. 2020. Recommendations and User Agency: The Reachability of Collaboratively-Filtered Information. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 436–445. https://doi.org/10.1145/3351095.3372866

[13] John J. Dudley and Per Ola Kristensson. 2018. A Review of User Interface Design for Interactive Machine Learning. *ACM Trans. Interact. Intell. Syst.* 8, 2, Article 8 (jun 2018), 37 pages. https://doi.org/10.1145/3185517

[14] Efthimis N. Efthimiadis. 1993. A User-Centred Evaluation of Ranking Algorithms for Interactive Query Expansion. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pittsburgh, Pennsylvania, USA) *(SIGIR '93)*. Association for Computing Machinery, New York, NY, USA, 146–159. https://doi.org/10.1145/160688.160710

[15] Blaž Fortuna, Carolina Fortuna, and Dunja Mladenić. 2010. Real-Time News Recommender System. In *Machine Learning and Knowledge Discovery in Databases*, José Luis Balcázar, Francesco Bonchi, Aristides Gionis, and Michèle Sebag (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 583–586.

[16] Dominic Girardi, Josef Kueng, and Andreas Holzinger. 2015. A Domain-Expert Centered Process Model for Knowledge Discovery in Medical Research: Putting the Expert-in-the-Loop. In *Brain Informatics and Health*, Yike Guo, Karl Friston, Faisal Aldo, Sean Hill, and Hanchuan Peng (Eds.). Springer International Publishing, Cham, 389–398.

[17] Benjamin Gras, Armelle Brun, and Anne Boyer. 2016. Identifying Grey Sheep Users in Collaborative Filtering: A Distribution-Based Technique. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization* (Halifax, Nova Scotia, Canada) *(UMAP '16)*. Association for Computing Machinery, New York, NY, USA, 17–26. https://doi.org/10.1145/2930238.2930242

[18] Asela Gunawardana and Guy Shani. 2009. A Survey of Accuracy Evaluation Metrics of Recommendation Tasks. *J. Mach. Learn. Res.* 10 (dec 2009), 2935–2962.

[19] Tony Harcup and Deirdre O'Neill. 2017. What is News? *Journalism Studies* 18, 12 (2017), 1470–1488. https://doi.org/10.1080/1461670X.2016.1150193 arXiv:https://doi.org/10.1080/1461670X.2016.1150193

[20] M. A. Hearst. 2006. Design recommendations for hierarchical faceted search interfaces. In *Proc. SIGIR 2006, Workshop on Faceted Search*. Association for Computing Machinery, Seattle, WA, 26–30.

[21] Marti A. Hearst and Chandu Karadi. 1997. Cat-a-Cone: An Interactive Interface for Specifying Searches and Viewing Retrieval Results Using a Large Category Hierarchy. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Philadelphia, Pennsylvania, USA) *(SIGIR '97)*. Association for Computing Machinery, New York, NY, USA, 246–255. https://doi.org/10.1145/258525.258582

[22] Bernie Hogan. 2015. *From Invisible Algorithms to Interactive Affordances: Data After the Ideology of Machine Learning*. Springer International Publishing, Cham, 103–117. https://doi.org/10.1007/978-3-319-05467-4_7

[23] Andreas Holzinger, Markus Plass, Michael Kickmeier-Rust, Katharina Holzinger, Gloria Cerasela Crişan, Camelia-M. Pintea, and Vasile Palade. 2019. Interactive Machine Learning: Experimental Evidence for the Human in the Algorithmic Loop. *Applied Intelligence* 49, 7 (jul 2019), 2401–2414. https://doi.org/10.1007/s10489-018-1361-5

[24] Shih-Wen Huang, Pei-Fen Tu, Wai-Tat Fu, and Mohammad Amanzadeh. 2013. Leveraging the Crowd to Improve Feature-Sentiment Analysis of User Reviews. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces* (Santa Monica, California, USA) *(IUI '13)*. Association for Computing Machinery, New York, NY, USA, 3–14. https://doi.org/10.1145/2449396.2449400

[25] Muhammad Ibrahim and Mark Carman. 2016. Comparing Pointwise and Listwise Objective Functions for Random-Forest-Based Learning-to-Rank. *ACM Trans. Inf. Syst.* 34, 4, Article 20 (aug 2016), 38 pages. https://doi.org/10.1145/2866571

[26] Suraya Nurain Kalid, Keng-Hoong Ng, Gee-Kok Tong, and Kok-Chin Khor. 2020. A Multiple Classifiers System for Anomaly Detection in Credit Card Data With Unbalanced and Overlapped Classes. *IEEE Access* 8 (2020), 28210–28221. https://doi.org/10.1109/ACCESS.2020.2972009

[27] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. IEEE Institute of Electrical and Electronics Engineers, San Jose, CA, USA, 3–10. https://doi.org/10.1109/VLHCC.2013.6645235

[28] Andrew Kuznetsov, Joseph Chee Chang, Nathan Hahn, Napol Rachatasumrit, Bradley Breneisen, Julina Coupland, and Aniket Kittur. 2022. Fuse: In-Situ Sensemaking Support in the Browser. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 34, 15 pages. https://doi.org/10.1145/3526113.3545693

[29] Lei Li, Li Zheng, Fan Yang, and Tao Li. 2014. Modeling and broadening temporal user interest in personalized news recommendation. *Expert Systems with Applications* 41, 7 (2014), 3168–3177. https://doi.org/10.1016/j.eswa.2013.11.020

[30] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. Isolation Forest. In *2008 Eighth IEEE International Conference on Data Mining*. IEEE, Pisa, Italy, 413–422. https://doi.org/10.1109/ICDM.2008.17

[31] Jiahui Liu, Peter Dolan, and Elin Rønby Pedersen. 2010. Personalized News Recommendation Based on Click Behavior. In *Proceedings of the 15th International Conference on Intelligent User Interfaces* (Hong Kong, China) *(IUI '10)*. Association for Computing Machinery, New York, NY, USA, 31–40. https://doi.org/10.1145/1719970.1719976

[32] Ping Liu, Karthik Shivaram, Aron Culotta, Matthew A. Shapiro, and Mustafa Bilgic. 2021. The Interaction between Political Typology and Filter Bubbles in News Recommendation Algorithms. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 3791–3801.

[33] Weiyang Liu, Bo Dai, Ahmad Humayun, Charlene Tay, Chen Yu, Linda B. Smith, James M. Rehg, and Le Song. 2017. Iterative Machine Teaching. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 2149–2158. https://proceedings.mlr.press/v70/liu17b.html

[34] Mykola Makhortykh and Mariëlle Wijermars. 2021. Can Filter Bubbles Protect Information Freedom? Discussions of Algorithmic News Recommenders in Eastern Europe. *Digital Journalism* 0, 0 (2021), 1–25. https://doi.org/10.1080/21670811.2021.1970601 arXiv:https://doi.org/10.1080/21670811.2021.1970601

[35] Gary Marchionini. 2006. Exploratory Search: From Finding to Understanding. *Commun. ACM* 49, 4 (apr 2006), 41–46. https://doi.org/10.1145/1121949.1121979

[36] Andrew McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. 1999. A Machine Learning Approach to Building Domain-Specific Search Engines. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2* (Stockholm, Sweden) *(IJCAI'99)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 662–667.

[37] Swati Mishra and Jeffrey M. Rzeszotarski. 2021. Crowdsourcing and Evaluating Concept-Driven Explanations of Machine Learning Models. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 139 (apr 2021), 26 pages. https://doi.org/10.1145/3449213

[38] Swati Mishra and Jeffrey M Rzeszotarski. 2021. Designing Interactive Transfer Learning Tools for ML Non-Experts. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 364, 15 pages. https://doi.org/10.1145/3411764.3445096

[39] Swati Mishra and Jeffrey M Rzeszotarski. 2023. Human Expectations and Perceptions of Learning in Machine Teaching. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization* (Limassol, Cyprus) *(UMAP '23)*. Association for Computing Machinery, New York, NY, USA, 13–24. https://doi.org/10.1145/3565472.3595612

[40] Toni-Jan Keith Palma Monserrat, Shengdong Zhao, Kevin McGee, and Anshul Vikram Pandey. 2013. NoteVideo: Facilitating Navigation of Blackboard-Style Lecture Videos. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 1139–1148. https://doi.org/10.1145/2470654.2466147

[41] Raymond J. Mooney and Loriene Roy. 2000. Content-Based Book Recommending Using Learning for Text Categorization. In *Proceedings of the Fifth ACM Conference on Digital Libraries* (San Antonio, Texas, USA) *(DL '00)*. Association for Computing Machinery, New York, NY, USA, 195–204. https://doi.org/10.1145/336597.336662

[42] Xi Niu, Xiangyu Fan, and Tao Zhang. 2019. Understanding Faceted Search from Data Science and Human Factor Perspectives. *ACM Trans. Inf. Syst.* 37, 2, Article 14 (jan 2019), 27 pages. https://doi.org/10.1145/3284101

[43] Maria Panteli. 2019. Recommendation Systems Compliant with Legal and Editorial Policies: The BBC+ App Journey. In *Proceedings of the 13th ACM Conference on Recommender Systems* (Copenhagen, Denmark) *(RecSys '19)*. Association for Computing Machinery, New York, NY, USA, 529. https://doi.org/10.1145/3298689.3346961

[44] Jaakko Peltonen, Jonathan Strahl, and Patrik Floréen. 2017. Negative Relevance Feedback for Exploratory Search with Visual Interactive Intent Modeling. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces* (Limassol, Cyprus) *(IUI '17)*. Association for Computing Machinery, New York,

NY, USA, 149–159. https://doi.org/10.1145/3025171.3025222

[45] W.D. Penny and S.J. Roberts. 1999. Dynamic logistic regression. In *IJCNN'99. International Joint Conference on Neural Networks. Proceedings (Cat. No.99CH36339)*, Vol. 3. IEEE, Washington, DC, USA, 1562–1567 vol.3. https://doi.org/10.1109/IJCNN.1999.832603

[46] Robert Plutchik. 1980. Chapter 1 - A GENERAL PSYCHOEVOLUTIONARY THEORY OF EMOTION. In *Theories of Emotion*, Robert Plutchik and Henry Kellerman (Eds.). Academic Press, 3–33. https://doi.org/10.1016/B978-0-12-558701-3.50007-7

[47] Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. 2022. News Recommendation with Candidate-Aware User Modeling. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 1917–1921. https://doi.org/10.1145/3477495.3531778

[48] Hema Raghavan, Omid Madani, and Rosie Jones. 2006. Active Learning with Feedback on Features and Instances. *J. Mach. Learn. Res.* 7 (dec 2006), 1655–1686.

[49] Gonzalo Ramos, Christopher Meek, Patrice Simard, Jina Suh, and Soroush Ghorashi. 2020. Interactive machine teaching: a human-centered approach to building machine-learned models. *Human–Computer Interaction* 35, 5-6 (2020), 413–451.

[50] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. 1994. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work* (Chapel Hill, North Carolina, USA) *(CSCW '94)*. Association for Computing Machinery, New York, NY, USA, 175–186. https://doi.org/10.1145/192844.192905

[51] Tuukka Ruotsalo, Giulio Jacucci, and Samuel Kaski. 2020. Interactive faceted query suggestion for exploratory search: Whole-session effectiveness and interaction engagement. *Journal of the Association for Information Science and Technology* 71, 7 (2020), 742–756. https://doi.org/10.1002/asi.24304 arXiv:https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.24304

[52] Chirag Shah and Emily M. Bender. 2022. Situating Search. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval* (Regensburg, Germany) *(CHIIR '22)*. Association for Computing Machinery, New York, NY, USA, 221–232. https://doi.org/10.1145/3498366.3505816

[53] Dev Shah, Haruna Isah, and Farhana Zulkernine. 2018. Predicting the Effects of News Sentiments on the Stock Market. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 4705–4708. https://doi.org/10.1109/BigData.2018.8621884

[54] Patrice Y Simard, Saleema Amershi, David M Chickering, Alicia Edelman Pelton, Soroush Ghorashi, Christopher Meek, Gonzalo Ramos, Jina Suh, Johan Verwey, Mo Wang, et al. 2017. Machine teaching: A new paradigm for building machine learning systems. *arXiv preprint arXiv:1707.06742* (2017).

[55] Daniel J. Simons and Daniel T. Levin. 1997. Change blindness. *Trends in Cognitive Sciences* 1, 7 (1997), 261–267. https://doi.org/10.1016/S1364-6613(97)01080-2

[56] Xiaoyuan Su and Taghi M. Khoshgoftaar. 2009. A Survey of Collaborative Filtering Techniques. *Adv. in Artif. Intell.* 2009, Article 4 (jan 2009), 1 pages. https://doi.org/10.1155/2009/421425

[57] Jennifer Wortman Vaughan. 2017. Making Better Use of the Crowd: How Crowdsourcing Can Advance Machine Learning Research. *J. Mach. Learn. Res.* 18, 1 (2017), 7026–7071.

[58] Robert Villa, Nicholas Gildea, and Joemon M. Jose. 2008. FacetBrowser: A User Interface for Complex Search Tasks. In *Proceedings of the 16th ACM International Conference on Multimedia* (Vancouver, British Columbia, Canada) *(MM '08)*. Association for Computing Machinery, New York, NY, USA, 489–498. https://doi.org/10.1145/1459359.1459424

[59] Geoffrey I Webb, Michael J Pazzani, and Daniel Billsus. 2001. Machine learning for user modeling. *User modeling and user-adapted interaction* 11 (2001), 19–29.

[60] Kanit Wongsuphasawat, Dominik Moritz, Anushka Anand, Jock Mackinlay, Bill Howe, and Jeffrey Heer. 2016. Voyager: Exploratory Analysis via Faceted Browsing of Visualization Recommendations. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (jan 2016), 649–658. https://doi.org/10.1109/TVCG.2015.2467191

[61] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. NPA: Neural News Recommendation with Personalized Attention. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 2576–2584. https://doi.org/10.1145/3292500.3330665

[62] Chuhan Wu, Fangzhao Wu, Yongfeng Huang, and Xing Xie. 2023. Personalized News Recommendation: Methods and Challenges. *ACM Trans. Inf. Syst.* 41, 1, Article 24 (jan 2023), 50 pages. https://doi.org/10.1145/3530257

[63] Chuhan Wu, Fangzhao Wu, Yongfeng Huang, and Xing Xie. 2023. Personalized News Recommendation: Methods and Challenges. *ACM Trans. Inf. Syst.* 41, 1, Article 24 (jan 2023), 50 pages. https://doi.org/10.1145/3530257

[64] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. 2020. MIND: A Large-scale Dataset for News Recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 3597–3606. https://doi.org/10.18653/v1/2020.acl-main.331

[65] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. 2019. Explainable AI: A brief survey on history, research areas, approaches and challenges. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*. Springer, Springer, 563–574.

[66] Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2022. Fairness in Ranking, Part I: Score-Based Ranking. *ACM Comput. Surv.* 55, 6, Article 118 (dec 2022), 36 pages. https://doi.org/10.1145/3533379

[67] Wenbin Zhang and Steven Skiena. 2010. Trading Strategies to Exploit Blog and News Sentiment. *Proceedings of the International AAAI Conference on Web and Social Media* 4, 1 (May 2010), 375–378. https://doi.org/10.1609/icwsm.v4i1.14075

[68] Jian Zhao, Steven M. Drucker, Danyel Fisher, and Donald Brinkman. 2012. TimeSlice: Interactive Faceted Browsing of Timeline Data. In *Proceedings of the International Working Conference on Advanced Visual Interfaces* (Capri Island, Italy) *(AVI '12)*. Association for Computing Machinery, New York, NY, USA, 433–436. https://doi.org/10.1145/2254556.2254639

[69] Xiaojin Zhu. 2015. Machine Teaching: An Inverse Problem to Machine Learning and an Approach toward Optimal Education. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (Austin, Texas) *(AAAI'15)*. AAAI Press, 4083–4087.